frontiers | Research Topics

Fuzzy boundaries: Ambiguity in speech production and comprehension

Edited by
Christopher Carignan, Georgia Zellou, Eleanor Chodroff
and Joseph V. Casillas

Published in
Frontiers in Communication
Frontiers in Psychology

ISSN 1664-8714
ISBN 978-2-83252-102-1
DOI 10.3389/978-2-83252-102-1

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Fuzzy boundaries: Ambiguity in speech production and comprehension

Topic editors

Christopher Carignan — University College London, United Kingdom
Georgia Zellou — University of California, Davis, United States
Eleanor Chodroff — University of York, United Kingdom
Joseph V. Casillas — Rutgers, The State University of New Jersey, United States

Citation

Carignan, C., Zellou, G., Chodroff, E., Casillas, J. V., eds. (2023). *Fuzzy boundaries: Ambiguity in speech production and comprehension*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83252-102-1

Table of contents

TYPE Editorial
PUBLISHED 23 December 2022
DOI 10.3389/fcomm.2022.1112753

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Xiaolin Zhou,
Peking University, China

*CORRESPONDENCE
Christopher Carignan
✉ c.carignan@ucl.ac.uk
Joseph V. Casillas
✉ joseph.casillas@rutgers.edu
Eleanor Chodroff
✉ eleanor.chodroff@york.ac.uk
Georgia Zellou
✉ gzellou@ucdavis.edu

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

RECEIVED 30 November 2022
ACCEPTED 09 December 2022
PUBLISHED 23 December 2022

CITATION
Carignan C, Casillas JV, Chodroff E and Zellou G (2022) Editorial: Fuzzy boundaries: Ambiguity in speech production and comprehension. *Front. Commun.* 7:1112753.
doi: 10.3389/fcomm.2022.1112753

Editorial: Fuzzy boundaries: Ambiguity in speech production and comprehension

Christopher Carignan[1]*, Joseph V. Casillas[2]*, Eleanor Chodroff[3]* and Georgia Zellou[4]*

[1]Department of Speech, Hearing and Phonetic Sciences, University College London, London, United Kingdom, [2]Department of Spanish and Portuguese, Rutgers, The State University of New Jersey, New Brunswick, NJ, United States, [3]Department of Language and Linguistic Science, University of York, York, United Kingdom, [4]Department of Linguistics, University of California, Davis, Davis, CA, United States

KEYWORDS

speech, production, perception, sound change, data science, ambiguity

Editorial on the Research Topic
Fuzzy boundaries: Ambiguity in speech production and comprehension

Language is a system of discrete and abstract elements. Yet, we can rarely (if ever) identify predictable, linear, or clear one-to-one relationships between the speech signal and linguistic categories. Rather, the relationship between speech and language consists of fuzzy boundaries between categories and myriad sources of ambiguity. Early research may have attributed much of this ambiguity to equipment error, less than ideal recording conditions, population under-sampling, or other sources of spurious behavior in the data. Upon closer inspection, however, many researchers have identified a richness and systematicity in the fuzzy mapping from speech to language: ambiguity may play a crucial role in the development, evolution, and realization of language itself. Listeners may benefit from acoustic variability when learning phonological categories and generalizing from them across phonological contexts. Ambiguity about the source of acoustic effects can serve as a catalyst of sound change actuation. Speakers adapt their productions when the environment could make their speech ambiguous to listeners. Gradiency in linguistic representations could allow greater flexibility for listeners to adjust to cross-speaker and cross-situational variation.

The current research era presents opportunities for tackling this difficult topic in ways that have never before been possible or in some cases even imaginable. Recent trends and techniques involving co-registration of multiple data streams allow us to disentangle the articulatory source of observable acoustic effects of vocal tract dynamics, in spite of complicated many-to-one or even many-to-many articulatory-acoustic mappings. The interdisciplinary and trans-global collaborative research that is becoming increasingly popular in our virtual age encourages a wide range of interpretations and strategies for dealing with ambiguous data. Cutting edge machine learning techniques and statistical approaches can help dis-ambiguate fuzzy data patterns to uncover meaningful underlying structure. Virtual experiment platforms that have flourished in recent times can be used to collect participant response data at a scale that was previously unthinkable, allowing novel insight into group-level patterns that characterize the cognitive processing of potentially ambiguous speech signals.

Rather than consider the ambiguous relationship between speech and language as mere noise, or even avoid it entirely in study design and the interpretation of study results, this Frontiers Research Topic seeks to highlight ambiguity itself as a central aspect of the research and object of observation. Our call for papers resulted in 11 original contributions that represent a range of perspectives within the topic of ambiguity in speech production and perception. The articles in this collection all present empirical research that centered around four major themes.

The first theme covers research in perceptual cue-weighting and cue-trading. Four contributions fall under this theme. Guo and Kwon examine the relation between stop aspiration and post-stop F0 in the production and perception of the laryngeal contrast in Mandarin Chinese. They find variations in F0 perturbations across tones which they explain as due to interactions between aerodynamic forces, vocal fold tension, and tonal targets. Yet, in perception, listeners associate high F0 with aspirated plosives. The contribution of this paper for fuzzy boundaries is a detailed exploration of mismatches between production and perception for contrasts that involve complex laryngeal gestures.

Phillips examines the time course for how listeners use anticipatory coarticulation on /s/ for an upcoming rhotic segment. Coarticulation has been considered by some as contributing to "noise" in the speech signal, variation that makes sound categories more "fuzzy", yet this paper finds that listeners use coarticulatory variation immediately, as soon as those cues become available, and further that immediate integration strategies were strengthened when the coarticulatory cues of retraction were stronger and when they were more predictable.

Yu identifies top-down influences of the listener's perception of the talker's persona on the stop voicing contrast. The combination of the listener's gender and the listener's perception of the speaker's socio-indexical properties, such as attractiveness, gayness, or confidence, significantly influences stop categorization, even for the same acoustic stimulus. Perceptual boundaries can therefore be a bit blurred before taking into consideration the listener's in-the-moment perception of the speaker along various socio-indexical dimensions.

The final contribution under this theme comes from Lo in a study exploring the role of F0 as a cue to stop voicing in non-tonal and tonal languages. Lo analyzes the production and perception of stops in Mandarin-English bilinguals. F0 is considered a secondary cue to voicing in English, but serves as a critical acoustic correlate of tone in Mandarin. Participants completed two tasks: a reading production task and a two-alternative forced-choice identification task using stimuli drawn from a bilabial stop continuum in which VOT and F0 were manipulated orthogonally. The results of the production task show that post-stop F0 is consistently higher for voiceless stops when compared with voiced stops. This F0 disparity is larger in the bilinguals' English production than in Mandarin. Lo ascribes this difference to post-stop F0 receiving more weight in English. The perception data also reflect this weighting. Overall, stimuli with higher post-stop F0 are more likely to be identified as voiceless, but the probability of a voiceless response is even higher when the participants believe they are hearing English words. This study underscores a general flexibility, present not only in perceptual boundaries, but also in bilingual cue-weighting strategies, when producing and perceiving similar contrasts in typologically different languages.

The second theme of this collection targets the role of acoustic and/or perceptual ambiguity in sound changes in progress. Bi and Chen identify incomplete neutralization of two falling tones in Dalian Mandarin Chinese, tones 1 and 4. Though the phonetic form of these tones are typically transcribed with the same Chao tone numerals of 51, this study finds subtle but statistically significant differences in F0 contour and velocity profile across two generations of speakers. Lexical frequency and homophone neighborhood density also interact with the phonetic realization of each tone. These findings indicate incomplete neutralization, with additional fuzziness in the exact phonetic instantiation coming from influences of lexical frequency, homophone neighborhood density, as well as their interactions with speaker generation.

Zhang et al. evaluate the production-perception link in two marginal contrasts of Chicagoland English: [ɑ−ɔ] ("cot-caught") and [ʌi–aɪ] ("writer-rider"). The former represents a phonological merger in this variety, and the latter a phonemic split. Individuals from this speech community provided production data by reading cot-caught and writer-rider pairs embedded in sentences and in isolation. The perception data was derived from ABX and two-alternative forced-choice tasks. Zhang et al. provide evidence suggesting that the production/perception link may follow a different trajectory depending on the type of sound change in question, i.e., a phonological merger vs. a phonemic split. This study highlights the manner in which data from fuzzy contrasts can contribute to our understanding of sound change and language acquisition processes.

Zahner-Ritter et al. investigate the form and function of three rising-falling contours—L + H*, (LH)*, and L* + H—found in German *wh*-questions across Northern and Southern varieties of German. The production results indicate reasonable separation among contours, but also some degree of fuzziness, especially for Southern German speakers with respect to the L + H* and (LH)* contrast. The perception results reveal very distributed and somewhat fuzzy meaning associations for each of the contour types: for both dialects, L + H* and L* + H accents are largely interpreted as information-seeking, whereas

(LH)* has a more distributed meaning, and is much more likely to be interpreted in both dialects as a negative attitude or aversion.

The **third** theme of this collection involves perceptual adaptation to speech that is variable in both time and space. Temporal boundaries of speech perception may be fuzzy: speech unfolds in time and variations in the duration and coordination of temporal events can affect how speech is perceived. Inappropriate gaps between syllables is a core diagnostic feature of childhood apraxia of speech (CAS), yet no baseline exists in the literature concerning how adults perceive inappropriate gaps in the speech of typically developing children. O'Farrell et al. address this issue by investigating the perceptual threshold for inter-syllabic temporal gaps from 84 adult listeners, using speech samples from typically developing children digitally altered to insert gaps. They find that 80% accuracy in detecting inappropriate gaps occurs for intervals between 100 and 125 ms, and 90% accuracy for intervals between 125 and 150 ms. This finding provides the first evidence of the perceptual limen of syllable segregation, which can provide a threshold for a therapy goal for treatment of CAS.

"Spatial" boundaries of speech perception may also be fuzzy: perceptual boundaries between categories are malleable and can shift as speech production traverses through myriad domains of sensory input. Previous studies have shown that repeated exposure to a particular acoustic stimulus can shift a listener's perceptual boundary toward that stimulus, a phenomenon known as selective adaptation. Ito and Ogane use orofacial skin stretching to investigate whether the category boundary between /ε/ and /a/ is similarly affected by repeated *somatosensory* exposure. They find that exposure to a particular somatosensory stimulus (in this case, pulling the skin upward in a manner consistent with the production of /ε/) results in selective adaptation in the same way as acoustic exposure: participants perceive /a/ more than /ε/ after repeated somatosensory training, suggesting that the perceptual boundary is shifted toward the repeated exposure stimulus, /ε/. These results may simulate the natural sensory pairing which occurs during speech production and, thus, support the idea that somatosensory inputs contribute to the formation of sound representations.

The **fourth** theme deals with the perception-production link specifically by looking at "own speech". Two contributions examine how listeners' perception of their own speech can shed light on questions of speech representation. This line of research stems from the fact that speakers are generally more accurate and efficient when processing familiar accents and voices. Cheung and Babel examine the own-voice benefit utilizing Cantonese-English bilinguals' productions of minimal pairs to generate personalized two-alternative forced-choice perception tasks. That is, the bilingual listeners identify instances of Cantonese words which were manipulations of their own voice, as well as productions of other speakers. Cheung and Babel find that the bilinguals are more successful identifying instances of their own manipulated voice than when they are presented with tokens from other speakers, even when said speakers maintain the same degree of acoustically contrastive minimal pairs. Cheung and Babel conclude that phonological contrasts may be primarily shaped by the distributions of our own phonetic realizations. This study highlights the variability present in bilingual speech for producing contrasts. Importantly, it sheds light on how this variability relates to perception, particularly with regard to our understanding of how familiarity aids speech processing, even in presence of a more ambiguous signal.

Finally, Baxter et al. provide a partial replication study in which they evaluate the claim that one's own speech processing can be affected when interacting with L2 speakers. Specifically, this thread of research suggests that processing costs due to increased cognitive effort can affect one's memory of a conversation. In their study, L1 English speakers interact with other L1 English speakers as well as L2 English speakers of intermediate and advanced proficiency. The results suggest speakers display more accurate recall when interacting with L1 speakers in some conditions. The authors conclude that recall accuracy may be modulated by the degree of processing costs incurred and, in turn, result in fuzzier lexical/semantic representations of their own speech.

The contributions to this Research Topic provide wide-ranging and varied perspectives on ambiguity in speech production and perception. The contributions open questions and provide many ripe avenues for future research in this area.

Author contributions

CC, JC, EC, and GZ contributed equally to the conceptualization, writing, and article summaries of the editorial. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Communication

ORIGINAL RESEARCH
published: 25 March 2022
doi: 10.3389/fcomm.2022.839415

The Adult Perceptual Limen of Syllable Segregation in Typically Developing Paediatric Speech

Ciara O'Farrell, Patricia McCabe, Alison Purcell and Rob Heard*

School of Health Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW, Australia

OPEN ACCESS

Edited by:
Christopher Carignan,
University College London,
United Kingdom

Reviewed by:
Min Ney Wong,
Hong Kong Polytechnic University,
Hong Kong SAR, China
Theresa Schoelderle,
Ludwig Maximilian University of
Munich, Germany

****Correspondence:***
Patricia McCabe
tricia.mccabe@sydney.edu.au

Specialty section:
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Communication

Received: *19 December 2021*
Accepted: *25 February 2022*
Published: *25 March 2022*

Citation:
O'Farrell C, McCabe P, Purcell A and
Heard R (2022) The Adult Perceptual
Limen of Syllable Segregation in
Typically Developing Paediatric
Speech. Front. Commun. 7:839415.
doi: 10.3389/fcomm.2022.839415

Inappropriate gaps between syllables are one of the core diagnostic features of both childhood apraxia of speech and acquired apraxia of speech. However, little is known about how listeners perceive and identify inappropriate pauses between syllables (gap detection). Only one previous study has investigated the perception of inappropriate pauses between syllables in typical adult speakers and no investigations of gap detection in children's speech have been undertaken. The purpose of this research was to explore the boundaries of listener gap detection to determine at which gap length (duration) a listener can perceive that an inappropriate pause is present in child speech. Listener perception of between-syllable gaps was explored in an experimental design study using the online survey platform Qualtrics. Speech samples were collected from two typically developing children and digitally manipulated to insert gaps between syllables. Adult listeners ($n = 84$) were recruited and could accurately detect segregation on 80% of presentations at a duration between 100 and 125 ms and could accurately detect segregation on 90% of presentations at a duration between 125 and 150 ms. Listener musical training, gender and age were not correlated with accuracy of detection, but speech pathology training was, albeit weakly. Male speaker gender, and strong onset syllable stress were correlated with increased accuracy compared to female speaker gender and weak onset syllable stress in some gap conditions. The results contribute to our understanding of speech acceptability in CAS and other prosodic disorders and moves towards developing standardised criteria for rating syllable segregation. There may also be implications for computer and artificial intelligence understanding of child speech and automatic detection of disordered speech based on between syllable segregation.

Keywords: speech disorder, auditory perception, child, apraxia, artificial intelligence

INTRODUCTION

Childhood Apraxia of Speech (CAS) is "a neurological childhood speech sound disorder in which the precision and consistency of movements underlying speech are impaired in the absence of neuromuscular deficits" (American Speech-Language-Hearing Association, 2007). These difficulties in planning and sequencing speech movements result in decreases in the precision, consistency, and intelligibility of speech.

This core deficit of motor planning can be identified by observable speech behaviours, including 'inconsistent errors on consonants and vowels on repeated productions of syllables and words, lengthened and disrupted coarticulatory transitions between sounds and syllables, and inappropriate prosody particularly in lexical or phrasal stress' (American Speech-Language-Hearing Association, 2007, Definitions of CAS section, para 2.). CAS is thought to have a genetic origin (e.g., Fedorenko et al., 2016), and many single genes have been implicated as causal (Hildebrand et al., 2020) however to date, idiopathic cases predominate. The gold standard of CAS diagnosis in clinical practise is the judgement of perceptual speech features including inappropriate pauses or gaps on transitions between sounds or syllables judged by expert listeners (Murray et al., 2015).

Syllable segregation occurs within a word when the movement from one syllable to the next is disrupted by an inappropriate pause (Brown et al., 2018). Syllable segregation is a hallmark diagnostic feature of CAS representing the reported difficulty transitioning between syllables. Syllable segregation was identified by Murray et al. (2015) as a key symptom of CAS diagnosis, along with poor lexical stress matches, reduced percentage phonemes correct in polysyllabic words, and reduced articulatory accuracy on repetition of a diadochokinetic speech task. Syllable segregation is therefore both a key identifying feature in CAS, and important in differential diagnosis of CAS from other speech disorders (Murray et al., 2015).

Despite the significance of syllable segregation as a diagnostic feature of CAS, there has been little examination of the perceptual characteristics of between-syllable segregation in the speech of children. There are currently no accepted criteria against which to rate segregation (Brown et al., 2018), and there is little research literature regarding how between-syllable segregation in children's speech is perceived by listeners. One study (Shriberg et al., 2017) investigated between-word segregation, however within-word segregation may be a more valuable diagnostic tool, particularly in minimally verbal children. Reporting of segregated speech currently relies on perceptual judgement and there is no existing standard value for the duration of between-syllable segregation which would be considered disordered. In order to know what is perceived to be distorted or disordered, we must first know what is typical. It is therefore important to understand the perception of syllable segregation in the speech of typically developing children as a potential standard from which we can determine disordered production.

Previous research exploring perception of between-syllable pauses has primarily focused on "gap detection," which refers to a listener's ability to detect a noiseless temporal gap between two stimuli (e.g., Mishra et al., 2014). Research has typically focused on either "within-channel" gap detection where the non-speech sounds on the boundary of the gap are spectrally symmetrical, or "between-channel," where the non-speech sounds bordering the gap are spectrally asymmetrical and therefore more closely resemble speech signals. Gap detection thresholds within the literature vary with stimuli and listener. For example, Heldner (2011) reported that the gap detection threshold varied from 58 to 204 ms.

One study has investigated adult perception of syllable segregation *per se*. Brown et al. (2018) investigated perception of syllable segregation in adult speech and found that the perceptual limen of syllable segregation for adult listeners when listening to words with inserted gaps created from ambient noise was 80 ms at an 80% accuracy threshold. In Brown's study, a fixed anchor method was used in which an anchor stimulus with no manipulated gap and a stimulus with the artificial gap were presented in series. Participants judged the second stimulus as to whether they could hear a gap within the word. This was a type of modified just noticeable difference (JND). Full JND was not used in Brown's study for pragmatic reasons, that is, to reduce the number of presentations. Full JND would have taken 475 presentations, significantly greater than the 80 presentations used. Such a JND approach was therefore not used to answer the fundamental question of the research which was to establish the level at which any segregation is perceived by the majority of listeners. This level is known as the perceptual limen of syllable segregation.

Importantly for CAS diagnosis, the perceptual limen of syllable segregation in typically developing children's speech has not yet been studied. There are known suprasegmental differences between adult and child speech (e.g., Lee et al., 1999), which may result in a higher limen of perception for syllable segregation. Child and adult speech differ significantly in the following ways: children's speech is characterised by higher fundamental and secondary formant frequencies, increased duration of fricative consonant length, higher consonant-vowel duration ratios, and more similar spectral characteristics of different phones than adult productions of the same sounds (Gerosa et al., 2006). Children's speech is also slower, with the movement of articulators less coordinated than in adult speech (e.g., Cychosz et al., 2019), and children produce more consonant distortions as part of typical development (e.g., Storkel, 2019). Similarly, durational variability for children's speech is greater than adult's speech, converging to adult levels around age 13 years (Gerosa et al., 2006). These features may contribute to a lengthened perceptual limen for between-syllable segregation compared to perception of the same phenomena in adult speakers.

Musical training, speech pathology training, age and gender have been identified as factors which may impact perception of auditory features (Pichora-Fuller et al., 2006; Giannela-Samelli and Schochat, 2008; Mishra et al., 2014; Brown et al., 2018). Musicians have significantly lower between-channel gap detection thresholds compared to non-musicians (Mishra et al., 2014; Elangovan et al., 2016), with one study finding that between channel gap detection thresholds in musicians were on average half those in non-musicians (Mishra et al., 2014). However, it is important to note that within-channel gap detection stimuli do not fully represent the complexity of speech sound signals and therefore cannot be readily generalised to the perception of between-syllable segregation (Brown et al., 2018). Only one study has examined differences in accuracy of gap detection resulting from speech pathology training. This study compared accuracy between untrained listeners and experienced speech pathologists rating the presence of syllable segregation and found a difference in accuracy of identification at the 90% accuracy threshold (Brown et al., 2018). Younger age is also correlated with increased accuracy of gap detection (Pichora-Fuller et al., 2006). Gap detection thresholds have been found to be greater for older listeners (67–82 years old, mean 75 years) than for younger listeners (21–35 years old, mean 24 years) (Pichora-Fuller et al., 2006). Few studies have examined the relationship between listener gender and accuracy of gap detection. One study reported males performed slightly better in the gaps-in-noise test, which uses white noise as a stimulus (Giannela-Samelli and Schochat, 2008) and is therefore of limited utility to between-syllable gap detection.

There is also limited previous research regarding stimuli factors which may influence perception of auditory features. Speaker gender has been identified as a factor which may influence perception. Existing research suggests that female speakers may be overall more intelligible than male speakers (Markham and Hazan, 2004; Yoho et al., 2018) in both subjective and objective measures, however no existing research has investigated the interaction of gender and perception of syllable segregation. Similarly, stress pattern of spoken stimuli may influence perception, although there is limited research investigating syllable stress pattern and perception of syllable segregation. One previous study (Brown et al., 2018) found that syllable stress pattern was weakly correlated with accuracy of gap detection. These factors therefore warrant further investigation.

Despite the known differences in production between adult and child speakers, no comparison between what is acceptable in adult and child speech has been undertaken regarding within word pauses. That is, it is unknown whether the limen of perception in child speech is similar to that reported for adult speakers or not. Such a comparison may provide valuable information for speech pathologists in working with children with CAS including assisting in identifying the need to train listeners to what is typical in child speech for therapy accuracy. Additionally, it may assist in the design of speech recognition systems, which are largely trained on adult speech (Shahin et al., 2020). These systems have shown a substantial degradation in performance when tested on child speech, due to the linguistic and acoustic mismatches outlined above (Shahin et al., 2020). There is therefore a gap in the existing literature regarding the differences in the perceptual limen of adult speech compared to child speech.

Non-words may be most appropriate to investigate listeners' perception of syllable segregation for multiple reasons, including that a listeners' pre-existing idea of words' pronunciation may cause potential confounds with their perception of the word (Gierut et al., 2010) and non-words separate perception from any semantic context. Importantly, previous research in detection of syllable segregation used non-words to investigate listener perception (Brown et al., 2018).

Despite syllable segregation being a diagnostic feature of CAS, understanding the duration of the gap between syllables is an emerging field. If a value for the perceptual limen of between-syllable gaps is identified, this may be used to contribute to the development of standardised training and rating tools which could be used in both diagnosis and treatment of CAS. The purpose of this study was therefore to explore the perceptual boundaries of adult listeners when judging artificial syllable segregation in the speech of typically developing children.

Research Questions

1a. What is the threshold for accurate detection of a between-channel gap within non-words?

1b. What is the strength of relationship between gap duration and between-channel gap detection accuracy?

2. Do stimulus factors impact the listener perceptual limen of between-syllable segregation?

    a. Do non-words with a strong onset syllable stress pattern have a shorter perceptual limen than weak onset syllable stress patterns?
    b. Does speaker gender affect the perceptual limen of syllable segregation?

3. Do listener factors impact the listener perceptual limen of between-syllable segregation?

    a. Do listeners with musical training have a shorter perceptual limen compared to listeners without musical training?
    b. Do listeners with speech pathology training have a shorter perceptual limen compared to listeners without speech pathology training?
    c. Do younger listeners have a shorter perceptual limen compared to older listeners?
    d. Does listener gender affect the perceptual limen of syllable segregation?

METHOD

This study used a cross sectional experimental design using the online platform Qualtrics (Qualtrics, 2021). The research was approved by The University of Sydney Human Research Ethics Committee (2021/753). There were two groups of participants involved in this study—child speaker participants (hereafter referred to as "speakers") and adult listener participants (hereafter "listeners"). All participants gave informed consent to participate.

Speakers

Eligibility and Recruitment

To be considered eligible for this study, children were required to speak English with an Australian accent, have hearing within the normal range, have typically developing speech and language, and no structural or neuromuscular deficits as determined by an oral-musculature assessment completed by an experienced qualified speech-language pathologist (the second author). Two children were recruited and parents provided written consent. Speakers were therefore 1 male child and 1 female child, aged 10 and 8 respectively.

Stimuli

Following Brown et al. (2018), a set of 4 non-words from the Syllable Repetition Task (Shriberg et al., 2009) (ma'da, 'maba, da'ba, 'bada) were selected as target productions. These words were chosen as they were two syllable words suitable for acoustic manipulation which contained a variety of stress patterns including two strong onset words ('bada, 'maba) and two weak onset words (da'ba, ma'da) to examine any differences in listener perception as a result of stress pattern.

Single word utterances were used in preference to connected speech to reduce the influence of other words in the utterance on the listener's perception of a word. Two syllable non-words were used to ensure comparability of results with previous research (Brown et al., 2018).

Speakers were asked to imitate an adult female (second author) saying the stimuli. Samples were recorded using Audacity 2.4.2® (Audacity Team, 2021) in a quiet space using a head mounted AKG microphone at a mouth to microphone distance of 5 cm and a Roland Quad Capture sound card attached to a laptop computer. No audible distortions were found in the stimuli when reviewed.

Stimuli Preparation

Sample preparation followed Brown et al. (2018). Samples were edited using Audacity 2.4.2® software (Audacity Team, 2021). All samples were normalised to −1.0 dB to ensure volume was consistent across samples and the "noise reduction" feature in Audacity was applied to remove background noises or distortions in the clip that could interfere with a listener's perception of the recording. The inserted recorded gap was copied from periods of ambient sound in the clip instead of pure silence which, if used, may have resulted in detectable sound distortions. Gaps of the selected ambient sound, ranging in duration from 25 to 200 ms, were then inserted into the single word samples. Gaps were inserted at the pre-voice onset pause between the first and second syllable of the four non-words.

Length and Number of Gaps in Stimuli

Two small pilot studies (total $n = 7$) were initially conducted with gaps of 50 ms increments (50–200 ms) based on previously reported gap detection research (Brown et al., 2018). All listeners were able to detect segregation at 200 ms. The pilot results suggested that the 80% accuracy threshold was at least 100 ms and no higher than 150 ms, and the 90% accuracy threshold at least 150 ms and no higher than 200 ms, indicating the need for smaller increments to reliably determine the limen of perception as well as the need for a gap condition of 175 ms. These findings combined with prior research regarding gap detection (Brown et al., 2018) indicated 25 ms was the most appropriate gap increment. An upper limit of 200 ms was therefore chosen as a gap all listeners should be able to detect reliably. A total of nine gap conditions were therefore used (1) no gap, (2) 25 ms gap, (3) 50 ms, (4) 75 ms, (5) 100 ms, (6) 125 ms, (7) 150 ms, (8) 175 ms, and (9) 200 ms. Pilot participants did not participate in the primary study.

Listener Eligibility and Recruitment

Listeners were then recruited to judge the stimuli. Listeners were required to be between 18 and 59 years of age. This age range was selected to reduce the impact of presbycusis and age-related cognitive decline. Listeners were required to have no current or previous history of hearing loss, no self-reported current ear infection, no self-reported current or history of cognitive impairment, and to be an Australian English speaker. All listeners were asked to undertake a hearing screen using Hearing Australia Online Hearing Assessment (Hearing Australia, 2021) and self-report a result within the normal range. Listeners were recruited via social media, word of mouth, and advertising within The University of Sydney.

Listeners and Data Preparation

A total of 140 listeners aged 18–59 consented to participate in the study. No identifying information was collected about listeners.

Some 49 listeners started the survey but did not complete any listening tasks. These listeners were removed from the data set. Three (3) listeners who answered either "yes, segregated," or "no, not segregated" to all questions were removed from the data set. Three (3) listeners who only answered one question were removed. One listener achieved a mean score of 29.1% compared to the mean of all listeners, which was 69.5%. The apparent difficulty this listener had with the task suggested they may not actually meet the inclusion criteria, and so they were removed from the data set. A total of 56 listeners were therefore removed from the data set without analysis. Five (5) listeners partially completed the listening tasks but failed to complete the entire study. These listener responses were included in the data analysis and consequently some analyses have varying participant numbers.

A total 84 listeners (61 women, 22 men, 1 other) were therefore included in the data analysis. The mean age was 28.4 years (SD 11.3; range 18–59). Nineteen (19) listeners indicated that they had received musical training, which was defined as either having received musical training within the previous 5 years or practising as a professional musician, and 46 had received speech pathology training. Speech pathology training was defined as a listener being either a qualified speech pathologist or a speech pathology student. Of these, 42 were speech pathology students and 4 were qualified speech pathologists. Demographic data regarding age, speech pathology training, musical training, listener gender was collected and is reported in **Table 1**.

Procedure

A set of 80 (4 stimuli x 9 gap conditions x 2 speaker genders + 10% repeats) were played in two randomised orders. All modified words spoken by the male child were placed in a random order block and all modified words spoken by the female child were similarly blocked. The order of the two blocks was switched halfway through the data collection period to reduce any order effect associated with the gender of the speaker. Participants were asked to respond to "Indicate if you did hear segregation or did not hear segregation." Binary choice answer options were "yes, segregated" and "no, not segregated." Binary choice has been found to reduce bias in ratings (Harvey, 2016). No feedback was provided.

Data Analysis

To answer research question 1a, the percentage of stimuli detected accurately for each gap condition was calculated across all listeners and graphed, to indicate trends by gap condition. The

TABLE 1 | Description of subgroups in the data.

	Speech Pathology Training (n; percent of total)	Musical Training (n; percent of total)	Neither Speech Pathology nor Musical Training (n; percent of total)	Total
Female	40 (47.6%)	17 (20.2%)	4 (4.8%)	61
Male	5 (5.9%)	2 (2.4%)	15 (17.9%)	22
Other	1 (1.2%)	0 (0.0%)	0 (0.0%)	1
Total	46 (55.9%)	19 (22.6%)	19 (41.4%)	84

limen of perception for listeners was marked at both 80 and 90% accuracy thresholds, to include both accuracy thresholds used in syllable segregation research previously (Brown et al., 2018).

A second measure of accuracy was used to answer research questions 1b, 2a, and 2b. Because participants made eight responses for each gap condition (four words by two speaker genders), it was possible to calculate a proportion of correct responses at each gap condition. The 0 ms gap (control) condition was excluded in statistical analyses, to investigate only perception of inserted gaps. Kolmogorov-Smirnov tests of normality (Chakravarti et al., 1967) showed distributions were heavily skewed for some gap durations. The design was repeated measures because participants had accuracy scores for all gap conditions. The non-parametric Friedman test (Friedman, 1937) tested equality of median accuracy across gap durations. Strength of relationship between accuracy and gap condition was calculated by converting the Friedman *p* value to a correlation r value. The conversion was done by finding the z score on the standard normal distribution which corresponded to the Friedman *p* value, then applying the formula $r = |z|/\sqrt{n}$ (Ratner, 2009).

To address research questions 2a and 2b, the relationships between the stimulus factors (word stress pattern and speaker gender) and accuracy, accuracy was first graphed to show gap durations with separations in accuracy for strong/weak onset words vs. weak/strong onset words, and between female and male speakers. Differences at these gap durations were then analysed using Wilcoxon signed-rank tests (Wilcoxon, 1945). The test statistics were converted to correlation r values, as an effect size index, using the formula $r = |z|/\sqrt{n}$.

To address the listener factor research questions 3a to 3d, accuracy across all 8 gap conditions was averaged for each listener and then correlated with the dichotomous listener factors using parametric point biserial correlations (r_{pb}) (Cureton, 1956). Kolmogarov-Smirnov tests indicated average accuracy was normally distributed, meaning parametric tests could be used. Listener age was grouped into (1) younger listeners (aged 18–32; 77.4% of listeners) and (2) older listeners (aged 37–59, 22.6% of listeners). These age bands were selected as this was where the data showed a natural break in age distribution.

Supplementary analysis of inter-rater reliability used intraclass correlation coefficients, two-way random with absolute agreement (ICC 2,1) across the average accuracy scores of all 83 raters who rated all gap conditions (one rater did not rate all gap conditions) (Bartko, 1966). ICC values between 0.5 and 0.75 indicated moderate reliability, values between 0.75 and 0.9 indicated good reliability and values >0.9 indicated excellent reliability (Koo and Li, 2016).

Intra-rater reliability of responses was analysed using Cohen's Kappa (Cohen, 1960) due to the binary data collected. A result of >0.8 indicated very good agreement; 0.61–0.8 good agreement; 0.41–0.60 moderate agreement; 0.21–0.40 fair agreement and <0.20 poor agreement (Landis and Koch, 1977; Altman, 1991).

A *post hoc* analysis was conducted to compare the limen of perception in child speech and the limen of perception reported in adult speech (Brown et al., 2018). An individual participant data meta-analysis was conducted to determine the accuracy of listener detection of syllable segregation at each gap duration using raw data obtained from Brown et al. (2018) and the data included in this study. Wilcoxon Rank Sum tests were used to compare accuracy of listeners when listening to adult speech, as in data collected by Brown and colleagues, and when listening to child speech, as collected by the present study. Non-parametric point biserial correlations were used to measure the strength of these differences.

Analyses were conducted using SPSS Version 27.0 (IBM Corp, 2020) and R version 3.1.1 (R Core Team, 2021). For all correlation effect sizes, a small effect was indicated by an r between 0.1 and 0.3, a medium effect between 0.3 and 0.5 and a large effect by >0.5 (Fritz et al., 2011).

RESULTS

The Threshold for Accurate Detection of Between-Channel Gaps in Non-words

Across all listening tasks listeners achieved 60.1% "accurate yes"; 9.5% "accurate no," 0.6% "inaccurate yes"; and 29.7% "inaccurate no." **Figure 1** shows the mean 80 and 90% accuracy thresholds across all listener groups.

The listener limen of perception at 80% accuracy was at least 100 ms and no higher than 125 ms. At 90% accuracy, the limen of perception was at least 125 ms and no higher than 150 ms. Within all sub-groups of listeners, the limen of perception at 80% accuracy was also at least 100 ms and no higher than 125 ms. At the 90% accuracy thresholds, sub-groups differed in their limen of perception. **Table 2** outlines the 80 and 90% accuracy thresholds for each sub-group of listeners. **Figure 2** shows the proportion of accurate gap detections by gap duration.

A Friedman's test with follow up pairwise comparisons showed that there was an increase in accuracy up to 150 ms. After this, there was no statistically significant increase in accuracy of detection. There was a strong relationship between

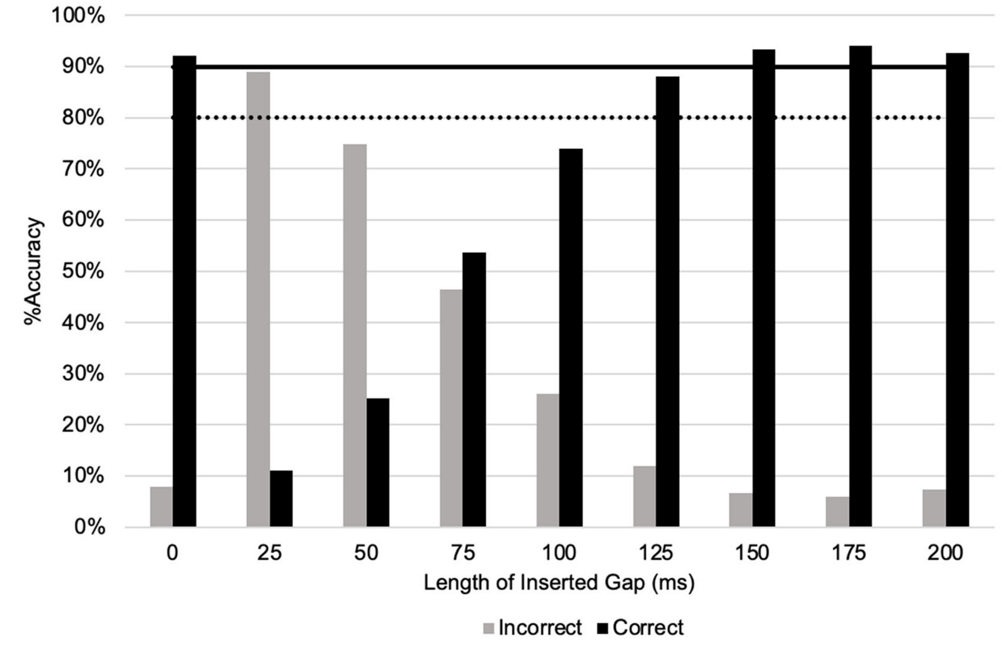

FIGURE 1 | Barchart of mean accuracy across all gap durations and 80 and 90% accuracy thresholds.

TABLE 2 | Eighty and 90% accuracy thresholds for all listeners and for each sub-group of listeners.

Listener Group	Range of 80% Accuracy (ms)	Range of 90% Accuracy (ms)
All listeners (*n* = 84)	100–125	125–150
Female listeners (*n* = 61)	100–125	125–150
Male listeners (*n* = 22)[a]	100–125	150–175
Older Listeners (age 37–59) (*n* = 17)	100–125	100–125
Younger Listeners (age 18–32) (*n* = 67)	100–125	125–150
Listeners with Speech Pathology Training (*n* = 47)	100–125	125–150
Listeners without Speech Pathology Training (*n* = 37)	100–125	150–175
Listeners with Musical Training (*n* = 19)	100–125	125–150
Listeners without Musical Training (*n* = 65)	100–125	125–150
Listeners with neither Speech Pathology nor Musical Training (*n* = 28)	100–125	125–150

[a] *One listener identified as "other" and was therefore not included in this analysis.*

increased gap duration and accuracy of detection ($X_7^2 = 446.56$, $p < 0.01$). This converts to an r effect size measurement of 0.79. The mean and median scores, standard deviation and interquartile range of each gap condition are shown in **Supplementary Material 2**.

There was a positive correlation between increased length of inserted gap and increased accuracy of gap detection, which is shown in **Supplementary Material 3**.

Listener Factors Affecting the Perceptual Limen of Syllable Segregation

An independent samples *t*-test was used to compare the overall accuracy of the listener groups at each duration. This revealed no significant differences between overall accuracy of perception of syllable segregation between male (mean = 69.50%, *SD* = 12.55) and female listeners (mean = 69.66%, *SD* = 10.50, r_{pb} = 0.01); listeners with musical training (mean = 68.66%, *SD* = 11.61, r_{pb} = 0.05) and listeners without musical training (mean = 69.99%, *SD* =10.82, r_{pb} = 0.05); younger listeners (mean = 69.81%, *SD* =11.47), and older listeners (mean = 69.22%, *SD* = 8.91, r_{pb} = 0.02). Listeners with speech pathology training (mean = 71.81%, *SD* = 10.34) were more accurate than listeners without speech pathology training (mean= 66.99%, *SD* = 11.26, r_{pb} = 0.22).

Stimulus Factors Affecting the Perceptual Limen of Syllable Segregation

Graphical screening was used to identify the gap duration with the largest differences between strong and weak onset words, which were then analysed for statistical significance using Wilcoxon signed-ranks test. **Figure 3** was used to visually determine gap conditions for further analysis.

Based on **Figure 3**, 50, 100, and 125 ms were selected for further analysis using the Wilcoxon signed ranks test. This revealed a statistically significant difference in accuracy of detection between strong and weak onset words at these gap conditions. Listeners were more accurate when listening to strong onset words with 100 ms gaps ($r = 0.33$, $p < 0.01$) and 125 ms gaps ($r = 0.25$, $p = 0.01$). Listeners may be more accurate when listening to weak onset words with 50 ms gaps ($r = 0.18$, $p = 0.05$). This is inconclusive.

Graphical screening was used to identify the gap durations with the largest differences between female and male speakers, which were then analysed for statistical significance using the Wilcoxon Signed Ranks test. **Figure 4** was created to determine these points of interest.

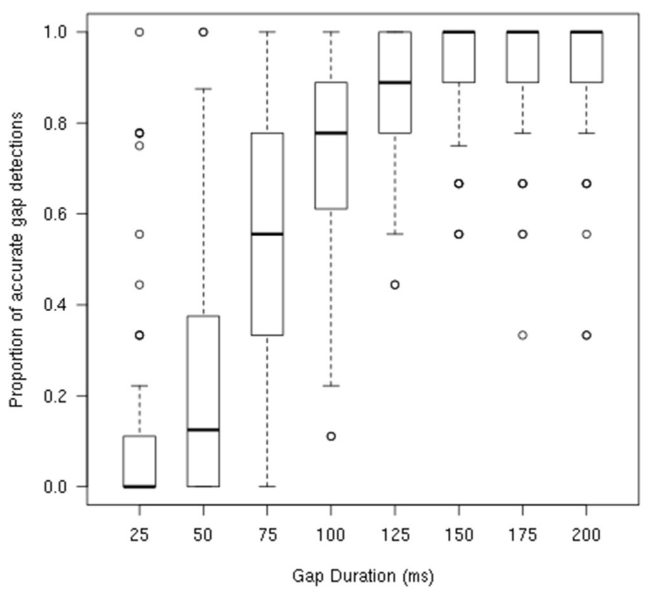

FIGURE 2 | Boxplot showing proportions of accurate gap detections by gap duration.

Based on graphical screening shown in **Figure 4**, gap durations with the largest differences between female and male speakers were 25, 75 and 100 ms. The Wilcoxon Signed Ranks Test revealed a difference in accuracy of listener detection between male and female speakers at these gap conditions. Listeners were more accurate when listening to female speech at 75 ms ($r = 0.29$, $p = 0.005$), and to male speech at 100 ms ($r = 0.22$, p =0.024) and 150 ms ($r = 0.34$, $p = 0.001$).

Supplementary Analysis: Listener Reliability

Listeners had an average intra-rater reliability of $K = 0.709$ (95% CI 0.65–0.77) suggesting listeners had a good level of agreement within their own judgements. Listeners had an average inter-rater reliability of ICC = 0.727 (95% CI 0.545–0.908), suggesting they also had moderate reliability with each other.

The impact of musical training, speech pathology training, listener age and listener gender on inter-rater reliability was investigated. There was no statistically significant difference between these groups. **Table 3** outlines the inter-rater reliability of each group.

Supplementary Analysis: Stimulus Factors Affecting Listener Reliability

The effect of different stimuli factors on listener inter-rater reliability was also investigated. **Table 4** outlines the inter-rater reliability of listeners when listening to different stimulus factors.

Post-hoc Analysis: The Limen of Perception in Child vs. Adult Speech

Table 5 outlines the comparison between listener accuracy when rating children's or adult's speech. The gap conditions with the

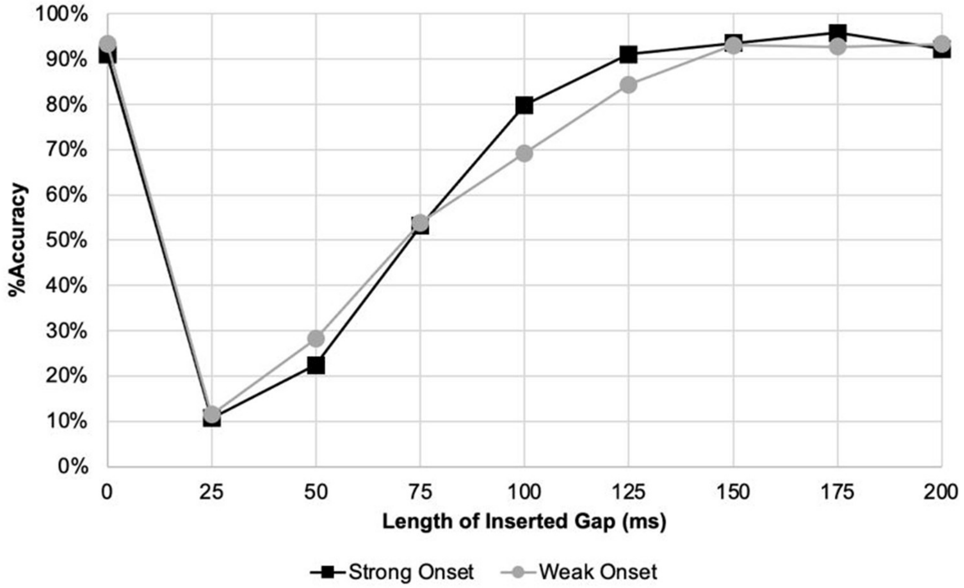

FIGURE 3 | Line graph of listener accuracy when rating strong vs. weak onset stimuli.

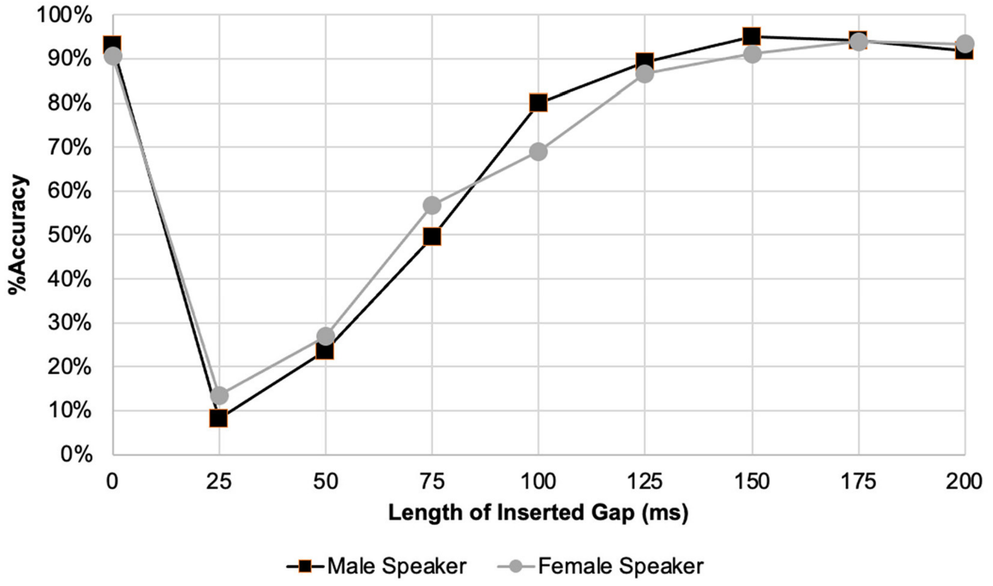

FIGURE 4 | Line graph of listener accuracy when rating male vs. female speaker stimuli.

TABLE 3 | Listener factors and inter-rater reliability.

Stimulus factor	Inter-rater reliability (ICC)	95% Confidence interval	Interpretation of reliability
Musical Training (*n* = 19)	0.74	0.57–0.92	Moderate
Speech Pathology Training (*n* = 47)	0.73	0.54–0.91	Moderate
Neither Musical nor Speech Pathology Training (*n* = 28)	0.70	0.51–0.90	Moderate
Age—Younger (18–32 years) (*n* = 67)	0.70	0.50–0.89	Moderate
Age—Older (37–59 years) (*n* = 17)	0.83	0.670.95	Good
Gender—Female (*n* = 61)	0.76	0.59–0.92	Good
Gender—Male (*n* = 22)	0.63	0.42 –0.86	Moderate

greatest differences in listener accuracy were 75 and 100 ms, where listeners were more accurate when rating adult speech. For all conditions where a significant finding is reported, listeners were more accurate with adult than child samples.

DISCUSSION

The limen of perception of syllable segregation and the listener and speaker features which impact accurate detection of such segregation were variables of interest.

TABLE 4 | Stimulus factors and listener inter-rater reliability.

Stimulus factor	Inter-rater reliability (ICC)	95% Confidence interval	Interpretation of reliability
Stress Pattern —Weak Onset	0.63	0.43–0.86	Moderate
Stress Pattern—Strong Onset	0.73	0.54–0.91	Good
Speaker Gender—Male	0.77	0.60–0.93	Good
Speaker Gender—Female	0.61	0.41–0.85	Moderate

The first question was: what is the perceptual limen of adult listeners for syllable segregation, and how strong is the relationship between syllable segregation (gap) duration and the accuracy of its detection? As expected, the limen of perception of syllable segregation in children's speech was higher than that reported for adult speech (Brown et al., 2018). The threshold of accurate detection of syllable segregation in children's speech at the 80% accuracy threshold was at least 100 ms and no higher than 125 ms. At 90% accuracy, this limen was at least 125 ms and no higher than 150 ms. These values are higher than those reported in Brown et al. (2018) where the 80% threshold was 80 ms and the 90% threshold was 90 ms. The *post-hoc* analysis showed that there were statistically significant differences in listener accuracy when rating adult vs. child speech at gap lengths 75, 100, 125 and 200 ms. The statistically significant difference in accuracy found at 200 ms may be a statistical artefact of a greater proportion of listeners in Brown and colleagues' study correctly

TABLE 5 | Listener accuracy when hearing adult vs. child speech.

Gap duration[a] (ms)	Wilcoxon rank sum test (W)	*P*-Value	Spearman's rho	Effect Size Interpretation
25	1,177	0.85	0.02	Small
50	1,178	0.86	0.02	Small
75	705.5	<0.01	0.31	Medium
100	544.5	<0.01	0.42	Medium
125	843.5	0.01	0.25	Small
150	983	0.07	0.17	Small
200	896	0.01	0.25	Small

Brown et al. (2018) did not include a 175 ms gap in that study. Therefore this gap duration is excluded from comparison.

identifying a gap than in the present study, or may be due to listener fatigue, as this present study included a greater number of presentations ($n = 80$) than Brown and colleagues' study ($n = 32$).

The impact of stimulus factors on the listener perceptual limen of syllable segregation was investigated as the second research question. Research questions regarding which stimulus factors would impact detection were: Do strong onset words have a shorter perceptual limen than weak onset words; and does speaker gender affect the point at which listeners can identify segregation? These findings suggest that adults may be more accurate when detecting gaps in adult speech than in child speech. This potentially higher perceptual limen may be due in part to the different perceptual characteristics of children's speech described previously. When judging syllable segregation in children's speech a different standard may be required when compared with the same judgement for adult speech. That is, compared to adult speech, children's speech may need to have a greater pause duration between syllables to be considered segregated.

Listener Factors Affecting Detection of Syllable Segregation

Unlike Brown et al. (2018), this study used child speech with both male and female speakers as well as a greater number of listeners and a larger number of samples per listener. Thus, these results may contribute more information regarding the speaker and listener factors which influence accurate detection of syllable segregation.

The third question sought to answer: do listener factors impact the listener perceptual limen of between—syllable segregation? These listener factor research questions were: Do listeners with musical training have a shorter perceptual limen compared to listeners without musical training; Do listeners with speech pathology training have a shorter perceptual limen compared to listeners without speech pathology training; Do younger listeners have a shorter perceptual limen compared to older listeners; and, does listener gender affect limen of perception in children's speech?

Musical training, age and gender did not contribute to an individual's overall perceptual accuracy while a weak correlation was found between speech pathology training and accuracy of gap detection. This is in contrast to Brown et al., who found that speech pathology training did not result in a significant difference in perceptual accuracy of syllable segregation (Brown et al., 2018). This may be due to the larger listener group ($n = 84$) used in this study compared to Brown's study ($n = 30$). That is, Brown and colleagues' sample size may not have been sufficiently large to detect a correlation between speech pathology training and accuracy of gap detection.

Listener age and gender were not also correlated with increased accuracy of perception overall. This is in contrast to existing literature regarding perceptual accuracy and age, which suggests that accuracy of detection of perceptual features declines with increased age (Snell and Frisina, 2000). The current finding on age may be due to the listener age restriction in study design and the requirement for listeners to pass a hearing screen prior to beginning the listening tasks, which may have mitigated the effect of any presbycusis present in other studies. Other possible sources of age variation were not examined in this study. The current literature is divided regarding the effect of gender on accuracy of detection of perceptual features. A larger sample may be required to confirm the current finding of no difference.

Stimuli Factors Affecting Detection of Syllable Segregation

Accuracy of detection overall was not correlated with either speaker gender or onset stress pattern across all listener responses. This may be of clinical relevance and of importance to the development of computer and artificial intelligence tools, as this suggests that the speech of both male and female children may be held to the same standard when judging syllable segregation although the small sample size should be acknowledged. Accuracy of detection was correlated with speaker gender and stress onset pattern at some gap durations.

There was a statistically significant difference in accuracy of detection at the 100 ms gap condition for both factors. As the limen of perception at 90% accuracy was between 125 and 150 ms, this difference in detection occurred at a gap length lower than the limen. This suggests that listeners may be more accurate when detecting syllable segregation in strong-weak stress pattern words (compared to weak-strong stress pattern words), and in male speakers, at least in the present sample, when the inserted gap is shorter.

Limitations and Future Directions

This study recruited two typically developing children as speakers, resulting in the need to artificially insert gaps to mimic natural segregation. However, it is possible that these artificial gaps do not truly reflect the natural syllable segregation that occurs in CAS, as other speech features (such as inappropriate lexical stress and speech sound errors) may be involved in listeners' judgments of the presence of syllable segregation (Murray et al., 2015). It must also be considered that the stimuli used were two syllable non-words. This potentially limits our ability to readily generalise these results to naturally occurring syllable segregation in a range of speakers across a range of words.

Inclusion criteria for this study differed from previous studies which examined listener factors and gap detection. This study collected information on musical training, which was defined here as a listener who had received music lessons within the previous five years, or who practised as a professional musician. Other studies which have examined musicians have used more specific selection criteria, including having commenced musical training in childhood and receiving specific academic training (e.g. Mishra et al., 2014; Elangovan et al., 2016). Similarly, of the 84 listeners who were included in the data analysis here, only 19 of these were musicians by the current definition. These factors increase the risk of a type II error as does the limited number of older adults were recruited for the study regarding the age variable. Similarly, most of the listeners with speech pathology training included in this study were students, who have more limited experience in detecting auditory features compared to qualified practising speech pathologists. This may have contributed to the weak differences in accuracy of perception between these groups.

Whilst the reported accuracy thresholds of 80 and 90% for the limen of perception are appropriate for use in a research context, there remains the question of whether these are sufficiently sensitive or specific for a clinical context. Judgments of syllable segregation are most likely to occur in real time in clinical settings, without an anchor stimulus for comparison, and in combination with other speech errors. Additionally, various distractors are present in a clinical setting including background noise and child behaviour. Clinical practise often requires a clinician to rate multiple speech features simultaneously. Perhaps the accuracy threshold for a limen in clinical contexts, and in children with actual CAS, would be higher than reported here.

While it was beyond the scope of this paper, future research should investigate perception of syllable segregation using a wider range of speakers and stimuli. This includes testing non-words with a greater range of phonemes, testing real words, testing polysyllabic real and non-words with a range of lengths and stress patterns, and testing in languages other than English. Future research should also explore listeners' perception of natural syllable segregation occurring in the speech of children with CAS. Such research could provide valuable information regarding listener perception of this feature which could be applied to the development of standardised diagnostic tools, computer and artificial intelligence use in treatment and diagnosis of CAS. Future research should also consider examining the threshold of gap detection using smaller gap increments, for example 5 ms, within the ranges identified as significant here.

Clinical and Practical Implications

This research has a number of clinical implications relevant to the diagnosis and treatment of CAS. Firstly, it provides data on the pause duration at which listeners can perceive segregation in child speech. This data could be used to determine what level of segregation may constitute a significant therapy goal and be used to train clinicians to rate these features more accurately and reliably. For example, Rapid Syllable Transition (ReST) treatment (McCabe et al., 2017) is one of a limited set of evidence-based treatments for CAS. This treatment relies on a clinician's real time perception of syllable segregation. Training clinicians using real and modified samples around the limen could increase the accuracy and speed of such decisions and potentially the efficacy of the intervention. A refined limen of perception of syllable segregation in children with CAS could also be used to develop computer-aided tools which could be used for diagnosis and treatment of CAS.

This research may also aid the development of an AI tool for diagnosis of CAS and other prosodic disorders. Given the limited accessibility of Speech-Language Pathologists (McGill et al., 2020) children may benefit from computer-aided speech therapy tools as a means to reduce waiting lists and increase access generally (Shahin et al., 2020). However, accuracy of automated disordered speech analysis tools is not yet reliable enough to be used clinically (Shahin et al., 2020). Identifying the threshold of accurate gap detection could be used to improve computer-aided tools for the diagnosis and treatment of CAS and other prosodic disorders. Such results may also have implications for further development of computer and artificial intelligence recognition of children's speech more broadly. Current speech recognition systems trained on adult speech show a degradation in performance when used for child speech, due to linguistic and acoustic mismatches between adult and child speech (Shahin et al., 2020). These findings may therefore be useful in improving artificial intelligence in the treatment and diagnosis of children's speech disorders and in understanding child speech in general.

Listener factors of musical training, age and gender were not significantly correlated with accuracy of detection of syllable segregation while speech pathology training was weakly related to increased accuracy of gap detection. This has clinical implications for speech pathology practise and suggests that specific training may be required for clinicians treating CAS or other prosodic disorders which feature syllable segregation. The weak relationship between speech pathology training and average accuracy does however suggest that members of the community may be able to identify syllable segregation in the speech of children with CAS with accuracy not far below clinicians. As increased therapy dosage is related to generalisation of skills (Edeal and Gildersleeve-Neumann, 2011), this finding has implications for service delivery models and utilisation of family members as therapists into speech-language therapies.

CONCLUSION

This study suggests that the limen of perception of syllable segregation in children's speech is at least 125 ms and no higher than 150 ms. This study also suggests that the limen of perception of children's speech is higher than that of adult speech, and that adult listeners are less accurate when detecting syllable segregation at gap lengths of 75, 100, 125 and 200 ms in children's speech compared to adult speech although the latter finding needs confirmation. There is no evidence that there is any difference in listener accuracy or reliability related to musical training, age or gender in their perception of syllable segregation in typical children's speech. There is some evidence which suggests that speech pathology training may result in improved accuracy of gap detection. Overall, the findings provide useful information that may contribute to the development of a standardised rating tool for syllable segregation to be used in the assessment, diagnosis and management of CAS as well as contribute to the further development of computer and artificial intelligence for use in treatment and diagnosis of speech disorders.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The University of Sydney Human Research Ethics Committee. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

CO'F designed the study with support from PM and AP. CO'F and PM collected the data with support from AP. CO'F cleaned and prepared the data. CO'F and RH analysed the data. CO'F wrote the manuscript with support from PM, AP, and RH. All authors approved the final version of the manuscript. CO'F, PM, and RH contributed to the revisions.

ACKNOWLEDGMENTS

The authors would like to thank Tayla Brown for permission to re-analyse the data in Brown et al. (2018). We would also like to thank the participants and Callie Swayze for help during her 2018 internship.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.839415/full#supplementary-material

REFERENCES

Altman, D. (1991). *Practical Statistics for Medical Research*. London: Chapman and Hall. doi: 10.1201/9780429258589

American Speech-Language-Hearing Association. (2007). *Childhood Apraxia of Speech [Technical report]*. Available online at: http://www.asha.org/policy (accessed September 2, 2021)

Audacity Team (2021). *Audacity(R): Free audio editor and recorder. Version 2.4.2*. Available online at: https://audacityteam.org/ (accessed November 30, 2021)

Bartko, J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychol. Rep*. 19, 3–11. doi: 10.2466/pr0.1966.19.1.3

Brown, T., Murray, E., and McCabe, P. (2018). The boundaries of auditory perception for within-word syllable segregation in untrained and trained adult listeners. *Clin. Linguist. Phon*. 32, 979–996. doi: 10.1080/02699206.2018.1463395

Chakravarti, M., Laha, R., and Roy, J. (1967). *Handbook of Methods of Applied Statistics, Volume I*. Hoboken, NJ: John Wiley and Sons. 392–394.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas*. 20, 37–46. doi: 10.1177/001316446002000104

Cureton, E. (1956). Rank-biserial correlation. *Psychometrika* 21, 287–290. doi: 10.1007/BF02289138

Cychosz, M., Edwards, J., Munson, B., and Johnson, K. (2019). Spectral and temporal measures of coarticulation in child speech. *J. Accoust. Soc*. 146, 516–522. doi: 10.1121/1.5139201

Edeal, M., and Gildersleeve-Neumann, C. (2011). The importance of production frequency in therapy for childhood apraxia of speech. *Am. J. Speech Lang. Pathol*. 20, 95–110. doi: 10.1044/1058-0360(2011/09-0005)

Elangovan, S., Payne, N., Smurzynski, J., and Fagelson, A. (2016). Musical training influences auditory temporal processing. *J. Hear. Sci*. 6, 37–44. doi: 10.17430/901913

Fedorenko, E., Morgan, A., Murray, E., Cardinaux, A., Mei, C., Tager-Flusberg, H., et al. (2016). A highly penetrant form of childhood apraxia of speech due to deletion of 16p11.2. *Eur. J. Hum. Genet*. 24, 302–306. doi: 10.1038/ejhg.2015.149

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc*. 32, 675–701. doi: 10.1080/01621459.1937.10503522

Fritz, C., Morris, P., and Richler, J. (2011). Effect size estimates: current use, calculations, and interpretation. *Exp. Psychol. Gen*. 141, 2–18. doi: 10.1037/a0024338

Gerosa, M., Lee, S., Giuliani, D., and Narayanan, S. (2006). "Analyzing children's speech: an acoustic study of consonants and consonant-vowel transition [Conference presentation]," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings* (Toulouse).

Giannela-Samelli, A., and Schochat, E. (2008). The gaps-in-noise test: gap detection thresholds in normal-hearing young adults. *Int. J. Audiol*. 47, 238–245. doi: 10.1080/14992020801908244

Gierut, J., Morrisette, M., and Ziemer, S. (2010). Nonwords and generalisation in children with phonological disorders. *Am. J. Speech Lang. Pathol*. 19, 167–177. doi: 10.1044/1058-0360(2009/09-0020)

Harvey, C. (2016). Binary choice vs ratings scales: a behavioural science perspective. *Int. J. Mark. Res*. 58, 647–648. doi: 10.2501/IJMR-2016-041

Hearing Australia (2021). *Hearing Australia: Online Hearing Assessment*. Available online at: https://www.hearing.com.au/onlineassessment (accessed October 31, 2021)

Heldner, M. (2011). Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. *J. Acoust. Soc*. 130, 508–513. doi: 10.1121/1.3598457

Hildebrand, M., Jackson, V., Scerri, T., Reyk, O., Coleman, M., Braden, R., et al. (2020). Severe childhood speech disorder: gene discovery highlights transcriptional dysregulation. *Neurology* 94, 2148–2167. doi: 10.1212/WNL.0000000000009441

IBM Corp. (2020). *IBM SPSS Statistics for Windows, Version 27.0*. Armonk, NY: IBM Corp.

Koo, T., and Li, M. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med*. 15, 155–163. doi: 10.1016/j.jcm.2016.02.012

Landis, J., and Koch, D. (1977). The measurement of observer agreement for categorical data. *Biometrica* 33,159–164. doi: 10.2307/2529310

Lee, S., Potamianos, A., and Narayanan, S. (1999). Acoustics of children's speech: developmental changes of temporal and spectral parameters. *J. Acoust. Soc. Am*. 105, 1455–1468. doi: 10.1121/1.426686

Markham, D., and Hazan, V. (2004). The effect of talker- and listener-related factors on intelligibility for a real-word, open-set perception test. *J. Speech Lang. Hear Res*. 47, 725–737. doi: 10.1044/1092-4388(2004/055)

McCabe, P., Thomas, D., Murray, E., Crocco, L., and Madill, C. (2017). Rapid syllable transition treatment-ReST. Sydney, NSW: The University of Sydney.

McGill, N., Crowe, K., and Mcleod, S. (2020). "Many wasted months": Stakeholders' perspectives about waiting for speech-language pathology services. *Int. J. Speech. Lang. Pathol*. 22, 313–326. doi: 10.1080/17549507.2020.1747541

Mishra, S., Panda, M., and Herbert, C. (2014). Enhanced auditory temporal gap detection in listeners with musical training. *J. Acoust. Soc. Am*. 136, 173–178. doi: 10.1121/1.4890207

Murray, E., McCabe, P., Heard, R., and Ballard, K. (2015). Differential diagnosis of children with suspected childhood apraxia of speech. *J. Speech Lang. Hear. Res*. 58, 43–60. doi: 10.1044/2014_JSLHR-S-12-0358

Pichora-Fuller, M., Schneider, B., Benson, N., Hamstra, S., and Storzer, E. (2006). Effect of age on detection of gaps in speech and nonspeech markers varying in duration and spectral symmetry. *J. Accoust. Soc. Am*. 119, 1143–1155. doi: 10.1121/1.2149837

Qualtrics (2021). *Qualtrics Version 10/21*. Provo, UT: Qualtrics.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/ (accessed February 16, 2022)

Ratner, B. (2009). The correlation coefficient: its values range between +1/−1, or do they? *J. Target. Meas. Anal. Mark*. 17, 139–142. doi: 10.1057/jt.2009.5

Shahin, M., Zafar, U., and Ahmed, B. (2020). The automatic detection of speech disorders in children: challenges, opportunities, and preliminary results. *IEEE J. Sel. Top. Signal Process*. 14, 400–412. doi: 10.1109/JSTSP.2019.2959393

Shriberg, L., Lohmeier, H., Campbell, T., Dollaghan, C., Green, J., and Moore, C. (2009). A nonword repetition task for speakers with misarticulations: the syllable repetition task (SRT). *J. Speech Lang. Hear, Res*. 52, 1189–1212. doi: 10.1044/1092-4388(2009/08-0047)

Shriberg, L., Strand, E., Fourakis, M., Jakielski, K., Hall, S., Karlsson, H., et al. (2017). A diagnostic marker to discriminate childhood apraxia of speech from speech delay: I. development and description of the pause marker. *J. Speech Lang. Hear. Res*. 60, 1096–1117. doi: 10.1044/2016_JSLHR-S-16-0148

Snell, K., and Frisina, D. (2000). Relationships among age-related differences in gap detection and word recognition. *J. Acoust. Soc. Am*. 107, 1615–1626. doi: 10.1121/1.428446

Storkel, H. (2019). Using developmental norms for speech sounds as a means of determining treatment eligibility in schools. *Perspect. ASHA Special Interest Groups*. 4, 67–75. doi: 10.1044/2018_PERS-SIG1-2018-0014

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometr. Bull*. 1, 80–83. doi: 10.2307/3001968

Yoho, S., Borrie, S., Barret, T., and Whittaker, D. (2018). Are there sex effects for speech intelligibility in American English? Examining the influence of talker, listener and methodology. *Atten. Percept. Psychophys*. 81, 558–570. doi: 10.3758/s13414-018-1635-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 14 April 2022
doi: 10.3389/fcomm.2022.838955

Three Kinds of Rising-Falling Contours in German *wh*-Questions: Evidence From Form and Function

Katharina Zahner-Ritter[1,2]*, *Marieke Einfeldt*[2], *Daniela Wochner*[2], *Angela James*[2], *Nicole Dehé*[2] *and Bettina Braun*[2]

[1] *Phonetics, Department II, University of Trier, Trier, Germany,* [2] *Department of Linguistics, University of Konstanz, Konstanz, Germany*

The intonational realization of utterances is generally characterized by regional as well as inter- and intra-speaker variability in f0. Category boundaries thus remain "fuzzy" and it is non-trivial how the (continuous) acoustic space maps onto (discrete) pitch accent categories. We focus on three types of rising-falling contours, which differ in the alignment of L(ow) and H(igh) tones with respect to the stressed syllable. Most of the intonational systems on German have described two rising accent categories, e.g., L+H* and L*+H in the German ToBI system. L+H* has a high-pitched stressed syllable and a low leading tone aligned in the pre-tonic syllable; L*+H a low-pitched stressed syllable and a high trailing tone in the post-tonic syllable. There are indications for the existence of a third category which lies between these two categories, with both L and H aligned within the stressed syllable, henceforth termed (LH)*. In the present paper, we empirically investigate the distinctiveness of three rising-falling contours [L+H*, (LH)*, and L*+H, all with a subsequent low boundary tone] in German *wh*-questions. We employ an approach that addresses both the **form** and the **function** of the contours, also taking **regional variation** into account. In Experiment 1 (**form**), we used a delayed imitation paradigm to test whether Northern and Southern German speakers can imitate the three rising-falling contours in *wh*-questions as distinct contours. In Experiment 2 (**function**), we used a free association task to investigate whether listeners interpret the pragmatic meaning of the three contours differently. Imitation results showed that German speakers—both from the North and the South—reproduced the three contours. There was a small but significant effect of regional variety such that contours produced by speakers from the North were slightly more distinct than those by speakers from the South. In the association task, listeners from both varieties attributed distinct meanings to the (LH)* accent as opposed to the two ToBI accents L+H* and L*+H. Combined evidence from **form** and **function** suggests that three distinct contours can be found in the acoustic and perceptual space of German rising-falling contours.

Keywords: intonation, pitch accent, category, fuzziness, imitation, meaning, German

OPEN ACCESS

Edited by:
Eleanor Chodroff,
University of York, United Kingdom

Reviewed by:
Pärtel Lippus,
University of Tartu, Estonia
Oliver Niebuhr,
University of Southern Denmark, Denmark

****Correspondence:***
Katharina Zahner-Ritter
k.zahner-ritter@uni-trier.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *18 December 2021*
Accepted: *22 February 2022*
Published: *14 April 2022*

Citation:
Zahner-Ritter K, Einfeldt M, Wochner D, James A, Dehé N and Braun B (2022) Three Kinds of Rising-Falling Contours in German wh-Questions: Evidence From Form and Function. Front. Commun. 7:838955. doi: 10.3389/fcomm.2022.838955

INTRODUCTION

In spoken communication, speakers use intonation, primarily cued by f0, to mark sentence type (e.g., question vs. statement), information structure (e.g., focus and topic), information status (e.g., given vs. new information), hierarchical discourse structure, and attitudinal meaning (cf. Lehiste, 1975; Ladd, 2008; Prieto, 2015). The intonational realization of utterances is characterized by a lot of variability in f0—both within and across speakers (Atkinson, 1976; Gandour et al., 1991; Niebuhr et al., 2011; Grice et al., 2017), as well as across regional varieties, as shown, for instance, for German (Atterer and Ladd, 2004; Ulbrich, 2005; Braun, 2007; Mücke et al., 2009), or English (Grabe, 2004; Fletcher et al., 2005; Smith and Rathcke, 2020). Such variability includes, among others, the alignment of tonal targets [i.e., the position of low (L) and high (H) turning points with respect to the segmental string], their scaling (i.e., the tonal height), or the shape of intonational events (e.g., the slope or curvature). Category boundaries hence remain "fuzzy" and the question of whether and how the acoustic space can be split into distinct categories is non-trivial (cf. Arvaniti, 2019; Lohfink et al., 2019, for discussion). On the one hand, these categories need to do justice to the variability in the signal; on the other hand, they need to allow for generalizations. In the present paper, we contribute to this debate by addressing the distinctiveness of rising-falling f0 contours in German in an integrative approach that accounts for both **form** and **function**.

Previous research has demonstrated that nuclear intonation contours in German crucially differ with respect to f0-peak alignment (Kohler, 1991b; Grice et al., 2005; Niebuhr, 2022). The f0 peak may either precede the stressed syllable (H+L*, early-peak accent), or follow it (L*+H, late-peak accent), or be aligned within the stressed syllable (L+H*, medial-peak accent). While in H+L* the accentual movement falls onto the stressed syllable, L+H* and L*+H accents are considered rising accents, with the rising movement being perceptually very prominent (Baumann and Röhr, 2015; Baumann and Winter, 2018). In the present paper, we focus on the two rising accents L+H* and L*+H with a subsequent low boundary tone, along with a third rising-falling contour that lies between the two [henceforth (LH)*, which is our own descriptive label], see **Figure 1**. An earlier study has highlighted the potential existence of a third category between L+H* and L*+H in German (termed "late-medial peak," Kohler, 2005; cf. Niebuhr, 2022). The present study is designed to corroborate this preliminary evidence and sharpen the scope of the category, specifically with reference to rhetorical questions where (LH)* was recently observed in production (Braun et al., 2019).

Rising-falling contours in German have received different phonological representations in intonational phonology (Kohler, 1991a; Mayer, 1995; Grice et al., 2005; Peters, 2014). Most of these descriptions distinguish between two kinds of rising-falling contours (**Figures 1A,C**), transcribed as L+H* and L*+H in the German ToBI system (Grice et al., 2005). These accents have been related to differences in meaning: new information vs. self-evident information/information conflicting a speaker's belief (Kohler, 1991b; Grice and Baumann, 2002; Niebuhr, 2007b; Kügler and Gollrad, 2015) or attitudinal information, such as sarcasm (Lommel and Michalsky, 2017). Recent production data on German rhetorical questions (Braun et al., 2019) and verb-first exclamatives (Wochner, 2021) reveal another accent type which falls between the two more established ones (**Figure 1B**). In this contour, both the low and the high tonal target are realized within the stressed syllable (**Figure 1B**), similar to the late-medial peak reported in Kohler (2005, p. 90). This alignment pattern differentiates (LH)* from the more established accents L+H* and L*+H. Here, we take a fresh look at the acoustic and interpretative space of German rising-falling contours to discuss whether there is evidence to model three kinds of rising-falling contours: L+H*, L*+H, and (LH)*. To this end, we employ combined evidence from imitative productions (**form**) and judgments on the connotative meaning (**function**) to determine whether (LH)* is a pitch accent category on its own in German or, alternatively, whether (LH)* might be a variant of one of the two other pitch accents (L+H* or L*+H). If there are three distinct pitch accents in speakers' mental grammars, we expect three distinct contours in production (**form**, Experiment 1) and different connotative meaning attributions in perception (**function**, Experiment 2). The overall aim of our study is hence to probe the fuzziness in German rising-falling contours and discuss ways to model them appropriately.

In section "Background", we first provide background information on rising-falling contours in German in the different systems of intonational description, before we review approaches

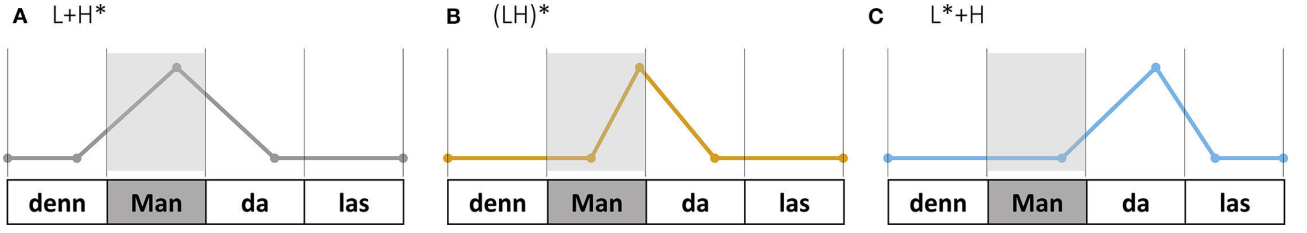

FIGURE 1 | Schematic representation of three rising-falling contours in German realized on a four-syllable sequence *denn Mandalas* "PRT Mandalas;" gray shading indicates the stressed syllable (tonic syllable) with which the pitch accent is associated. **(A–C)** show the three different alignment configurations analyzed in the present study.

attempting to model variability in intonational contours from a broader perspective. Section "The Present Study: Rationale and Hypotheses" outlines the rationale of our study and our hypotheses. In sections "Experiment 1: Delayed Imitation Study" and "Experiment 2: Paraphrasing of connotative question meaning", we present the two experiments before discussing the combined experimental results in section "General Discussion".

BACKGROUND

Rising-Falling Contours in German

According to the autosegmental-metrical (AM) theory of intonation (Arvaniti and Fletcher, 2020 for overview; Pierrehumbert, 1980; Ladd, 2008), pitch accents are represented by sequences of low and high tonal targets. In intonation languages, pitch accents are associated with the metrically stressed syllable, which is a lexical property of the word functioning as an anchor point for pitch accents (highlighted with gray shading in **Figure 1**). The actual alignment of the tonal targets with regard to the position in the stressed syllable varies, which results in different pitch accent types. Different models of German intonation, i.e., *German Tones and Break Indices* (GToBI, Grice et al., 2005), *The Kiel Intonation Model* (KIM, Kohler, 1991a; Niebuhr, 2022), *Intonationsgrammatik des nördlichen Standarddeutschen* "Intonation grammar of the Northern Standard German" (Peters, 2006, 2014) and *Transcription of German Intonation: The Stuttgart System* (STGTsystem, Mayer, 1995) have separated the space of possible pitch accents in rising-falling contours differently: The stylized realizations in **Figures 1A,C** are modeled after the tonal contrast in GToBI (Grice et al., 2005), which is widely used in research on German intonation and which is easily comparable to other ToBI systems in different languages. The realization depicted in **Figure 1B**, i.e., the contour found in rhetorical questions and exclamatives, (LH)*, does not occur in this system. The STGTsystem (Mayer, 1995), an alternative version of GToBI developed in Stuttgart, lists only one accent type for rising-falling contours (L*HL), whose phonetic description resembles the stylization in **Figure 1C** (L*+H in GToBI). Peters (2014, pp. 45–48) describes the contour in **Figure 1A** as a fall (H*L) and the contour in **Figure 1C** as L*H. KIM (Kohler, 1991a; Niebuhr, 2022) is a contour-based account and distinguishes so called "medial peaks" from "late peaks," whereby the medial peak can be projected onto H*/L+H* and the late peak onto L*+H in an AM framework (cf. Niebuhr and Ambrazaitis, 2006; Niebuhr, 2007b). Importantly, KIM added a contour to its system, based on subsequent research done on the model (Kohler, 2005; Niebuhr, 2022, for overview): In particular, Kohler (2005) used semantic scales to show that the peak alignment continuum contains an additional category that falls between the medial (L+H*) and late peak (L*+H)—a contour called "late-medial peak" in KIM and described as (LH)* in the present paper.

Recent production data further provide evidence for a consistent and meaningful use of (LH)* as a pitch accent signaling rhetorical illocution in *wh*-questions (Braun et al., 2019). This accent was specific to rhetorical questions and did not occur in information-seeking *wh*-questions (Dehé et al., 2022). Crucially, it did not only occur in contexts of tonal crowding, but also when post-tonic syllables were available. (LH)* has also been observed with verb-first exclamatives, while string-identical information-seeking questions were realized with a high-rising contour (Wochner, 2021). The occurrence of (LH)* in these (non-canonical) utterance types, as opposed to information-seeking questions, suggests that it may be phonemic, rather than a phonetic variant of another accent. At the same time, (LH)* has been described as an allophonic variant of the established accents described above, occurring in suboptimal segmental contexts, in which there are not enough syllables to realize the late-peak contour (cf. Mayer, 1995; Kügler, 2007; Peters, 2014), hence resembling the configuration in **Figure 1B**. In practice, (LH)* realizations caused difficulties in transcription because they share the alignment of the L tone with L*+H and that of the H tone with L+H*. Taken together, the status of (LH)* in German (i.e., whether it is phonetic or phonological) is by far not clear and it raises issues for the mapping between acoustic realization and phonological categories. The specific question of the present paper is how many distinct (meaningful) contours need to be modeled within the broad category of rising-falling contours in German.

A complicating factor for this question is that natural productions are not as clearly distinct as the stylizations in **Figure 1** may suggest, but are subject to variability both within and across speakers (Atkinson, 1976; Gandour et al., 1991; Niebuhr et al., 2011; Grice et al., 2017; Lohfink et al., 2019; Roessig et al., 2019; Roessig, 2021).[1] Clearly, such individual variation blurs the boundaries of intonational categories. Regional variety, which is one of the foci of the present paper, additionally pushes the notion of categories to its limits as distributions between categories might overlap (Atterer and Ladd, 2004; Grabe, 2004; Gilles, 2005; Peters, 2006; Braun, 2007; Mücke et al., 2009): Indeed, a main discriminating aspect of the above-cited models on German intonation are their geographical origins. On a north-south axis, KIM (Kohler, 1991a) is located farthest in the north, followed by the system developed by Jörg Peters in Oldenburg (Peters, 2006, 2014). The STGTsystem (Mayer, 1995), in turn, originates in the South of Germany (Stuttgart). GToBI is a collaborative approach developed at universities in Saarbrücken, Stuttgart, Munich, and Braunschweig (Grice et al., 2005, p. 62). It is possible that the apparent differences in intonation labels and pitch accent contrasts are in part influenced by differences in regional variety (cf. Gilles, 2005; Peters, 2006; Kügler, 2007).

In fact, there is experimental evidence that Southern German speakers produce pitch accents in declarative sentences differently from Northern German speakers, at least in prenuclear position: Atterer and Ladd (2004), for instance, reported that Southern German speakers (from Bavaria) aligned

[1] Niebuhr et al. (2011), for instance, showed that German speakers differ in the magnitude of the alignment with which they differentiate H* from H+L*, with a weaker alignment contrast being compensated by adjustments in contour shape. Grice et al. (2017) showed that speakers consistently use the phonetic cues *alignment* and *scaling* when differentiating focus types, however only for some of the speaker did these differences lead to a difference in the intonational event (H* vs. L+H*), see also Braun (2006) on similar findings for contrastive vs. non-contrastively used prenuclear accents.

prenuclear accentual rises significantly later than speakers from the North-West of Germany (cf. Braun, 2007; Mücke et al., 2008); in nuclear position, alignment differences went in the same direction but were not significant (Mücke et al., 2009). In his analysis of the tonal inventory of Swabian, an Alemannic variety in the South of Germany, Kügler (2007) shows that speakers predominantly produced L*+H accents in declarative sentences (see also Kügler, 2004). Distributional analyses of Northern German speakers (Kiel), in turn, reveal medial peaks to occur more frequently than late peaks (Peters et al., 2005). This suggests that the distribution frequency in tonal inventories might also differ across regions (cf. Fitzpatrick-Cole, 1999; Leemann, 2012, on Swiss German), such that the acoustic space of rising-falling contours in Southern German speakers is shifted toward the right end of the spectrum [recall that the southern STGTsystem (Mayer, 1995) only accounts for one rising-falling contour]. In the present paper, we directly compare speakers from two different regions (North vs. South) on the three-way tonal alignment contrast.

Modeling Intonational Categories

The question of how phonological representations—typically thought of as distinct categories—and phonetic modification—typically understood as a gradual change—interrelate has been an issue of on-going debate (e.g., Ohala, 1990; Niebuhr, 2007a; Pierrehumbert, 2016; Arvaniti, 2019; Barnes et al., 2021; Roessig, 2021). It is uncontroversial that some form of generalization is necessary to systematize interfaces with other core areas, such as semantics or pragmatics. At the same time, clear-cut boundaries cannot be maintained given the variability in the speech signal and the fuzziness of the mapping between acoustic **form** and phonological category.

In intonational research, different tasks have been employed to study the relation between the continuous signal and intonational categories (cf. Prieto, 2012 for overview): Focusing on **intonational form**, identification and discrimination tasks have been used in classic categorical perception paradigms (Kohler, 1987, 1991b; Ladd and Morton, 1997; Schneider and Lintfert, 2003; Niebuhr, 2007b). Kohler (1991b), whose work is directly related to our question, showed categorical perception for early vs. medial peaks (i.e., H+L* vs. L+H*), two accents that differ in the direction of the accentual movement. The difference between the two rising-falling contours (medial vs. late peaks, i.e., L+H* vs. L*+H), in turn, was less clear-cut. Categorical perception results were similar for speakers from Northern and Southern Germany (Kiel vs. Munich, Kohler, 1991b, p. 149ff.). Beyond tonal alignment, the shape of the contour also seems to influence the categorical perception of rising-falling contours, leading to a more or less clear-cut perception between L+H* and L*+H (Niebuhr, 2007a, for effects of peak shape and intensity transitions); see also Barnes et al. (2021) for a study corroborating the relevance of the shape of the interpolation between L and H for the distinction between L+H* and L*+H in English. Another paradigm testing the distinctiveness in intonational form is imitation, which is based on the idea of a perception-production loop (e.g., Pierrehumbert and Steele, 1989; Braun et al., 2006; Dilley and Brown, 2007; Dilley, 2010; Chodroff and Cole, 2019b; Petrone et al., 2021). In imitation tasks, participants are typically presented with one stimulus at a time and have to imitate it. The productions are analyzed in terms of the overlap (or non-overlap) in the distributions of relevant parameters (such as tonal alignment) or overall shape. In imitation, task difficulty or working memory seem to affect outcome patterns: Braun et al. (2006), for instance, employed an iterative imitation paradigm in which speakers first imitated a set of randomly generated f0 tracks and then iteratively repeated their previous productions. They showed that speakers retained some detail in immediate imitation, which was lost, however, over successive repetitions. The authors argue for attractors in the perceptual space of intonation that function as a perceptual magnet. Participants in Chodroff and Cole (2019b), American English speakers, had to imitate one of eight nuclear contours and transfer the respective contour to a novel sentence with the same rhythmic structure, hence making generalization necessary. In their study, speakers primarily maintained the distinction between rising and falling contours, similar to the attractor contours in Braun et al. (2006). Petrone et al. (2021) showed that when working memory capacity is smaller, speakers have difficulties in reproducing contours correctly: Specifically, speakers with high working memory capacity were more accurate in the imitation of phonological events, both for obligatory events (pitch accents and boundary tones) and optional events. In sum, the harder the imitation task (due to either task demands or cognitive capacities), the smaller the set of reproduced contours. A challenging imitation task hence seems to us an appropriate method for the question of whether there are three distinct rising-falling contours.

Studies that have addressed the **functional distinction** between intonational contours have employed semantic scales (e.g., Dombrowski, 2003; Kohler, 2005; Dombrowski and Niebuhr, 2010; Kügler and Gollrad, 2015; Wochner, 2021), free association tasks (Kohler, 1991b), acceptability judgment tasks (Baumann and Grice, 2006), or psycholinguistic methods such as eye-tracking (e.g., Braun and Biezma, 2019). Kügler and Gollrad (2015), for instance, showed that German listeners differentiated between a contrastive and a broad focus reading based on differences in the scaling of the H tone (the L tone did not affect perceptual ratings, but see Ritter and Grice, 2015). Based on a free association task, Kohler (1991b) reports that medial peaks were associated with information that was new to the discourse in declaratives and with an information-seeking notion in questions. Late peaks also signaled new information (similar to medial peaks) but also added attitudinal meanings, such as astonishment or self-evidence (see also Grice et al., 2005; Lommel and Michalsky, 2017). Kohler (2005) corroborated these findings using semantic differentials; the contour that falls between the medial and late peak, the late-medial peak, tended to be associated with unexpectedness or surprise. Braun and Biezma (2019) used an eye-tracking paradigm to investigate the contrastive nature of nuclear L+H*, prenuclear L+H*, and prenuclear L*+H. They showed that listeners interpreted prenuclear L*+H and nuclear L+H* contrastively (more fixations to a referent that contrasted with the accented word) as opposed to prenuclear L+H*. Here, we use a combined approach of **form** and **function** to understand the sources of fuzziness surrounding rising-falling accents in German and to model it successfully.

THE PRESENT STUDY: RATIONALE AND HYPOTHESES

To test the distinctiveness of the three kinds of nuclear rising-falling contours in German [L+H*, L*+H, (LH)*; cf. **Figure 1**], we employ *wh*-questions and make use of two tasks: a delayed imitation task and a free association task. The delayed imitation task requires a kind of storage (beyond access to echoic memory), and hence taps into phonological representations of intonational contours. The working memory model by Baddeley and Hitch (1974) assumes that acoustic information decays after ∼2 s (phonological short-term memory, cf. Plomp, 1964; Gathercole et al., 1997), unless it is refreshed by a sub-vocal articulatory rehearsal process (Baddeley and Hitch, 1974; Baddeley, 1986, 2003). Moreover, Crowder (1982) reports that in terms of discrimination accuracy for vowel formants "[t]he auditory memory loss seems to be asymptotic at about 3 s" (Crowder, 1982, p. 197). Hence, there seems to be a threshold of about 2 to maximally 3 s up to which acoustic information is readily available and after which acoustic information decays. Based on this threshold, we designed our delayed imitation task with a 2,000 ms delay and a following sine tone with a duration of 500 ms. The free association task seems to be the best-suited paradigm for our study since the functional scope of (LH)* is not clear yet, which makes it hard to establish pre-defined connotative meanings required in other tasks.

For both experiments, speakers from Southern and Northern Germany were recruited in order to investigate regional variation (Mayer, 1995; Atterer and Ladd, 2004; Ulbrich, 2005; Braun, 2007; Kügler, 2007; Mücke et al., 2009). Speakers were allocated to either the Northern or the Southern German group according to where they were born and grew up. The Northern German group comprised speakers north of the Benrath line, an isogloss separating Low German and High German dialects (based on the High German consonant shift). The Southern German group comprised speakers from Baden-Wuerttemberg and Bavaria (south of the Speyer line, an isogloss that additionally separates Upper German dialects from Central German dialects), cf. Waterman (1991/1966).

In **Experiment 1 (form)** participants imitate three resynthesized nuclear rising-falling contours on *wh*-questions [L+H*, (LH)*, and L*+H] in a delayed imitation paradigm addressing phonological processing (cf. Baddeley and Hitch, 1974; Crowder, 1982; Baddeley, 1986, 2003). Methodologically, we input a three-way alignment contrast, and analyze the productions of Experiment 1 holistically, using general additive mixed models (GAMMs, Wood, 2006, 2017) on time-normalized utterances. This method allows us to capture the f0 contours as a whole and compare when in time two contours differ from each other significantly (cf. Wieling, 2018; van Rij et al., 2019; Sóskuthy, 2021). Using GAMMs hence not only provides information about tonal alignment (Atterer and Ladd, 2004), but also about tonal onglides (Ritter and Grice, 2015; Roessig et al., 2019), f0 excursions and scaling, and the overall shape of the contour (Niebuhr, 2007b; Niebuhr et al., 2011; Barnes et al., 2012, 2013, 2021). GAMMs furthermore allow us to test for interactions between intonation condition and regional variety over time, hence informing us on whether regional variation affects the distinctions between contours differently. In that sense, GAMMs represents an ideal statistical technique to disambiguate the fuzzy data patterns existent in f0 contours in order to unravel the meaningful underlying structure of intonational phonology.

Participants' imitative productions will be informative on how many **distinct contours** we need to model in the acoustic space of German rising-falling contours: Three distinct rising-falling contours in the imitative productions of the speakers will provide evidence for (LH)* as a third kind of rising-falling contour in German next to the two more established L+H* and L*+H contours, hence corroborating the three-way-contrast initially laid out in Kohler (2005) and also observed in Braun et al. (2019). Given that our delayed imitation task requires storage of the contours, the evidence would go beyond phonetic details and clearly speak in favor of phonological processing. If, on the other hand, speakers reproduce two contours in their imitative productions, this will provide evidence in favor of collapsing the range of rising-falling contours into two contours (cf. Braun et al., 2006, on English; Chodroff and Cole, 2019b). Reproduction of only one contour would suggest that the task is too hard (since there is plenty of independent evidence in favor of two rising-falling contours in German, see sections "Introduction" and "Background"). With respect to **regional variation**—although a direct comparison of studies reporting occurrence frequency is difficult, medial-peak contours (H*/L+H*) have been shown to be more frequent than late-peak contours (L*+H) for Northern German speakers (Peters et al., 2005). For Southern German speakers, in turn, rising accents with a late L and H alignment have been reported to occur frequently (described as L*+H in Kügler, 2004; see also Truckenbrodt, 2007, for the prenuclear position). Based on these differences in occurrence frequency, it is conceivable that L+H* functions as a perceptual attractor (magnet) for Northern German speakers, while L*+H serves this function for Southern German speakers, along the lines of what is known on magnets on the segmental level (Anderson et al., 2003; cf. Braun et al. (2006) and Roessig et al. (2019) for attractor-based accounts of intonation). Given that (LH)* may be less strongly anchored in the intonational grammar due to its more restricted function and hence less frequent occurrence, it may be yet more prone to merger effects (cf. Braun et al., 2006). Under this assumption, we predict the distinction of contours to differ between regions, with a smaller distinction between L+H* and (LH)* in the North than in the South [L+H* as merger with (LH)*], and conversely, a smaller distinction between L*+H and (LH)* in the South than in the North [L*+H as merger with (LH)*].

Experiment 2 (**function**) tests whether the three rising-falling contours in *wh*-questions are interpreted differently. To this end, we conducted a qualitative study in which participants, different from the ones in Experiment 1, paraphrased the connotative meaning of the stimuli in Experiment 1 in their own words. The paraphrases of Experiment 2 were recoded into superordinate categories and analyzed using conditional inference trees (CTrees, Hothorn et al., 2006). This method allows us to test whether and how intonation condition and regional background affect participants' responses. We predict that if there are in fact three distinct contours, they will lead to different interpretations: Drawing on the available literature, we expect that L+H* leads to descriptions relating to an information-seeking nature (Kohler, 1991b; Baumann and Grice, 2006; Braun et al., 2019), while L*+H is expected to trigger descriptions related to contrast and/or attitudinal meanings (Grice et al., 2005; Niebuhr, 2007b; Lommel and Michalsky, 2017); (LH)* is hypothesized to be interpreted as rhetorical (Braun et al., 2019), or to signal surprise or obviousness (Wochner, 2021), or unexpectedness (Kohler, 2005). In terms of regional variation, we cannot make strong predictions regarding meaning—recall that Kohler (1991b, p. 149ff.) showed constant semantic judgments for different contours across Northern and Southern German listeners. If anything, we expect a merger effect in terms of meaning for the most frequent accent type (L+H* in Northern and L*+H in Southern German speakers).

EXPERIMENT 1: DELAYED IMITATION STUDY

Methods

Participants

In total, 28 monolingual native German participants, half from Northern Germany (mean age = 25.7 years, SD = 5.0 years, 10 female, 4 male) and half from Southern Germany (mean age = 25.5 years, SD = 4.4 years, 1 diverse, 8 female, 5 male), who had not learned a second language before the age of six, took part in the imitation study. Speakers from the Southern German group spent most of their lives in Baden-Wuerttemberg (*N* = 14), while speakers in the Northern German group came from Berlin (*N* = 1), Brandenburg (*N* = 1), Hamburg (*N* = 1), Mecklenburg-Vorpommern (*N* = 1), Lower Saxony (*N* = 3), North Rhine-Westphalia (*N* = 2), and Schleswig-Holstein (*N* = 5), all north of the Benrath Line. Due to restrictions imposed by COVID-19, testing started in the lab for eight Southern German speakers and then was continued via the online platform *SosciSurvey* (https://www.soscisurvey.de, Leiner, 2018) for all Northern speakers and the six remaining Southern German speakers.

Materials

Four target *wh*-questions were constructed that consisted of the *wh*-word *wer* "who," a monosyllabic verb, the particle *denn*, and a trisyllabic object noun with initial stress, see (1).

(1)

a. *Wer heißt denn Melanie?* [ˈmɛ.la.ni] ("Whose name is Melanie?")

b. *Wer spielt denn Libero?* [ˈliː.bə.ʁo] ("Who plays sweeper?")

c. *Wer malt denn Mandalas?* [ˈman.da.las] ("Who draws/colors mandalas?")

d. *Wer trinkt denn Malibu?* [ˈmaː.li.bu] ("Who drinks Malibu cocktails?")

Nouns with two post-tonic syllables were chosen to avoid tonal crowding (Prieto, 2011; Hanssen, 2017; Rathcke, 2017); also, their segments were as sonorous as possible, especially in the first two syllables, to ease f0 analysis. The propositions of the questions were chosen so that they did not elicit strong (positive or negative) feelings but were mainly perceived as neutral[2]. The four questions were recorded by a female native speaker from Northern Germany (31 years at time of recording), who grew up with a Southern German parent and who is familiar with intonational phonology. She produced the *wh*-questions in two conditions: (i) with a nuclear L+H* accent and (ii) with a nuclear L*+H accent (see **Supplementary Material S1** for acoustic analysis). She was instructed to focus on the alignment of the tonal targets. The recordings were then manipulated in three steps (splicing, duration manipulation, f0 manipulation) using *Praat* (Boersma and Weenink, 2016). First, splicing ensured that pitch accent realizations were not affected differently by the preceding part of the *wh*-question. To this end, for each item, the auditorily best "precontext" (*wh-word, verb, particle*) was selected. Likewise, the best productions of the object nouns (one for L+H*, one for L*+H for each item) were selected and cut at positive zero-crossings. Both parts were scaled to 63 dB. To reduce variability across items, the precontexts were manipulated in terms of duration using PSOLA resynthesis. This way, the constituents had an equal average duration for each item (in the three intonation conditions). The same was true for the three syllables of the noun. Second, the precontexts were cross-spliced to the nouns. Finally, the alignment of the tonal targets of the noun (L1: start of the f0 rise, H: f0 peak, L2: end of the f0 fall) was manipulated, based on the alignment of the naturally recorded stimuli for L+H* and L*+H and the values reported in Braun et al. (2019) for (LH)*, see **Table 1**. **Table 1** shows the locations of the three tonal targets (L1, H, L2) within the rising-falling contour (in the particle *denn,* the first, second or third syllable in the object noun). Percentages refer to the total duration of the respective unit, e.g., the f0 peak (H) occurred after 71% of first syllable of the noun in L+H*, and after 94% in (LH)*; for L*+H, it occurred after 71% of the second syllable of the noun. Note that **Figure 1** shows a visual representation of **Table 1**. The f0 values in the rising-falling contours were set at 166 Hz for L1, at 273 Hz for H, and at 170 Hz for L2 (based on the mean values in natural productions), leading to a pitch range of 8.6 semitones (st) for the rising part and 8.2 st for the falling part of the contour.

[2] Materials underwent a check in which the propositions of eight questions (four of which were the selected target questions and four of which were filler questions) were presented to 15 listeners (native speakers of German, mostly student assistants) who judged each proposition [e.g., *drinking Malibu* (target item), *going to the cinema* (filler) etc.] as either positive (*I like*), negative (*I don't like*), or neutral (*I don't have an opinion*). The propositions of the selected questions were predominantly judged as neutral (68%); in 27% participants had a positive attitude and in 5% of the cases a negative attitude toward the proposition of the sentence.

TABLE 1 | Alignment of tonal targets (L1, H, L2) in rising-falling contours in experimental stimuli; L+H* and L*+H values based on natural recordings, (LH)* values based on Braun et al. (2019).

	L1	H	L2
L+H*	In [n] from *denn* (22.0%)	In syllable 1 of noun (71.2%)	In syllable 2 of noun (69.6%)
(LH)*	In syllable 1 of noun (45.4%)	In syllable 1 of noun (94.4%)	In syllable 2 of noun (69.6%)
L*+H	In syllable 1 of noun (76.3%)	In syllable 2 of noun (71.1%)	In syllable 3 of noun (31.6%)

Percentages refer to the total duration of the respective unit.

The f0 contours of naturally produced L+H* accents and naturally produced L*+H accents (4 items each) were resynthesized into the three intonation conditions, leading to a total of 24 test stimuli (4 items × 3 target contours × 2 manipulation origins). We used two manipulation origins to exclude the possibility that spectral effects could have affected the imitations, which was not the case (see below). Natural recordings of (LH)* were avoided because this contour might be realized with breathy voice in *wh*-questions (Braun et al., 2019), which might be a confounding cue. Furthermore, recording the contours at the ends of the continuum (**Figures 1A,C**) will allow us to resynthesize intermediate steps in future studies. In the present study, we start with three contours, the two more established L+H* vs. L*+H, and one intermediate contour (LH)*. We further selected four additional *wh*-questions to be used as practice trials. They had the same syntactic structure but different target words (*Thymian* "thyme", *Komiker* "comedian", *Kolibris* "hummingbirds", *Tombolas* "tombolas"). These questions were resynthesized into the more established accents L+H* and L*+H.

There were two experimental lists with a pseudo-randomized order of trials to avoid priming of contours between trials. Lists did not contain sequences with the same item or the same contour in a row. The second experimental list was a mirror list of the first list such that the first trial in list 1 was the last in list 2. This was done to avoid order effects. Experimental lists were randomly assigned to the participants. Prior to the 24 experimental trials, there were four practice trials to familiarize participants with the procedure and voice of the speaker.

Procedure

Each trial was initiated by a sine tone (at 300 Hz, 500 ms duration) to signal the beginning of the trial. Participants listened to the questions via headphones. Each target question was also orthographically displayed on screen. Participants were instructed to imitate the utterances as closely as possible with a special focus on their speech melody. They were told to choose a pitch register that appeared suitable for them. This was done to avoid a mimicry of pitch and vocal characteristics of the speaker.

Each utterance was played only once, followed by a 2,000 ms period of silence and a sine tone of 500 ms (presented pseudo-randomly at 450 or 150 Hz) before participants started to imitate the question. The sine frequencies meet the floor and ceiling register frequencies of the speaker who produced the stimuli; the sine tones were played to overwrite any acoustic trace that might be kept after the 2,000 ms silence. We used two different frequencies for the sine tone, in random order, so that participants could not anticipate and adapt to it. After participants had imitated the respective utterance, they pressed a key to proceed to the next trial. Recordings were done via the microphone of the participants' computers in the remote setting. In the lab setting, recordings were done with a head-set microphone (DPA 4088F) onto a MacBookPro in a sound-attenuated booth.

TABLE 2 | Overview of imitated productions per group in final dataset of Experiment 1.

	Northern German speakers	Southern German speakers
L+H* contour	108	112
(LH)* contour	111	111
L*+H contour	110	111
Sum	**329**	**334**
Total		**663**

Data Processing and Statistical Analysis

Dataset

In total, we collected 672 sound files (28 participants × 24 imitated questions). Note that each sound file has one imitation. Nine files were excluded due to mispronunciations, bad sound quality, or a technical error on the online platform that led to data loss. The final data set for the analysis consisted of 663 sound files, see **Table 2** for a breakdown of the distribution of files across different groups.

Data Processing

Sound files were first segmented semi-automatically using the software *Web Maus* (Kisler et al., 2017) with boundaries being manually adjusted according to standard segmentation criteria (Turk et al., 2006). The critical segments were [n] in the particle *denn,* and the first, second, and third syllables of the object nouns. Subsequently, f0 values were extracted for these four intervals using *Prosody Pro* (Xu, 2013).[3] **Figure 2** shows an actual imitation of one target question in the three conditions, along with the annotation. We used 50 measurements per

[3] To reduce erroneous f0 values, the Hz range for higher pitched voices (mostly persons that identified as female) was set to 100–500 Hz, while it was changed to 50–300 Hz for lower pitched voices (persons that identified as male).

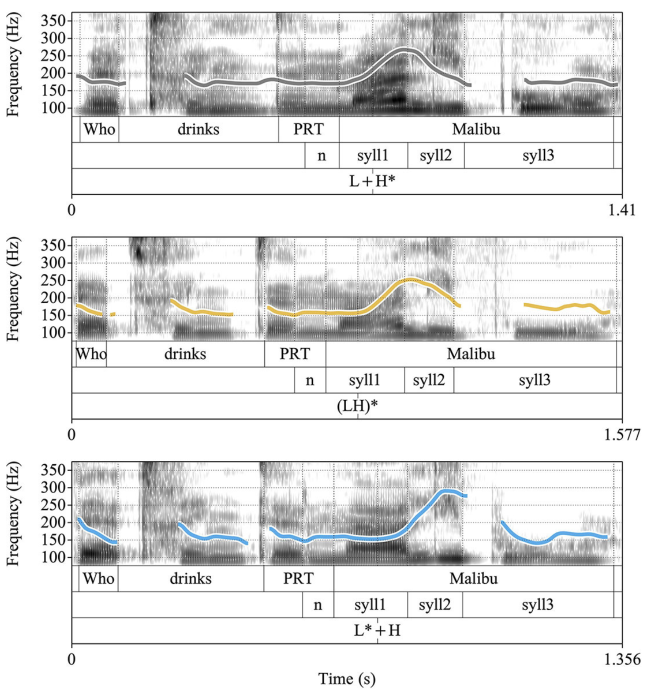

FIGURE 2 | Imitative productions of the German target question *Wer trinkt denn Malibu*? 'Who drinks Malibu?' in the three intonation conditions (vp19, Northern German group, female, 24 years). Top panel: L+H*, mid panel: (LH)*, bottom panel: L*+H. Tier 2 served as input tier for the extraction of f0 values; all other tiers are for illustration purposes only.

time interval. To detect and remove f0-tracking errors, which primarily occurred in word-final fricatives and word-medial stops, we used a custom-made algorithm in Python that replaced likely octave jumps by "NA" so that they were excluded from the analysis. In R (R Development Core Team, 2015), raw f0 values were transformed into semitones to ease interpretation of (perceptible) differences across contours and downsampled (10 values per interval for statistical analysis).[4]

Statistical Analysis

We used GAMMs (Wood, 2006, 2017) to test the distinctiveness of the three different intonational contours. GAMMs were chosen as they allow for a direct comparison between f0 contours by modeling non-linear dependencies of a response variable (here f0) and different predictors (here *intonation condition* and *region*) over time via smooth functions. They use a pre-specified number of base functions of different shapes (Baayen et al., 2018; Wieling, 2018; van Rij et al., 2019; Sóskuthy, 2021). GAMMs also allow us to model interactions over time (e.g., *condition* × *region*), which test whether the distinctiveness of contours differs between speakers of Northern and Southern German (cf. van Rij et al., 2019, p. 8ff.; Wieling, 2018, p. 106ff.). For the model fitting of the GAMMs, we used the R package *mgcv* (Wood, 2011, 2017); the package *itsadug* was used to plot the model results (van Rij et al., 2017), which is essential to interpret model outputs.

The response variable was the f0 value (in st) at different time points (10 values per interval), which was roughly normally distributed. All models were corrected for autocorrelation in the f0 data using an autocorrelation parameter rho, determined by the acf_resid()-function from the package *itsadug* (van Rij et al., 2017).[5] Models were initially fitted using the maximum likelihood (ML) estimation method in order to be able to compare models with different complexity (Sóskuthy, 2021, p. 16; Wieling, 2018, p. 89). We first tested whether the modeling of different curves for the three intonation conditions over time is warranted. Since this was the case, we then assessed the interaction between *intonation condition* and *region* (see below for details). Model fits were checked using gam.check() and the number of base functions (k) was adjusted if necessary. Also, models were re-run with the scaled *t* distribution (family = "scat", closely following the suggestion in van Rij et al., 2019, p. 17) due to tailed residuals. All steps of the analyses can be found in the **Supplementary Materials** to this paper (http://doi.org/10.17632/yhv7nmjmgf.2).

Results

Figure 3 shows the raw data, i.e., the average f0 contours on time-normalized utterances (in st) of imitated productions in the different intonation conditions for Northern German (left) and Southern German speakers (right). Data from both manipulation origins (i.e., whether the contours were resynthesized from L+H* or L*+H) were collapsed since the resynthesis procedure did not affect the realization of contours (see analysis on Mendeley for details). Note that syllable durations in the imitated questions did not differ across intonation conditions (all $p > 0.12$).

The initial GAMM included *condition* and *region* as parametric effects along with a smooth for the interaction of *intonation condition* over (normalized) time, s(Normtime, by = *intonation condition*), and factor smooths for *participants* and *items.* Model comparisons using the function compareML() revealed that this model was superior to a simpler model without the smooth for condition over time [$\chi^2_{(4.00)} = 1323.09$, $p < 0.0001$], corroborating the existence of different contours. We then assessed the interaction between *intonation condition* and *region* over time to test whether the distinction of contours differed across regions. To do so, we refitted the model including an interaction variable *RegCond* (6 levels, 2 regions × 3 intonation conditions). The interaction model had a better fit [$\chi^2_{(8.00)} = 91.40$, $p < 0.001$], indicating that the speakers from the North made

[4] For semitone conversion, the following formula was used: st = 12*log2(f0/f0ref). Based on visual inspection of the distribution of f0 values per gender, f0ref was set to 175 Hz for higher pitched voices (mostly persons that identified as female) and 100 Hz for lower pitched voices (persons that identified as male), resulting in mostly positive semitone values (mean = 2.6 st, sd = 3.0 st).

[5] Autocorrelation can also be reduced by fitting smooths for an event variable (i.e., a unique time series for each subject on each trial), cf. van Rij et al. (2019). However, this was computationally not possible. Instead, we fitted factor smooths for subject and item and controlled for autocorrelation.

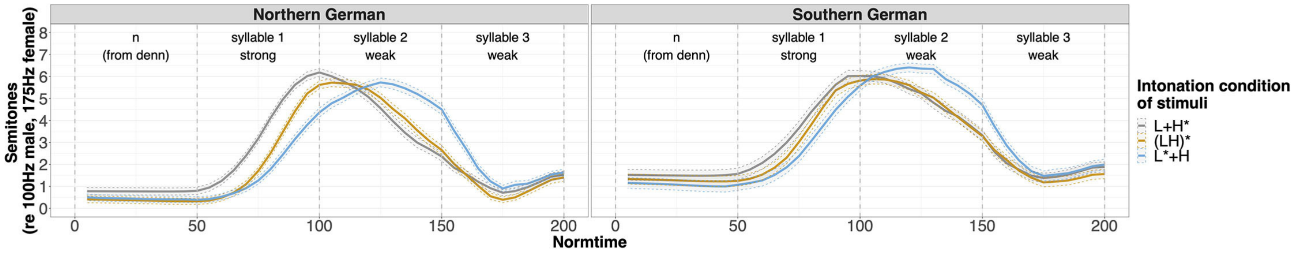

FIGURE 3 | Average f0 contours (in st) of imitative productions in the three different intonation conditions [L+H* in gray, (LH)* in orange, and L*+H in blue], for Northern German (left) and Southern German speakers (right). The x-axis displays the time-normalized questions (from [n] of the particle denn followed by the trisyllabic sentence-final object, e.g., *Mandalas*).

different distinctions than speakers from the South. The best-fitting model was re-run with the scaled *t* distribution specified (family = "scat"). Based on this new model, we again determined a value for the rho parameter to account for autocorrelation. The outcome of this final model[6] is visualized in **Figure 4**. It explained 68.5% of the deviance, see **Supplementary Material S2.1** for details on model evaluation in terms of residuals.[7]

Figure 4A shows the f0 contours in the three intonation conditions as predicted by the final GAMM, split by *region* (Northern German speakers are shown in the left and Southern German speakers in the right panel). The difference between two contours is directly displayed in so called difference curves, where f0 values in one condition are subtracted from f0 values in the other condition, see **Figure 4B** for the three pairwise contour comparisons for the two regions:

- **Comparison L+H* vs. (LH)*.** For speakers from Northern Germany, the imitative productions significantly differed already in the pre-tonic interval (the [n] of the particle *denn*), with L+H* contours being slightly higher than (LH)* contours. L+H* furthermore had an earlier low turning point than (LH)* in Northern German speakers (i.e., an earlier start of the rise), which accounted for a large difference in the stressed syllable. For speakers from Southern Germany, by contrast, the difference in alignment of the low turning point was less obvious and contours only differed for a small part of syllable 1 of the noun. The difference was around half a semitone, which is very subtle (Batliner, 1989). Similarly, a later alignment of the peak in (LH)* than in L+H* led to differences in the contour at the end of syllable 1 and the beginning of syllable 2 in the noun in the Northern German group. This difference was again less pronounced in the Southern German group. Overall, the distinction between L+H* and (LH)* seems to be larger for Northern German than for Southern German speakers (see **Supplementary Material S.2.2** for interaction model). Importantly though, both speaker groups distinguished between the two contours.
- **Comparison (LH)* vs. L*+H.** For both speaker groups, the two contours differed significantly in syllables 1 and 2 of the noun. The alignment of the low turning point was comparable but the (LH)* had a steeper rise with an earlier peak than the L*+H accent, which led to considerable differences in the stressed syllable and the post-tonic syllable. **Supplementary Material S.2.2** show that regional differences are minor for this distinction.
- **Comparison L+H* vs. L*+H.** As expected, the imitative productions of the two established accents differed significantly in syllables 1 and 2 of the noun, with a later alignment of the low turning point and the peak for L*+H as compared to L+H*. The f0 of the L+H* hence rose earlier than for L*+H, leading to strong differences in realization. Interestingly, the contours deviated already in the pre-tonic syllable ([n] of the particle *denn*), similarly in both varieties. The distinction between contours was generally more pronounced for Northern German speakers than for Southern German speakers (see **Supplementary Material S.2.2** for interaction model).

Taken together, the two more established accent types L+H* vs. L*+H are clearly distinct in their **form**, showing significant differences across both syllable 1 and 2 in both regional varieties (the difference was larger for Northern German speakers). Crucially though, the f0 contour in the (LH)* condition is also significantly different from the f0 contour in L+H* (syllable 1) and from the f0 contour in L*+H (syllables 1 and 2)—for both speaker groups, but the distinction between L+H* and (LH)* is smaller for speakers from the South than for speakers from the North; in fact, these two contours are very similar in the productions of Southern German participants. The distinction between L*+H and (LH)* is more pronounced

[6] The final model was specified as follows: model <- bam(st ∼ Regcond + s(Normtime, by = Regcond, k = 20) + s(Normtime, vp_index, bs = "fs", m = 1) + s(Normtime, manip_item, bs = "fs", m = 1), data = delayed_imitation_sorted, family = "scat", discrete = T, method = "fREML", rho = rhoval, AR.start = delayed_imitation_sorted$start_event).

[7] To corroborate the interaction between *condition* and *region* indicated in our final model, we constructed additional models containing binary difference smooths terms that capture the difference of the difference over time between two predictors, and hence their interaction (closely following the procedure described in van Rij et al., 2019, pp. 11–13; Wieling, 2018, p. 109 ff.), see **Supplementary Material S.2.2** for details. Again, these binary difference smooths support the interpretation that the difference between intonation conditions over time is different for speakers from Northern vs. Southern Germany but also, that speakers from both regions produced three distinct contours, **Supplementary Material S.2.2**.

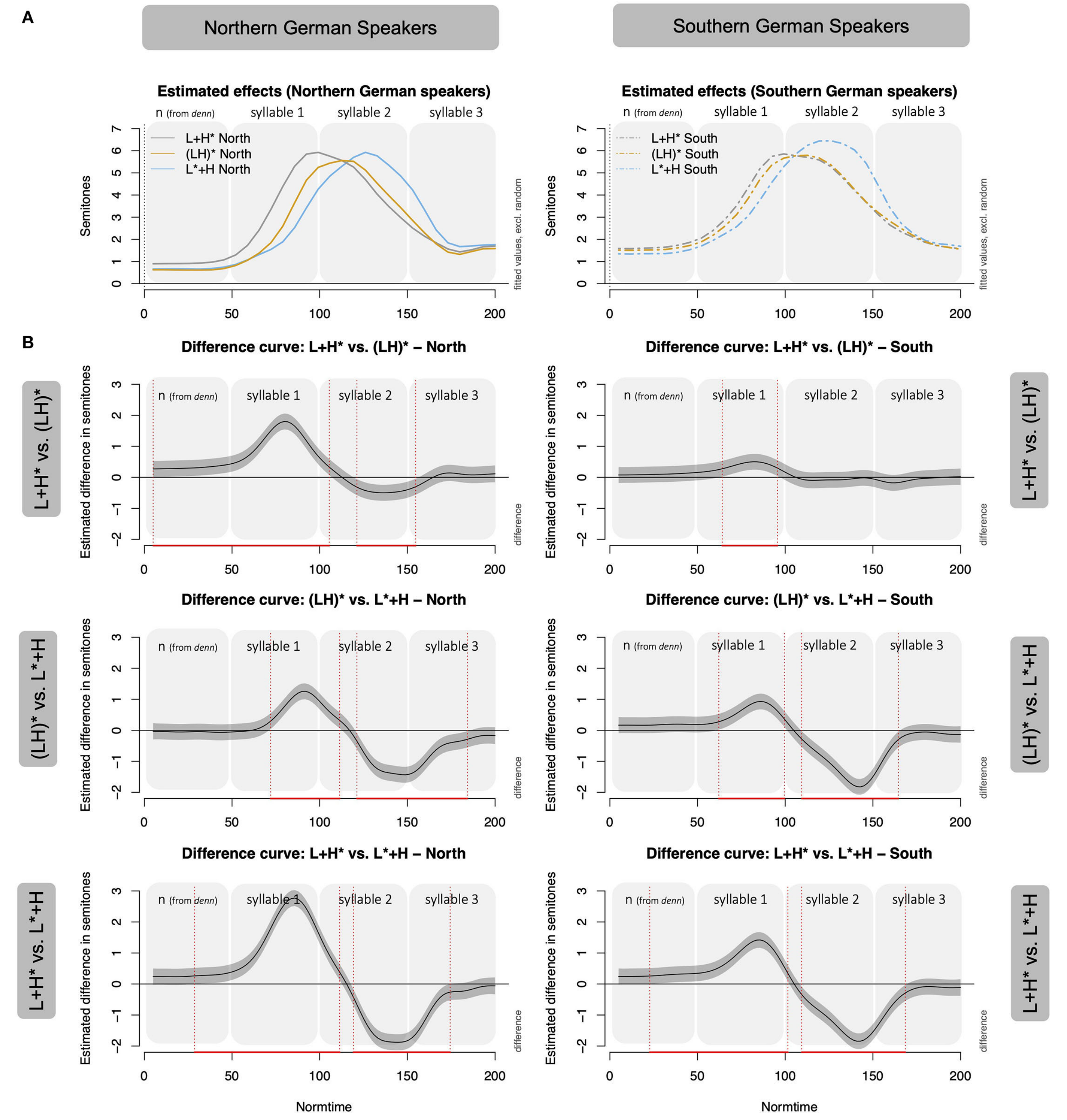

FIGURE 4 | GAMM results. **(A)** Predicted f0 values in the three intonation conditions [L+H* in gray, (LH)* in orange, and L*+H in blue] for Northern German speakers (left panel) and Southern German speakers (right panel). **(B)** Predicted difference curves (pairwise comparisons between contours), for Northern German speakers (left panel) and Southern German speakers (right panel). The gray shading displays the 95% CI (confidence interval) of the predicted mean difference. The difference becomes significant if zero is not included in the 95% CI. This is marked by the vertical red lines.

and clearly maintained in both regions, with differences in contours occurring both in syllables 1 and 2 of the object noun. The significant differences in contours across all pairwise comparisons suggest that speakers maintain a three-way contrast in rising-falling-contours, with regional variety modulating the extent of the distinction.

Interim Discussion

Experiment 1 tested the distinctiveness in realization of three rising-falling contours in Northern and Southern German speakers in a delayed imitation task that addressed phonological processing (Chodroff and Cole, 2019b; Petrone et al., 2021). Our findings show that all three contours are distinguished from each other. However, it is not unambiguously clear at this point whether participants imitated the tonal targets (i.e., pitch accent categories), a communicative function associated with the different contours (e.g., information-seeking vs. rhetorical question), or differences in perceived prominence. In prominence perception tasks, steeper slopes are judged more prominent than shallower ones (Rietveld and Gussenhoven, 1985; Baumann and Röhr, 2015; Baumann and Winter, 2018). Clearly, imitative productions had to be retrieved from stored representations, which may consist of aspects of tonal alignment properties, prominence, and meaning—possibly also a combination of the three.

Regional variety mediated the extent of the distinction between contours such that speakers from Northern Germany had more distinct productions than speakers from Southern Germany, especially regarding the distinction between L+H* vs. (LH)*. In section "The Present Study: Rationale and Hypotheses", we hypothesized about mergers toward the more frequent accent type, i.e., mergers toward L+H* in Northern German and toward L*+H in Southern German speakers due to a high occurrence frequency of these complementary accents in the respective varieties (Kügler, 2004, 2007; Peters et al., 2005)—an account that had predicted less clear-cut distinctions between L+H* and (LH)* in the North and between L*+H and (LH)* in the South. This prediction was clearly not borne out: Instead, Northern German speakers were more distinct in all pairwise comparisons than Southern German speakers, especially with regard to the distinction L+H* and (LH)*, which almost seemed to converge for large parts of the contours in speakers from the South. We will discuss this finding and its implications in more detail in the General Discussion, including results from the association task in Experiment 2. Summarizing the main findings from Experiment 1, our results indicate that speakers maintain a three-way contrast in their imitative productions. Crucially, (LH)* significantly differs in its **form** from the more established accents L+H* and L*+H in all experimental conditions, in particular for speakers from the North.

EXPERIMENT 2: PARAPHRASING OF CONNOTATIVE QUESTION MEANING

In Experiment 2, we tested whether the three rising-falling contours evoke different connotative meanings. To this end, listeners paraphrased the pragmatic meaning they associated with the stimuli from Experiment 1.

Methods

Participants

Overall, 66 native speakers of German were included in the study. None of the speakers had participated in Experiment 1. Twenty-eight of them were from Southern Germany (mean age: 24.1 years, SD = 4.3 years, 24 female, 4 male), that is, Baden-Wuerttemberg ($N = 20$) and Bavaria ($N = 8$), and 38 were from Northern Germany (mean age: 25.5 years, SD = 8.2 years, 33 female, 5 male), which includes the states of Saxony-Anhalt (1), Lower-Saxony ($N = 18$), Hamburg ($N = 3$), Bremen ($N = 1$), North-Rhine Westphalia ($N = 5$), and Schleswig-Holstein ($N = 10$). Data from five additional speakers was not considered since these participants could not unambiguously be assigned to the Southern or the Northern group [i.e., participants who were born in the North but grew up in the South, or vice versa ($N = 2$), or came from the Central German dialect area ($N = 3$)].

Materials

We selected 12 *wh*-interrogatives from the material set used in Experiment 1. Since the direction of resynthesis of stimuli (manipulated from L+H* vs. L*+H) did not have an effect on the imitation results in Experiment 1 (see Mendeley), we reduced the number of stimuli by taking only one manipulation direction into account. That is, for L+H* we used those stimuli that were originally recorded as L+H*; similarly, for L*+H we used those stimuli that were originally recorded as L*+H. For (LH)* contours, which were resynthesized half from originally recorded L+H* and half from L*+H in Experiment 1, we chose two of the four questions to be originally recorded as L+H*, and two as L*+H contours. This resulted in 12 items (4 items × 3 intonation conditions, only one manipulation direction). The 12 *wh*-interrogatives were ordered such that the stimuli with the same intonation condition and stimuli with the same lexicalization were separated by at least one other item to avoid priming.

Procedure

Participants were asked to paraphrase the intention they thought a speaker conveyed in the question. They were told that interrogative sentences may not only be used for inquiring information but can also serve other purposes (which were not further specified). Participants were furthermore explicitly instructed to focus on *how* the respective utterances sounded, that is, which connotative meaning the utterances expressed, disregarding their propositional content. The four different target *wh*-questions [cf. (1) above] were presented in written form in the instructions. This was done to familiarize participants with the syntactic and lexical composition of the target sentences and to focus them on the intonational realization of the utterances.

On each trial in the actual experiment, participants clicked on a "play" button to listen to one *wh*-question at a time. They then had to paraphrase the intention of the speaker in a free response field. The experiment was self-paced and participants were allowed to listen to a question as often as they wanted but were instructed to respond intuitively. In case they associated different intentions with the questions, they were allowed to give multiple responses; conversely, if participants did not identify an intention, they typed "NA" in the description field or left it blank. Participants moved on to the next trial by pressing a "continue" button. The study was conducted as a web-based experiment, which was created via *SoSci Survey*

TABLE 3 | Number of responses split by experimental item and intonation condition.

	Libero	Malibu	Mandalas	Melanie	Total
L+H*	59 (50)	66 (54)	55 (42)	81 (78)	**261 (224)**
(LH)*	75 (73)	88 (84)	71 (69)	72 (68)	**306 (294)**
L*+H	58 (49)	52 (47)	71 (57)	60 (54)	**241 (207)**
Total	**192 (172)**	**206 (185)**	**197 (168)**	**213 (200)**	**808 (725)**

Values in brackets indicate the number of instances that were statistically analyzed (after response categories with fewer than 12 instances per category had been removed).

(www.soscisurvey.com, Leiner, 2018), and ran on an in-house server. Participants took between 10 and 15 min to complete the study.

Data Treatment and Analysis

In total, 808 responses were given [261 for L+H*, 306 for (LH)* and 241 for L*+H], see **Table 3** for the distribution of responses across items and intonation condition. Listeners gave on average 12.2 responses for the 12 trials (SD = 3.7), with individuals ranging between 3 (i.e., responses to only a fourth or the trials) and 21 responses (i.e., almost two responses to every trial). There were 135 questions to which participants did not provide a response at all [50 times for L+H*, 29 times for (LH)*, and 56 times for L*+H].

We extracted superordinate categories from participants' responses. To this end, we took a sample of about a third of the data ($N = 262$) and grouped the responses into a set of superordinate categories (e.g., "information-seeking," "surprise," "p is odd"). These categories were generated bottom-up (i.e., data-driven) and were often explicitly mentioned by a number of participants (e.g., scepticism). The grouping was done together by two coders (a consensual coding between first and third author). **Table 4** lists the nine superordinate categories that contained $N >= 12$ instances each. It also includes example responses for each category. To verify the objectivity of the superordinate categories and the reliability with which they can be coded, a third coder (fourth author) independently coded a subset of 220 responses based on the keywords in **Table 4**. Agreement between the third coder and the consensual coding of the first two coders was assessed by calculating Cohen's kappa (Cohen, 1960) with the *irr package* in R. The interrater agreement was 88.2% ($\kappa = 0.83$), i.e., "almost perfect" (Landis and Koch, 1977, p. 165). The set of keywords was then used to code the remaining items by one of the three coders (first, third, and fourth author).

In total, responses fell into 27 superordinate categories, with instances in individual categories ranging between 1 and 370 responses. To keep the number of categories feasible for statistical analysis, we excluded categories with fewer than 12 instances per category, in total excluding 83 responses in 18 different categories (10.3% of the data). The statistical analysis was based on the nine different response categories of **Table 4** ($N = 725$ responses), see values in brackets in **Table 3** for distribution across condition and items.

TABLE 4 | Nine most frequent superordinate categories inferred from participants' responses.

Superordinate category	Exemplar responses
Aversion ($N = 77$)	*Abschätzige Meinung zu Mandalas*; ("pejorative opinion on mandalas") *Auf Mandalas malen als Beschäftigung wird herabgesehen;* ("disdaining coloring mandalas as an activity")
Information-seeking ($N = 370$)	*Wer malt gerade ein Mandala?*; ("Who is coloring a mandala at the moment?") *Herausfinden, wer Melanie heißt;* ("find out which of the persons is called Melanie") *Tatsächliches Interesse;* ("actual interest")
Irony ($N = 17$)	*Ironische Frage*; ("ironic question") *Ablehnung/ Spott ausdrücken*; ("to express rejection and mockery")
Negative attitude ($N = 54$)	*Melanie ist kein schöner Name;* ("Melanie is not a nice name.") *Kritische Äußerung zu Mandalas;* ("critical statement toward Mandalas")
P is odd ($N = 53$)	*Melanie ist ein ungewöhnlicher Name;* ("Melanie is an unusual name.") *Mandalas malen ist ungewöhnlich;* ("Drawing mandalas is unusual.") *Dass Leute, die Mandalas malen, komisch sind;* ("That people who draw mandalas are weird.")
Positive attitude ($N = 13$)	*Bewusst geduldig und freundlich auftreten*; ("to intentionally appear patient and friendly"), Malibu wird positiv bewertet; ("Malibu is rated positively.")
Rhetorical meaning ($N = 47$)	*Niemand heißt Melanie;* ("Nobody is called Melanie.") *Melanie als Name wird infrage gestellt. Wer heißt denn schon so?;* ("The name Melanie itself is questioned. Who is called Melanie?") *Rhetorische Frage;* ("rhetorical question")
Scepticism ($N = 12$)	*Zweifel*; ("doubt") *Skepsis*; ("scepticism")
Surprise ($N = 82$)	*Libero spielen ist etwas, das man nicht erwarten würde;* ("To play in the sweeper position is not something one would expect.") *Verwunderung*; ("astonishment/surprise")

Original responses for the superordinate category are shown on the right. Categories are presented in alphabetical order. The number in brackets gives the total number of responses in this category.

We used Conditional Inference Trees (CTrees) to test whether there was a significant clustering of response categories based on our two predictors *intonation condition* and *region*. CTrees are a non-parametric class of regression trees, applicable to all kinds of response variables (Hothorn et al., 2006). Different from other regression tree algorithms such as CART-based trees, CTrees employ a significance test procedure that grows only statistically significant splits. Hence, tree pruning is not needed in this approach [ctree() function description, Hothorn and Zeileis, 2015]. To fit the trees, we used the *partykit package* in R (Hothorn et al., 2006; Hothorn and Zeileis, 2015). To evaluate the generalization of the CTree, we used a 10-fold cross validation procedure: We split the data in 10 randomly sampled sets, training the tree on 85% of the data and testing it on the 15% of unseen data. For evaluation of the tree, the R package *caret* (Kuhn, 2020) was used.

Results

Figure 5 shows the distribution of the nine response categories across the three different intonation conditions, split by region ($N = 725$ responses).

Most of the questions were paraphrased as "information-seeking" ($N = 370$ of all 808 responses, 45.8%), which is not unexpected given their interrogative syntax. However, the specification "information-seeking" was more often ascribed with L+H* and L*+H accents, as compared to (LH)* accents: 59.4% of the L+H* accents, 59.8% of the L*+H accents, compared to 23.2% of the (LH)* accents. A similar distribution, but with much lower numbers, was found for the category "positive attitude." Conversely, all other response categories occurred more frequently in the (LH)* condition than in the L+H* and L*+H accents. That is, (LH)* was often paraphrased as "aversion," "surprise," "negative attitude," or "p is odd". A "rhetorical meaning" was also attributed to (LH)*, more often than for the two other accent types. However, this connotation was rare overall. The CTree shows one significant split only (**Figure 6**), which is caused by intonation condition, separating the (LH)* accent on the one hand from the L+H* and L*+H on the other. The L+H* and L*+H accents were not further subdivided. The factor *region* was not considered by the CTree, which mirrors the similar meaning attributions across regions shown in **Figure 5**. The evaluation of the unseen test set (15% of the data) revealed a mean accuracy of 51.8%, 95% CI [42.1%; 61.5%].[8]

It was surprising that the two established GToBI accents L+H* and L*+H were not distinguished in the meaning task. After all, these contours have often been claimed to differ in their communicative **function**: L+H* has been associated with new or contrastive information (Kohler, 1991b; Grice et al., 2005; Baumann and Grice, 2006), while L*+H has been associated with contrast and/or certain attitudinal meanings (Grice et al., 2005; Niebuhr, 2007b; Lommel and Michalsky, 2017). These meaning attributions refer to information structure, the information status of referents and to attitudes that are typically conveyed in utterances with a declarative syntax. It is conceivable that these intonational meanings are less obvious in the *wh*-question structure employed in our experimental sentences. An alternative explanation could be that the meaning contrast was not captured well by the superordinate categories. To follow up on these possibilities, we conducted a *post-hoc* study ($N = 15$ participants, 5 from Northern and 10 from Southern Germany) in which we used the same recordings of the target words (*Mandala, Malibu, Melanie*, and *Libero*), but spliced onto a declarative-sentence structure (*Das ist der/die* "That is the").[9] The instructions and the experimental procedure were the same as in Experiment 2. Participants' responses in this follow-up study were coded into keywords (which partly differed from the ones for *wh*-questions) and analyzed by the third and fourth author in a consensus coding. In this follow-up study, results revealed differences in interpretation between L+H* and L*+H: L+H* was more often paraphrased as "correction", "enforcement", "statement", "p is new", and "information-giving" than L*+H. Conversely, L*+H was more often paraphrased as "surprise" and "aversion" than L+H*. In line with the results of Experiment 2, (LH)* was interpreted more often as "correction", "surprise", and "aversion" than the other two accents (see **Supplementary Material S.3** for more details).

Interim Discussion

In Experiment 2, we assessed the connotative meanings listeners associate with the three different kinds of rising-falling contours, L+H*, (LH)*, and L*+H, schematized in **Figure 1**. This was done in a qualitative study in which participants freely paraphrased the perceived intention of the speaker. Our results showed that *intonation condition* clearly affects the connotative meanings associated with the questions, causing the only split in the CTree (see **Figure 6**). Importantly for the question of whether (LH)* forms its own category, the connotative meanings evoked by (LH)* were distinct from the two other accent types: While L+H* and L*+H contours were equally paraphrased as "information-seeking" in most of the *wh*-questions, (LH)* received more diverse meaning attributions, which were often paraphrased as "aversion", "surprise", "negative attitude", or "p is odd". An explicit "rhetorical" meaning was also ascribed to (LH)*, but this association was comparatively rare. Importantly, the pattern of results was the same across regions, suggesting that speakers from Northern and Southern Germany share the same set of connotative meanings for the three rising-falling contours.

We first discuss the connotations ascribed to the (LH)* accent before we turn to the finding that the meaning attributed to L+H* and L*+H did not differ in *wh*-questions. From a phonetic point of view, (LH)* differs from the other two accents in our study in that it exhibits a steeper slope of the rising movement since both the low and the high tonal target occur within the stressed syllable. This might have increased the perceptual salience of this accent type. As discussed briefly in section "Interim Discussion" of Experiment 1, previous prominence rating tasks have shown that rising nuclear accents are perceived

[8] Note that reducing the number of categories (excluding those with only few instances: irony, positive attitude and scepticism) led to the same results in the CTree, i.e., one single split grouping L+H*/L*+H on the one hand, and (LH)* on the other.

[9] For the item "Mandalas" we had to remove the final [s] to change it from a plural to a singular word form to match it to the referential expression *das* and the verb *ist*.

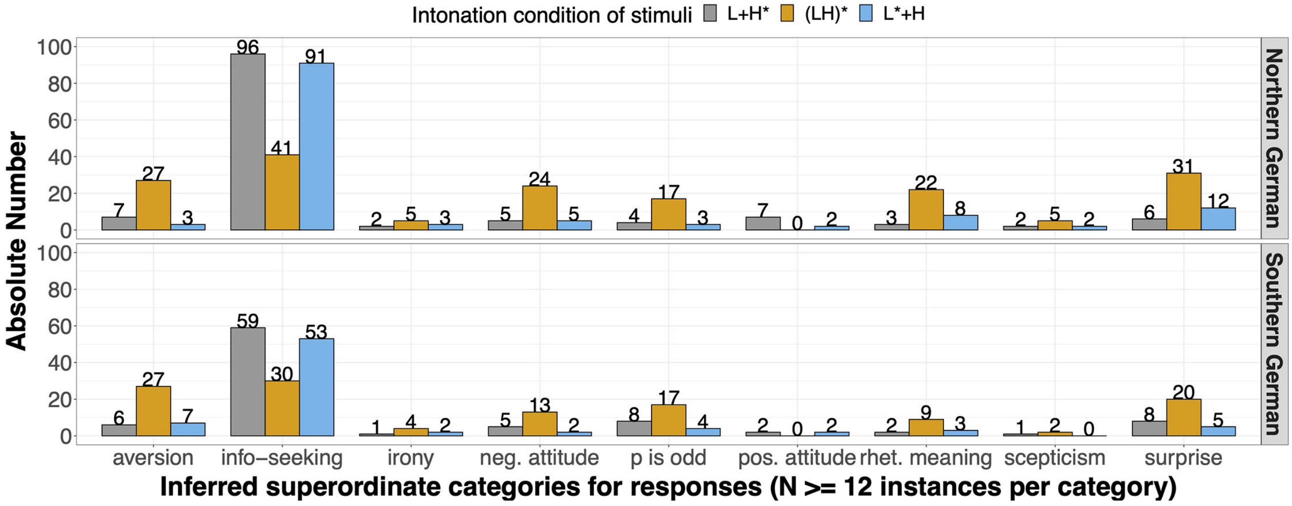

FIGURE 5 | Distribution of superordinate categories (inferred keywords from participants' responses), color-coded for the different intonation conditions [L+H* in gray, (LH)* in orange, and L*+H in blue]; split by region [upper panel for speakers from Northern Germany (*N* = 38 participants); lower panel for speakers from Southern Germany (*N* = 28 participants)].

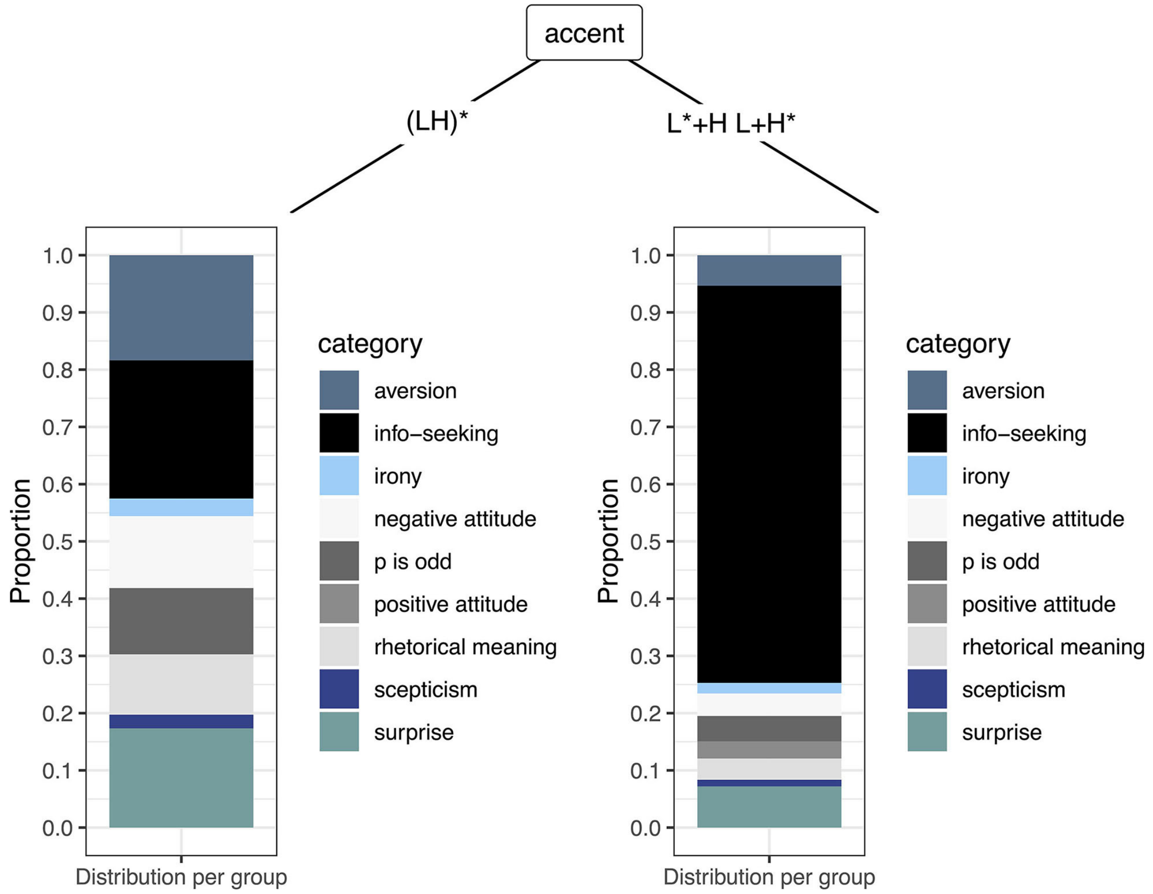

FIGURE 6 | Visualization of the Conditional Inference Tree. The predicted categories are shown in form of a stacked bar plot. The only split in the CTree was caused by intonation condition ($p < 0.001$), separating the (LH)* accent (left) from the GToBI accents L+H* and L*+H (right).

as more prominent than falling ones, H* accents as more prominent than L* accents, steeper slopes as more prominent than shallower ones, and larger f0 excursions as more prominent than smaller f0 excursions (Rietveld and Gussenhoven, 1985; Baumann and Röhr, 2015; Baumann and Winter, 2018). This implies the following decreasing prominence order among the accents of this study: (LH)* > L+H* > L*+H. This ranking is reproduced in only three of the nine categories ("p is odd," "negative attitude," "aversion"). Hence, while the perceptual prominence may have affected listeners' interpretations to some degree, differences in prominence alone cannot explain the findings. Clearly though, the steeper slope perceptually stands out. Higher peaks and concomitant steeper slopes have been shown to affect meaning interpretation: In particular, a steeper slope in rising movements has been associated with surprise in the literature (Ladd and Morton, 1997; Chen, 2009).

To further interpret the findings concerning the meaning attributions to (LH)*, it helps to access the pragmatics literature: As mentioned earlier, the accent (LH)* has been observed in a study on rhetorical questions (Braun et al., 2019). Several connotative meanings mentioned by the listeners are in fact compatible with a rhetorical question interpretation: Rhetorical questions are often described to have the illocutionary force of assertions (Han, 2002) or to be assertion-like (Caponigro and Sprouse, 2007; Biezma and Rawlins, 2017). The speaker of a rhetorical question commits her interlocutors to the proposition presupposed by the rhetorical question (Biezma and Rawlins, 2017). For positive *wh*-questions used in this paper, the presupposition denotes the empty set (e.g., *niemand* "nobody" for *Wer mag Mandalas?* "Who likes mandalas?"). At the same time, the speaker of a rhetorical question signals that the answer to the rhetorical question is obvious and she expects all interlocutors to know that it is obvious. A rhetorical question is not, a priori, connected to any specific kind of speaker emotion. However, it can convey a large range of emotional or attitudinal load. It may be used positively (e.g., *Mach dir keine Sorgen. Wer ist denn nicht nervös vor einer Prüfung?* "Don't worry. Who isn't nervous right before an exam?") or negatively (e.g., *Was weiß der schon?* "What does he know, after all?"). Hence, it is conceivable to assume that a rhetorical question may also trigger emotional stances such as aversion or negativity or a certain sense of surprise or oddness. The indeterminacy in attitudes also explains the larger variability in paraphrases in the (LH)* accent compared to the L+H* and L*+H accents. A frequent category for the (LH)* condition was "surprise" (e.g., that the speaker is surprised that there are people who like drawing mandalas). The surprise aspect ties in with observations that rhetorical questions may be marked by "mirativity markers" in some languages (for Basque: Alcázar, 2017). Also, rhetorical questions in English have been associated with surprise (Celle, 2018). Taken together, both the phonetic composition of (LH)* as well as pragmatic approaches explain why (LH)* evokes a different meaning than the two other accents. We now turn to the lack of distinction between L+H* and L*+H.

Contrary to our hypothesis, L+H* and L*+H contours in *wh*-interrogative structures did not lead to overall different judgments of meaning but were both predominantly interpreted as conveying the intent of requesting information from the addressee ("information-seeking"). The syntactic question form, i.e., the *wh*-verb-second-interrogative form, may have inflicted a strong bias, in particular for the two established accents L+H* and L*+H. Given ample evidence on different meaning contributions for L+H* and L*+H (Grice et al., 2005; Kohler, 2005; Niebuhr, 2007b; Lommel and Michalsky, 2017; Braun and Biezma, 2019), the lack of a distinction is indeed surprising, and may theoretically cast doubt on the validity of the study. This is not the case, however: A follow-up study conducted with the same design but with declarative sentences instead of *wh*-interrogatives showed meaning differences between L+H* and L*+H. Our findings corroborate current reports in the literature that challenge the view of a one-to-one mapping between intonational form and pragmatic meaning (Chodroff and Cole, 2019a; Roettger et al., 2019; Orrico and D'Imperio, 2020), but clearly show that differences in tonal alignment interact with propositional content and sentence structure to evoke different meaning interpretations.

Taken together, (LH)* differed in its interpretation from the GToBI accents L+H* and L*+H. The data may be partly explained in terms of a link between phonetic emphasis (steepness of the slope of the rise) and surprise. From a functional perspective, there is evidence that the (LH)* accent is interpreted differently from the L+H* and L*+H accents and that the latter two, regarding their meaning, do not differ in *wh*-questions—but, corroborating previous research, in declaratives.

GENERAL DISCUSSION

In the present study, we examined the distinctiveness in **form** and **function** of German nuclear rising-falling intonation contours. We focused on three rising-falling contours, which have been described in several different intonational frameworks and empirical studies on German intonation. Only one system previously discussed a three-way contrast (Kohler, 2005; Niebuhr, 2022); the other systems mostly provide a two-way distinction (Kohler, 1991a; Grice et al., 2005; Peters, 2006, 2014). The contours investigated in the present study are termed L+H* (H aligned in stressed syllable, L in preceding syllable), L*+H (L aligned in stressed syllable, H in following syllable), and (LH)* (both L and H aligned in the stressed syllable). In Experiment 1, we tested whether German speakers are able to imitate these three distinct rising-falling contours using a delayed imitation task, which taps into phonological processing (Baddeley and Hitch, 1974). Imitative productions were complemented by a qualitative study on the perceived intention of the speaker (Experiment 2). In both paradigms (**form** and **function**), we further investigated whether the regional background of the participants (Northern vs. Southern Germany) affects the ability to distinguish between the three contours in production and perception. The factor regional variety suggests itself (a) because Southern and Northern German speakers were found to align tonal targets differently in prenuclear rising accents—with later tonal targets in the South than the North (e.g., Atterer and Ladd, 2004) and (b) because

Northern German speakers use L+H* as most frequent accent type while Southern German speakers predominantly use L*+H (Peters et al., 2005; Kügler, 2007).

With respect to **form**, our results show that speakers of both varieties produced three distinct rising-falling contours. While imitated contours differed significantly in all experimental conditions, they were more distinct in Northern German speakers than in Southern German speakers, especially with regard to the contrast between L+H* and (LH)*. In none of the varieties did the contours totally converge onto a single or two contours. Based on the previous literature, one may have expected that frequent categories act as attractors (Anderson et al., 2003; Braun et al., 2006; Chodroff and Cole, 2019b). In this study, L+H* is the most frequent accent in Northern German (Peters et al., 2005), while L*+H is frequently used in Southern German (Kügler, 2007; Truckenbrodt, 2007). An attractor account would have predicted more mergers toward L+H* in Northern German and more mergers toward L*+H in Southern German (cf. Braun et al., 2006). However, we do not see evidence that one of these frequent accents acts as a perceptual magnet that is able to warp the perceptual space. It rather seems that even less frequent accent types can easily be held in memory and retrieved for production. It is likely that adult speakers of both regions have accumulated enough experience with different pitch accents, which allowed them to form the respective three categories (cf. Zahner et al., 2016, showing that already children who grow up in Southern Germany are exposed to the full German accent inventory from early on). Our data hence call for a three-way distinction between rising-falling contours for both regional varieties.

Let us nevertheless briefly speculate about the smaller distinction between L+H* and (LH)* contours for Southern German as compared to Northern German speakers. Since these two regional varieties have been shown to differ in the alignment of tonal targets, one may assume that this tendency is the cause for the small difference between L+H* and (LH)* in Southern German speakers. However, if tonal alignment differences uniformly applied to all accents in this speaker group, we would not have observed differences in the distinction of the contours across regions (but a main effect of region, later alignment in Southern than in Northern German speakers throughout). We argue that the contextually more restricted (LH)* accent is also more restricted in terms of its realization and needs to have both tonal targets realized within the stressed syllable. This requirement does not allow for variety-specific alignment differences. If (LH)* has fixed alignment (L toward the middle and H at the end of the stressed syllable), then a later aligned L+H* for Southern German speakers may lead to an overlap in production with (LH)*. The most important finding we take from Experiment 1 is that all contours significantly differed from each other, suggesting that the German acoustic space of rising-falling contours may be best described as a threefold partition.

The presence of three distinct contours is further corroborated by the data from Experiment 2 **(function)**. Listeners associated different meanings with (LH)* on the one hand, and with L+H* and L*+H on the other. We argued that the steep slope of the rise in (LH)* (both L and H in the stressed syllable with the same f0 excursion as L+H* and L*+H) may have evoked the perception of surprise (cf. Kohler, 2005; Chen, 2009). It is also conceivable that the increased prominence of this accent (Baumann and Röhr, 2015; Baumann and Winter, 2018) triggers implicatures through the effort code, such that an increased effort signals pragmatic relevance (cf. Hirschberg, 2002 on modeling intonational meaning in terms of implicatures). The attitudinally loaded meanings (aversion, negative attitude, unexpectedness, etc.) furthermore have been argued to be compatible with the pragmatics of rhetorical questions. In contrast, L+H* and L*+H were frequently paraphrased as information-seeking, but with no further differences in meaning, which was unexpected but seems to be due to the use of *wh*-questions (as differences were found when using declarative sentences). This, in turn, suggests that intonational meaning cannot entirely be dissociated from sentence type. In sum, we take the combined results from imitation and perception experiments as evidence for three types of distinct rising-falling contours in German.

This brings us to the question of whether and how we need to model three distinct contours in the acoustic space of German rising-falling contours. The present paper supports models that contain a three-way contrast (Kohler, 2005; Niebuhr, 2022) by providing combined evidence from production (**intonational form**) and perception (**function**). Our results show that (LH)* is distinct in **form** and **function**. For answering the question on the number of contours we need to model, it may be helpful to keep an open eye for the (LH)* pitch accent in future transcriptions of German intonation to learn more about its distribution and the (phonetic, phonological, syntactic, or pragmatic) conditions in which it occurs. As the meaning data suggest, its use is likely not restricted to German rhetorical questions (cf. Kohler, 2005; Braun et al., 2019; Wochner, 2021). Pilot data of a study on the realization of sarcastic irony suggest that this type of accent, (LH)*, also frequently occurs in sarcastic utterances of the sort *Das klappt ja super* "That works PRT great," accent on *super* (Fünfgeld et al., 2021). As those utterances are clearly attitudinally loaded, it is not surprising that speakers also employ (LH)* in ironic situations. Also, recent data suggest that (LH)* may also be found in exclamative sentences, e.g., *Kann die Lene malen* "Can Lene paint!" (Wochner and Dehé, 2018; Wochner, 2021), which have been described to express an attitude of surprise in the sense that a speaker conveys that the proposition of an utterance is unexpected (e.g., Fries, 1988). As the current experiment investigated solely *wh*-questions with a very homogeneous structure, more research needs to be done to answer the question of whether or not the (LH)* constitutes its own phonological category or not. With the current knowledge, it seems justifiable to posit three kinds of rising-falling accents in German.

An alternative view is that (LH)* is a meaningful modification of the more established accents L+H* or L*+H. The small differences between L+H* and (LH)* for Southern German speakers, in particular, might in fact allow such an interpretation. The idea of phonetic modifications of pitch accent types is not new (Ladd and Morton, 1997; Gussenhoven, 2004; Ladd, 2008, p. 155f). Other studies have also called for a gradual mapping between intonation and meaning, cf. Chen and Gussenhoven (2008) on aspects of gradience in the encoding of emphasis in a tone language, Orrico and D'Imperio (2020) on gradience in intonation-meaning mapping in biased questions or Dorokhova and D'Imperio (2019) on gradience in the interpretation of final rises in French. In terms of modeling, Ladd and Morton (1997) labeled rising-falling accents that differ regarding their pitch range with a binary feature [± emphatic]. The difference between "normal" and "emphatic" accents may be interpreted categorically (e.g., to signal uncertainty or incredulity) but not perceived as categorically different kinds of pitch accents (Ladd and Morton, 1997, p. 339). For our purposes, the feature [± emphatic] would serve the purpose to single out the (LH)* accent from the other two accents. Alternatively, acoustic features, such as [+ steep slope] could be used for achieving the (LH)* from an L*+H or L+H* accent. This proposal would gain support if we found this kind of modification, ideally with similar contributions to meaning, also for other accent types or contexts in German. Future research needs to test this proposal.

Taken together, our study addressed the question of how many rising-falling contours are needed to best describe the German perceptual space, thus serving as an attempt to clear up the "fuzzy" space for these contours. We hence addressed the distinctiveness in **form** and **function** of three different contours, two widely established accents L+H* vs. L*+H, and an accent which lies in-between, (LH)*. Our data show that speakers can differentiate between these three contours, both in perception and production, suggesting the existence of three kinds of rises. The most prototypical connotative meaning of the (LH)* accent is "surprise," which was frequently mentioned in *wh*-questions and declaratives for this accentual realization. Generally, the intonational contrasts are employed and interpreted in a consistent and meaningful way. Advocating (LH)* as a third category (beyond the two more established ones) might appear premature, as the same effect may be achieved by employing modifications of the other two accents. However, as it stands, it is difficult to decide whether (LH)* would be a variant of L*+H or of L+H*. If anything, the data from Southern German speakers, in particular, suggest that (LH)* is more likely to be a variant of L+H* than of L*+H. In future research, we plan to present contours that are more variable in the phonetic space (with more phonetic variability in the alignment and shape of accents and the steepness of the slope) to better map the "white spots" between contours. We further plan to test a wider variety of accentual realizations and sentence types to gain more insights into a broader range of contexts.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study, along with analyses scripts, can be found in the Mendeley repository: http://dx.doi.org/10.17632/yhv7nmjmgf.2.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by IRB Konstanz, 05/2021. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

KZ-R and ME processed and annotated the data for Experiment 1. KZ-R, DW, and AJ for Experiment 2. KZ-R led stimulus preparation (resynthesis), data analysis (annotation and processing), and statistical analyses (supported by BB). KZ-R drafted the manuscript. All authors contributed to the idea of using a combined approach (form-function) to study the distinctiveness of rising-falling contours and collectively developed the study designs. All authors contributed to the article, edited and wrote parts of the manuscript, and approved the submitted version and are responsible for it.

FUNDING

This research was funded by the DFG as part of research unit Questions at the Interfaces (FOR 2111, P6), Grant Numbers BR 3428/4-1,2 and DE 876/3-1,2.

ACKNOWLEDGMENTS

We are very grateful to the speaker of our experimental stimuli, Clara Huttenlauch. We also thank Johannes Heim, Jan Michalsky, Andrea Pešková, and Meg Zellers for distributing the link to the online experiment in Northern Germany, and Massimiliano Canzi and Kajsa Djärv for advice on the Conditional Inference Trees. We thank Moritz Jakob for testing (in the lab) and the implementation of the imitation study on SoSci Survey, Friederike Hohl for data processing with MAUS and the generation of the algorithm for f0 correction, and Justin Hofenbitzer, Katharina Hölzl, Elena Schweitzer, Friederike Hohl, Meike Rommel, Johanna Schnell, Meiqiao Li, and Svenja Willenborg for help with segmental annotation. We are very grateful to Cesko Voeten for practical advice on GAMM modeling, and to Melanie Loes and Vivien Meyer for help with checking the bibliography. We are grateful to the audiences at LabPhon2020 (Satellite Workshop "Situating phonological contrast within the production-perception loop") in Vancouver (virtual conference) for discussion and feedback on preliminary results of the imitation study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.838955/full#supplementary-material

REFERENCES

Alcázar, A. (2017). "A syntactic analysis of rhetorical questions," in *Proceedings of the the 34th West Coast Conference on Formal Linguistics* (Somerville, MA), 32–41.

Anderson, J. L., Morgan, J. L., and White, K. S. (2003). A statistical basis for speech sound discrimination. *Lang. Speech* 46, 155–182. doi: 10.1177/00238309030460020601

Arvaniti, A. (2019). "Crosslinguistic variation, phonetic variability, and the formation of categories in intonation," in *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)* (Melbourne, VIC), 1–6.

Arvaniti, A., and Fletcher, J. (2020). "The autosegmental-metrical theory of intonational phonology," in *The Oxford Handbook of Language Prosody*, eds C. Gussenhoven and A. Chen (Oxford: Oxford University Press), 78–95.

Atkinson, J. E. (1976). Inter- and intraspeaker variability in fundamental voice frequency. *J. Acoust. Soc. Am.* 60, 440–445. doi: 10.1121/1.381101

Atterer, M., and Ladd, D. R. (2004). On the phonetics and phonology of "segmental anchoring" of F0: evidence from German. *J. Phon.* 32, 177–197. doi: 10.1016/S0095-4470(03)00039-1

Baayen, R. H., van Rij, J., de Cat, C., and Wood, S. N. (2018). "Autocorrelated errors in experimental data in the language sciences: some solutions offered by Generalized Additive Mixed Models," in *Mixed Effects Regression Models in Linguistics*, eds D. Speelman, K. Heylen, and D. Geeraerts (Berlin: Springer), 49–69.

Baddeley, A. D. (1986). *Working Memory*. Oxford: Oxford University Press.

Baddeley, A. D. (2003). Working memory and language: an overview. *J. Commun. Disord.* 36, 189–208. doi: 10.1016/S0021-9924(03)00019-4

Baddeley, A. D., and Hitch, G. J. (1974). "Working memory," in *The Psychology of Learning and Motivation,* Vol. 8, ed G. H. Bower (London: Academic Press), 47–90.

Barnes, J., Brugos, A., Shattuck-Hufnagel, S., and Veilleux, N. (2013). "On the nature of perceptual differences between accentual peaks and plateaux," in *Understanding Prosody: The Role of Context, Function and Communication*, ed O. Niebuhr (Berlin; New York, NY: de Gruyter), 93–118.

Barnes, J., Brugos, A., Veilleux, N., and Shattuck-Hufnagel, S. (2021). On (and off) ramps in intonational phonology: rises, falls, and the Tonal Center of Gravity. *J. Phon.* 85. doi: 10.1016/j.wocn.2020.101020

Barnes, J., Veilleux, N., Brugos, A., and Shattuck-Hufnagel, S. (2012). Tonal Center of Gravity: a global approach to tonal implementation in a level-based intonational phonology. *Lab. Phonol.* 3, 337–383. doi: 10.1515/lp-2012-0017

Batliner, A. (1989). "Wieviel Halbtöne braucht die Frage? Merkmale, Dimensionen, Kategorie [How many semitones does a question need? Characteristics, dimensions, category]," in *Zur Intonation von Modus und Fokus im Deutschen [On the Intonation of Mode and Focus in German]*, Vol. 234, eds H. Altmann, A. Batliner, and W. Oppenrieder (Tübingen: Niemeyer), 111–153.

Baumann, S., and Grice, M. (2006). The intonation of accessibility. *J. Pragmat.* 38, 1636–1657. doi: 10.1016/j.pragma.2005.03.017

Baumann, S., and Röhr, C. (2015). "The perceptual prominence of pitch accent types in German," in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)* (Glasgow).

Baumann, S., and Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *J. Phonetics* 70, 20–38. doi: 10.1016/j.wocn.2018.05.004

Biezma, M., and Rawlins, K. (2017). Rhetorical questions: severing asking from questioning. *Proc. SALT* 27, 302–322. doi: 10.3765/salt.v27i0.4155

Boersma, P., and Weenink, D. (2016). *Praat: Doing Phonetics by Computer*. Version 6.1.42 (version depended on labeller) [Computer Program].

Braun, B. (2006). Phonetics and phonology of thematic contrast in German. *Lang. Speech* 49, 451–493. doi: 10.1177/00238309060490040201

Braun, B. (2007). "Effects of dialect and context on the realisation of German prenuclear accents," in *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)* (Saarbrücken), 961–964.

Braun, B., and Biezma, M. (2019). Prenuclear L*+H activates alternatives for the accented word. *Front. Psychol.* 20, 1993. doi: 10.3389/fpsyg.2019.01993

Braun, B., Dehé, N., Neitsch, J., Wochner, D., and Zahner, K. (2019). The prosody of rhetorical and information-seeking questions in German. *Lang. Speech* 62, 779–807. doi: 10.1177/0023830918816351

Braun, B., Kochanski, G., Grabe, E., and Rosner, B. (2006). Evidence for attractors in English intonation. *J. Acoustical Soc. Am.* 119, 4006–4015. doi: 10.1121/1.2195267

Caponigro, I., and Sprouse, J. (2007). "Rhetorical questions as questions," in *Proceedings of the Sinn und Bedeutung* (Barcelona: Universitat Pompeu Fabra), 121–133.

Celle, A. (2018). "Questions as indirect speech acts in surprise contexts," in *Tense, Aspect, Modality and Evidentiality. Crosslinguistic Perspectives*, eds D. Ayoun, A. Celle, and L. Laure (Amsterdam; Philadelphia, PA: John Benjamins), 213–238.

Chen, A. (2009). Perception of paralinguistic intonational meaning in a second language. *Lang. Learn.* 59, 367–409. doi: 10.1111/j.1467-9922.2009.00510.x

Chen, Y., and Gussenhoven, C. (2008). Emphasis and tonal implementation in Standard Chinese. *J. Phon.* 36, 724–746. doi: 10.1016/j.wocn.2008.06.003

Chodroff, E., and Cole, J. (2019a). "The phonological and phonetic encoding of information structure in American English nuclear accents," in *Proceedings of the International Congress of Phonetic Sciences (ICPhS)* (Melbourne, VIC), 1570–1574.

Chodroff, E., and Cole, J. (2019b). "Testing the distinctiveness of intonational tunes: Evidence from imitative productions in American English," in *Proceedings of the Proceedings the 20th Annual Conference of the International Speech Communication Association (Graz: Interspeech 2019)*, 1966–1970.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Crowder, R. G. (1982). Decay of auditory memory in vowel discrimination. *J. Exp. Psychol. Learn. Memory Cogn.* 8, 153–162. doi: 10.1037/0278-7393.8.2.153

Dehé, N., Braun, B., Einfeldt, M., Wochner, D., and Zahner-Ritter, K. (2022). The prosody of rhetorical questions: a cross-linguistic view. *Linguistische Berichte*, 3–42.

Dilley, L. C. (2010). Pitch range variation in English tonal contrasts: continuous or categorical? *Phonetica* 67, 63–81. doi: 10.1159/000319379

Dilley, L. C., and Brown, M. (2007). Effects of pitch range variation on f0 extrema in an imitation task. *J. Phon.* 35, 523–551. doi: 10.1016/j.wocn.2007.01.003

Dombrowski, E. (2003). "Semantic features of accent contours: Effects of F0 peak position and F0 time shape," in *Proceedings of the 15th International Congress of Phonetic Sciences* (Barcelona), 1217–1220.

Dombrowski, E., and Niebuhr, O. (2010). "Shaping phrase-final rising intonation in German," in *Proceedings of the Fifth International Conference on Speech Prosody* (Chicago, IL), 1–4.

Dorokhova, L., and D'Imperio, M. (2019). "Rise dynamics determines tune perception in French: The case of questions and continuations," in *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)* (Melbourne, VIC), 691–695.

Fitzpatrick-Cole, J. (1999). "The alpine intonation of Bern Swiss German," in *Proceedings of the 14th International Congress of the Phonetic Sciences (ICPhS)* (San Francisco, CA), 941–944.

Fletcher, J., Grabe, E., and Warren, P. (2005). "Intonational variation in four dialects of English: the high rising tune," in *Prosodic typology. The Phonology of Intonation and Phrasing*, ed S.-A. Jun (Oxford: Oxford Univeristy Press), 390–409.

Fries, N. (1988). "Ist Pragmatik schwer! - Über sogenannte Exklamativsätze im Deutschen [Pragmatics is difficult - On so called exclamatives in German]," in *Sprache und Pragmatik. Veröffentlichung des Lunder Projektes "Sprache und Pragmatik"*, ed I. Rosengren (Lund: Germanistisches Institut der Universität Lund), 1–18.

Fünfgeld, S., Braun, A., and Zahner-Ritter, K. (2021). *The Phonetics of Verbal Irony. A Contrastive Study of Two German Regional Accents.* Talk at the Common Phonetics Colloquium Trier University and Saarland University. Saarbrücken; Trier.

Gandour, J., Potisuk, S., Ponglorpisit, S., and Dechongkit, S. (1991). Inter- and intraspeaker variability in fundamental frequency of Thai tones. *Speech Commun.* 10, 355–372. doi: 10.1016/0167-6393(91)90003-C

Gathercole, S. E., Hitch, G. J., Service, E., and Martin, A. J. (1997). Phonological short-term memory and new word learning in children. *Dev. Psychol.* 33, 966–979. doi: 10.1037/0012-1649.33.6.966

Gilles, P. (2005). *Regionale Prosodie im Deutschen: Variabilität in der Intonation von Abschluss und Weiterweisung [Regional Prosody in German: Variability in the Intonation of Terminality and Continuation]*, Vol. 6. Berlin: de Gruyter. doi: 10.1515/9783110201611

Grabe, E. (2004). "Intonational variation in urban dialects of English spoken in the British Isles," in *Regional Variation in Intonation*, eds P. Gilles and J. Peters (Tübingen: Niemeyer), 9–31.

Grice, M., and Baumann, S. (2002). Deutsche Intonation und GToBI [German intonation and GToBI]. *Linguistische Berichte* 191, 267–298.

Grice, M., Baumann, S., and Benzmüller, R. (2005). "German intonation in autosegmental-metrical phonology," in *Prosodic Typology. The Phonology of Intonation and Phrasing*, ed J. Sun-Ah (Oxford: Oxford University Press), 55–83.

Grice, M., Ritter, S., Niemann, H., and Roettger, T. (2017). Integrating the discreteness and continuity of intonational categories. *J. Phon.* 64, 90–107. doi: 10.1016/j.wocn.2017.03.003

Gussenhoven, C. (2004). *The Phonology of Tone and Intonation.* Cambridge: Cabridge University Press.

Han, C.-H. (2002). Interpreting interrogatives as rhetorical questions. *Lingua* 112, 201–229. doi: 10.1016/S0024-3841(01)00044-4

Hanssen, J. (2017). *Regional Variation in the Realization of Intonation Contours in the Netherlands.* Utrecht: LOT.

Hirschberg, J. (2002). "The pragmatics of intonational meaning," in *Proceedings of the 1st International Conference on Speech Prosody, April 11-13 2002* (Aix-en-Provence), 65–68.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: a conditional inference framework. *J. Comput. Graph. Stat.* 15, 651–674. doi: 10.1198/106186006X133933

Hothorn, T., and Zeileis, A. (2015). partykit: a modular toolkit for recursive partitioning in R. *J. Mach. Learn. Res.* 16, 3905–3909. Available online at: http://jmlr.org/papers/v16/hothorn15a.html

Kisler, T., Reichel, U. D., and Schiel, F. (2017). Multilingual processing of speech via web services. *Comput. Speech Lang.* 45, 326–347. doi: 10.1016/j.csl.2017.01.005

Kohler, K. (1987). "Categorical pitch perception," in *Proceedings of the 11th International Congress of Phonetic Sciences (ICPhS)* (Tallinn), 331–333.

Kohler, K. (1991a). A model of German intonation. *Arbeitsberiechte des Instituts für Phonetik und Digitale Sprachverarbeitung der Universität Kiel (AIPUK)* 25, 295–360.

Kohler, K. (1991b). Terminal intonation patterns in single-accent utterances of German: phonetics, phonology and semantics. *Arbeitsberichte des Instituts für Phonetik und Digitale Sprachverarbeitung der Universität Kiel (AIPUK)* 25, 115–185.

Kohler, K. (2005). Timing and communicative functions of pitch contours. *Phonetica* 62, 88–105. doi: 10.1159/000090091

Kügler, F. (2004). "The phonology and phonetics of rising pitch accents in Swabian," in *Regional Variation in Intonation*, eds P. Gilles and J. Peters (Tübingen: Niemeyer), 75–98.

Kügler, F. (2007). *The Intonational Phonology of Swabian and Upper Saxon.* Tübingen: Niemeyer.

Kügler, F., and Gollrad, A. (2015). Production and perception of contrast: the case of the rise-fall contour in German. *Front. Psychol.* 6, 1254. doi: 10.3389/fpsyg.2015.01254

Kuhn, M. (2020). *Classification and Regression Training (caret). R package, Version 6.0-86.* Available online at: http://cran.r-project.org/web/packages/caret/index.html

Ladd, D. R. (2008). *Intonational Phonology, 2nd Edn.* Cambridge: Cambridge University Press.

Ladd, D. R., and Morton, R. (1997). The perception of intonational emphasis: continuous or categorical? *J. Phon.* 25, 313–342. doi: 10.1006/jpho.1997.0046

Landis, J. R., and Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 363–374. doi: 10.2307/2529786

Leemann, A. (2012). *Swiss German Intonation Patterns.* Amsterdam; Philadelphia, PA: John Benjamins Publishing Company.

Lehiste, I. (1975). "The phonetic structure of paragraphs," in *Structure and Process in Speech Perception*, eds A. Cohen and S. G. Nooteboom (Berlin: Springer), 195–203.

Leiner, D. J. (2018). *SoSci Survey (Current Version: Version 3.2.05-i in 2020).* Available online at: http://www.soscisurvey.com

Lohfink, G., Katsika, A., and Arvaniti, A. (2019). "Variability and category overlap in the realization of intonation," in *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)* (Melbourne, VIC), 701–705.

Lommel, N., and Michalsky, J. (2017). "Der Gipfel des Spotts. Die Ausrichtung von Tonhöhengipfeln als intonatorisches Indiz für Sarkasmus [Peak alignment as intonational cue to sarcasm]," in *Diversitas Linguarum*, Vol. 42, eds N. Levkovych and A. Urdze (Bremen: Universitätsverlag Dr. N. Brockmeyer), 33.

Mayer, J. (1995). *Transcription of German Intonation - The Stuttgart System*. University of Stuttgart. Available online at: https://www.ims.uni-stuttgart.de/institut/arbeitsgruppen/ehemalig/ep-dogil/joerg/labman/STGTsystem.html

Mücke, D., Grice, M., Becker, J., and Hermes, A. (2009). Sources of variation in tonal alignment: Evidence from acoustic and kinematic data. *J. Phon.* 37, 321–338. doi: 10.1016/j.wocn.2009.03.005

Mücke, D., Grice, M., Hermes, A., and Becker, J. (2008). "Prenuclear rises in northern and southern German," in *Proceedings of the 4th International Conference on Speech Prosody 2008* (Campinas), 245–248.

Niebuhr, O. (2007a). "Categorical perception in intonation: a matter of signal dynamics?" in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech)* (Antwerp), 109–112.

Niebuhr, O. (2007b). The signalling of German rising-falling intonation categories - the interplay of synchronization, shape, and height. *Phonetica* 64, 174–193. doi: 10.1159/000107915

Niebuhr, O. (2022). "The Kiel intonation model," in *Prosodic Theory and Practice*, eds J. Barnes and S. Shattuck-Hufnagel (Cambridge: MIT Press), 287–318.

Niebuhr, O., and Ambrazaitis, G. I. (2006). "Alignment of medial and late peaks in German spontaneous speech," in *Proceedings of the 3rd International Conference of Speech Prosody* (Dresden), 161–164.

Niebuhr, O., D'Imperio, M., Gili Fivela, B., and Cangemi, F. (2011). "Are there "shapers" and "aligners"? Individual differences in signalling pitch accent category," in *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS)* (Hong Kong), 120–123.

Ohala, J. (1990). There is no interface between phonology and phonetics: a personal view. *J. Phon.* 18, 153–171. doi: 10.1016/S0095-4470(19)30399-7

Orrico, R., and D'Imperio, M. (2020). Individual empathy levels affect gradual intonation-meaning mapping: the case of biased questions in Salerno Italian. *Lab. Phonol.* 11, 1–39. doi: 10.5334/labphon.238

Peters, B., Kohler, K., and Wesener, T. (2005). "Melodische Satzakzentmuster in prosodischen Phrasen deutscher Spontansprache - Statistische Verteilung und sprachliche Funktion [Melodic sentence accent patterns in prosodic phrases of German spontaneous speech - Statistical distribution and linguistic function]," in *Prosodic Structures in German Spontaneous Speech (AIPUK 35a)*, eds K. Kohler, F. Kleber, and B. Peters (Kiel: IPDS), 185–201.

Peters, J. (2006). *Intonation deutscher Regionalsprachen.* Berlin: de Gruyter.

Peters, J. (2014). *Intonation*. Heidelberg: Winter.

Petrone, C., D'Alessandro, D., and Falk, S. (2021). Working memory differences in prosodic imitation. *J. Phon.* 89, 101100. doi: 10.1016/j.wocn.2021.101100

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation* (PhD thesis). Massachusetts Institute of Technology; Department of Linguistics and Philosophy, Boston, MA, United States.

Pierrehumbert, J. B. (2016). Phonological representation: beyond abstract versus episodic. *Annu. Rev. Linguist.* 2, 33–52. doi: 10.1146/annurev-linguistics-030514-125050

Pierrehumbert, J. B., and Steele, S. A. (1989). Categories of tonal alignment in English. *Phonetica* 46, 181–196. doi: 10.1159/000261842

Plomp, R. (1964). Rate of decay of auditory sensation. *J. Acoust. Soc. Am.* 36, 277–282. doi: 10.1121/1.1918946

Prieto, P. (2011). "Tonal alignment," in *The Blackwell Companion to Phonology*, eds M. v. Oostendorp, C. Ewen, B. Hume, and K. Rice (Malden, MA; Oxford: Blackwell), 1185–1203.

Prieto, P. (2012). "Experimental methods and paradigms for prosodic analysis," in *The Oxford Handbook of Laboratory Phonology*, eds A. Cohn, C. Fougeron, and M. Huffman (Oxford: Oxford University Press), 528–537.

Prieto, P. (2015). Intonational meaning. *Wiley Interdiscipl. Rev. Cogn. Sci.* 6, 371–381. doi: 10.1002/wcs,.1352

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rathcke, T. (2017). How truncating are 'truncating languages'? Evidence from Russian and German. *Phonetica* 73, 194–228. doi: 10.1159/000444190

Rietveld, A. C. M., and Gussenhoven, C. (1985). On the relation between pitch excursion size and prominence. *J. Phon.* 13, 299–308. doi: 10.1016/S0095-4470(19)30761-2

Ritter, S., and Grice, M. (2015). The role of tonal onglides in German nuclear pitch accents. *Lang. Speech* 58(Pt 1), 114–128. doi: 10.1177/0023830914565688

Roessig, S. (2021). "Categoriality and continuity in prosodic prominence," in *Studies in Laboratory Phonology*, Vol. 10 (Berlin: Language Science Press).

Roessig, S., Mücke, D., and Grice, M. (2019). The dynamics of intonation: Categorical and continuous variation in an attractor-based model. *PLoS ONE* 14, e0216859. doi: 10.1371/journal.pone.0216859

Roettger, T. B., Mahrt, T., and Cole, J. (2019). Mapping prosody onto meaning - the case of information structure in American English. *Lang. Cogn. Neurosci.* 34, 841–860. doi: 10.1080/23273798.2019.1587482

Schneider, K., and Lintfert, B. (2003). "Categorical perception of boundary tones in German," in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, 631–634.

Smith, R., and Rathcke, T. (2020). Dialectal phonology constrains the phonetics of prominence. *J. Phon.* 78, 100934. doi: 10.1016/j.wocn.2019.100934

Sóskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *J. Phon.* 84, 101017. doi: 10.1016/j.wocn.2020.101017

Truckenbrodt, H. (2007). "Upstep on edge tones and on nuclear accents," in *Tones and Tunes. Volume 2: Experimental Studies in Word and Sentence Prosody*, eds C. Gussenhoven and T. Riad (Berlin: Mouton de Gruyter), 165–172.

Turk, A. E., Nakai, S., and Sugahara, M. (2006). "Acoustic segment durations in prosodic research: a practical guide," in *Methods in Empirical Prosody Research*, eds S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, and J. Schließer (Berlin; New York, NY: Walter de Gruyter), 1–27.

Ulbrich, C. (2005). *Phonetische Untersuchungen zur Prosodie der Standardvarietäten des Deutschen in der Bundesrepublik Deutschland, in der Schweiz und in Österreich [Phonetic studies on prosody in the Standard varieties in Germany, Switzerland, and Austria]*. Frankfurt am Main: Peter Lang.

van Rij, J., Hendriks, P., an Rijn,1, H., Baayen, R. H., and Wood, Simon, N. (2019). Analyzing the time course of pupillometric data. *Trends Hearing* 23, 1–22. doi: 10.1177/2331216519832483

van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (2017). *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs. R Package Version, 2.*

Waterman, J. C. (1991/1966). *A History of the German Language*. Long Grove, IL: Waveland Press Inc. (by arrangement with University of Washington Press).

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. *J. Phon.* 70, 86–116. doi: 10.1016/j.wocn.2018.03.002

Wochner, D. (2021). *Prosody meets Pragmatics: rhetorical questions, exclamatives and assertions* (PhD thesis). University of Konstanz, submitted to the Department of Linguistics, Konstanz.

Wochner, D., and Dehé, N. (2018). "Prosody meets pragmatics: a production study on German verb-first sentences, in *Proceedings of the 9th International Conference on Speech Prosody* (Poznan). 418–422.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction With R*. Boca Raton, MA: CRC Press.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B* 73, 3–36. doi: 10.1111/j.1467-9868.2010.00749.x

Wood, S. N. (2017). *Generalized Additive Models: An Introduction With R, 2nd Edn*. Boca Raton, MA: CRC Press.

Xu, Y. (2013). "ProsodyPro - a tool for large-scale systematic prosody analysis," in *Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)* (Aix-en-Provence), 7–10.

Zahner, K., Schönhuber, M., Grijzenhout, J., and Braun, B. (2016). "Konstanz prosodically annotated infant-directed speech corpus (KIDS corpus)," in *Proceedings of the 8th International Conference on Speech Prosody* (Boston, MA), 562–566.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Psychology

ORIGINAL RESEARCH
published: 20 April 2022
doi: 10.3389/fpsyg.2022.840291

Perceptual Cue Weighting Is Influenced by the Listener's Gender and Subjective Evaluations of the Speaker: The Case of English Stop Voicing

*Alan C. L. Yu**

Chicago Phonology Laboratory, Department of Linguistics, University of Chicago, Chicago, IL, United States

Speech categories are defined by multiple acoustic dimensions and their boundaries are generally fuzzy and ambiguous in part because listeners often give differential weighting to these cue dimensions during phonetic categorization. This study explored how a listener's perception of a speaker's socio-indexical and personality characteristics influences the listener's perceptual cue weighting. In a matched-guise study, three groups of listeners classified a series of gender-neutral /b/-/p/ continua that vary in VOT and F0 at the onset of the following vowel. Listeners were assigned to one of three prompt conditions (i.e., a visually male talker, a visually female talker, or audio-only) and rated the talker in terms of vocal (and facial, in the visual prompt conditions) gender prototypicality, attractiveness, friendliness, confidence, trustworthiness, and gayness. Male listeners and listeners who saw a male face showed less reliance on VOT compared to listeners in the other conditions. Listeners' visual evaluation of the talker also affected their weighting of VOT and onset F0 cues, although the effects of facial impressions differ depending on the gender of the listener. The results demonstrate that individual differences in perceptual cue weighting are modulated by the listener's gender and his/her subjective evaluation of the talker. These findings lend support for exemplar-based models of speech perception and production where socio-indexical features are encoded as a part of the episodic traces in the listeners' mental lexicon. This study also shed light on the relationship between individual variation in cue weighting and community-level sound change by demonstrating that VOT and onset F0 co-variation in North American English has acquired a certain degree of socio-indexical significance.

Keywords: speech perception, sociophonetics, cue weighting, English stop voicing, paralinguistic information, gender, personality traits, subjective evaluations

Edited by:
Eleanor Chodroff,
University of York, United Kingdom

Reviewed by:
Jessamyn Schertz,
University of Toronto Mississauga, Canada
Matthew Lehet,
Michigan State University, United States

****Correspondence:***
Alan C. L. Yu
aclyu@uchicago.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *21 December 2021*
Accepted: *28 February 2022*
Published: *20 April 2022*

Citation:
Yu ACL (2022) Perceptual Cue Weighting Is Influenced by the Listener's Gender and Subjective Evaluations of the Speaker: The Case of English Stop Voicing. Front. Psychol. 13:840291. doi: 10.3389/fpsyg.2022.840291

1. INTRODUCTION

Speech categories are defined by multiple acoustic dimensions. The acoustic and perceptual boundaries between speech categories are generally fuzzy in part because both speakers and listeners often give differential weighting to these dimensions in production and in perception. This study investigates how and why listeners may vary their perceptual weight of cues, with special focus on the voicing contrast in English initial stop, a prime example of the type of category fuzziness mentioned above.

The distinction between voiced and voiceless stops in English can be conveyed by as many as sixteen cues (Lisker, 1986). Voice onset time (VOT) and fundamental frequency (F0) at the onset of the following vowel, for example, have often been observed to co-vary in English word-initial plosives, with phonologically voiceless plosives followed by raised F0 at the vocalic onset, while phonologically voiced plosives (which are canonically realized with zero to weakly positive VOT) followed by lowered onset F0. Listeners have been found to be very sensitive to this type of onset F0 perturbations. Many studies have demonstrated that listeners can adjust their categorization of synthetic or digitally manipulated natural speech varying perceptually from voiced to voiceless stops depending on the F0 of the following vowel. Stimuli with lower F0's are more likely to be categorized as voiced whereas stimuli with higher F0's (but with otherwise identical acoustic characteristics) tend to be labeled as voiceless. A particular intriguing aspect of perceptual cue weighting, including the relative perceptual importance of VOT and onset F0 cues for stop voicing perception, is that it is not only language-specific (Schertz et al., 2015), there is also great individual-specific variation (Shultz et al., 2012; Clayards, 2018a) and such variation has been shown to be systematic across individuals (Idemaru et al., 2012; Schertz et al., 2015; Ou and Yu, 2021; Ou et al., 2021). What factors govern the differences in cue weighting between individuals remain under-investigated. In light of recent work that suggests socio-indexical information can influence speech perception, this study aims to elucidate the effects of listener's subjective evaluation of the talker on perceptual cue weighting, in particular the weighting between VOT and F0 for the English stop voicing contrast. The next section reviews important background information that motivates the current study. Section 3 introduces the experimental setup, followed by a discussion of the results in Section 4. Section 5 summarizes the study, discussing the implication of the present study for cue weighting research and for sound change theories.

2. BACKGROUND

2.1. Sources of Individual Variability in Cue Weighting

Researchers have attempted to explain onset F0 perturbations as a reflex of aerodynamic (Ladefoged, 1967) and/or articulatory (Halle and Stevens, 1971; Ohala, 1973; Löfqvist et al., 1989) byproducts of stop voicing production. More recently, many have argued that onset F0 perturbations in English is actively controlled by speakers, perhaps to enhance this specific phonological contrast (Kingston and Diehl, 1994; Keyser and Stevens, 2006; Kingston, 2007; Solé, 2007; Hanson, 2009). Specifically, studies of onset F0 perturbations have found that the extent of onset F0 perturbations is not only language-specific (Hombert et al., 1979; Francis et al., 2006; Dmitrieva et al., 2015), but it can also vary quite extensively across individuals (Shultz et al., 2012; Chodroff and Wilson, 2018; Clayards, 2018b). Also consistent with the controlled phonetic interpretation of onset F0 perturbations is the context-dependency of onset F0 perturbations. Hanson (2009) observed that, in high pitch environment within a given speaker's F0 range, F0 is greatly increased following voiceless obstruents relative to a baseline F0, but not following voiced ones. In low-pitch environment, F0 is slightly increased relative to a baseline following all obstruents. She interpreted this difference in onset F0 perturbations in high vs. low pitch contexts as an indication of contrast enhancement since VOT is less distinctive in high pitch context than in low ones (see also Kirby et al., 2020). Echoing the variability observed in the production domain, the perceptual importance of these cues has also been found to be quite variable. Not only do listeners adjust their cue reliance in different contexts (Haggard et al., 1981; Repp, 1982) and when they are under different cognitive loads (Gordon et al., 1993), many studies have also found a trading relationship between the perceptual weightings of the VOT and F0 cues across listeners. Specifically, English listeners who rely on the VOT cue are found to rely less on the onset F0 cue, indicating a trading relation between these cues (Kapnoula, 2016; Kapnoula et al., 2017; Ou et al., 2021). Crucially, individual differences in cue weight have been shown to be stable across time (Idemaru et al., 2012; Schertz et al., 2015; Kapnoula, 2016) and across contrasts (Clayards, 2018a; Ou et al., 2021).

What factors govern the differences in cue weighting between individuals remains a largely unanswered question. Variability might stem from differences in individual perceptual experiences, as evidenced by perceptual learning experimental results showing that listeners can adjust their perceptual cue weights in accordance with the cue distributions in the exposure stimuli (e.g., Francis et al., 2008; Lehet and Holt, 2017; Zhang and Holt, 2018). An experience-driven approach to individual variation in cue weighting seems insufficient, however, given the often elusive mapping between perception and production of cue weights. While phonetic imitation studies have found that some speakers may adjust their VOT production when exposed to a model talker with a different VOT distribution, results from studies that look at direct correspondences between perceptual and production cue weighting have been mixed. Shultz et al. (2012), for example, investigated the use of VOT and F0 in producing and perceiving the English stop voicing contrast. While they found a significant negative correlation between VOT and onset F0 in production (see also Dmitrieva et al., 2015; but see Chodroff and Wilson, 2018; Clayards, 2018b, who did not find such a significant correlation in production), but did not find a significant correlation in the corresponding perceptual weights. They also did not find a significant correlation between perceptual and production cue weights. Schertz et al. (2015) examined native Korean speakers' perception and production of stop contrasts in their native language (L1) and second language (L2, English) and found that Korean listeners use different cue weighting strategies for both Korean and English stop voicing contrasts. They identified three general patterns among the L1 Korean listeners. The so-called "VOT group" classified stimuli with a long VOT as voiceless and a short VOT as voiced irrespective of F0, while the "F0 group" classified stimuli with high F0 as voiceless and low F0 as voiced irrespective of VOT. Finally, the "VOT+F0" group classified only stimuli with high

F0 and long VOT as voiceless and all other stimuli as voiced. Of particular interest is that differences in perception were not predicted by individual variation in production patterns (Schertz et al., 2015). Such findings are problematic for input-driven accounts of speech categorization and cue weight setting that assume a tight perception-production loop since such models assume that speech classification and cue distributions are either estimated directly from the input (Pierrehumbert, 2002; Kronrod et al., 2016) or as a function of both the statistics of the input and the history of the learning system (Toscano and McMurray, 2010). Findings like those reported in Schertz et al. (2015) suggest that there might be other factors that mediate listeners' perceptual experiences that render the mapping between perception and production imperfect.

The fact that individual variation in cue weights is systematic across individuals (Idemaru et al., 2012; Schertz et al., 2015; Kapnoula, 2016; Ou et al., 2021) and not contrast-specific (Clayards, 2018a; Ou et al., 2021) suggests that such individual variability might stem from the influence of some general cognitive mechanism that modulates cue weights. Kong and Edwards (2016), for example, tied individual variability in perceptual cue trading between VOT and F0 to categorization gradience. Specifically, they found that listeners who exhibited a more gradient response pattern in a visual analog task also showed more sensitivity to F0 in an anticipatory eye movement task. Individual variability in categorization gradience might in turn stem from individual differences in neural encoding of the speech signal at the subcortical and cortical levels (Ou and Yu, 2021). Individual differences in cue weighting might also stem from individual variation in speech processing strategies. In their investigation of secondary cue weighting in two sets of English contrasts (/b/ vs. /p/ and /i/ vs. /ɪ/) using an eye-tracking paradigm, Ou et al. (2021) found that individuals who integrate secondary cues more extensively during processing are more likely to utilize a buffer processing strategy, suggesting a delayed reaction to the early-arriving cue until all relevant cues are available may facilitate the integration of multiple cues in the signal.

Another important source of individual variability that has yet to be explored in cue weighting research is the influence of socio-indexical and paralinguistic information on speech perception. The idea that socio-indexical information influences speech perception is not new *per se*. Strand (2000), for example, found that words are processed more quickly when the pitch of the talker is typical of his/her gender. Hay et al. (2006) investigated a case of merger in progress in New Zealand English (i.e., the merger of diphthongs /iɑ/ and /eɑ/) and found that the age and social class of the talker biased the listeners' perception of otherwise identical auditory stimuli. Staum Casasanto (2010) investigated the effect of listeners' experience with an ethnic dialect has on t/d deletion and found that listeners use social information about speakers (i.e., whether the face of the talker is Black or White) to make inferences about speech. Phonetic imitation/convergence research has also pointed to a significant influence of socio-indexical information on speech perception since whatever production adjustments in the direction of the model talker or interlocutors must presumably be perceptually detected in the first place. For example, Babel (2012) investigated the imitation of vowels in a lexical shadowing task and found that the degree to which vowels were imitated was subtly affected by how attractive the talker was rated by the participants; the listeners were given either no image, or saw either a Black talker or a White talker. Yu et al. (2013) investigated the imitation of VOT and found that the extent of phonetic imitation is modulated by the participant's subjective attitude toward the model talker, the participant's personality trait of openness, and the autistic-like trait associated with attention switching.

Evidence of socio-indexical information influencing speech perception and phonetic imitation/convergence lends support for models of speech perception and production where socio-indexical features are encoded as a part of the episodic traces in the listeners' mental lexicon and the activation of socio-indexical information will result in the activation of episodic traces that are consistent with, or linked to, the social category or feature (e.g., Sumner et al., 2014; Babel and Russell, 2015; McGowan, 2015). Thus, when a talker is perceived to be of a particular gender or has certain personality features such as being attractive or friendly, the listener's perception will be primed to interpret the speech signal in ways that are consistent with the social expectation (see also similar accounts under the rational exemplar-based model or the ideal adapter framework Kleinschmidt and Jaeger, 2015; Myslin and Levy, 2016; Kleinschmidt et al., 2018).

2.2. The Socio-Indexical and Paralinguistic Characteristics of VOT and F0

In addition to the fact that the likelihood of VOT imitation can be modulated by a listener's subjective evaluation of the talker, various converging evidence further lends support to the idea that the socio-indexical and paralinguistic characteristics of the talker may influence listeners' perception of English stop voicing. To begin with, Swartz (1992) reported that females have longer VOTs than males (see also Ryalls et al., 1997; Whiteside and Irving, 1997, 1998; Koenig, 2000; Whiteside and Marshall, 2001; Whiteside et al., 2004b; Robb et al., 2005; cf. Morris et al., 2008). Some attributed this gender-based VOT difference to anatomical differences in phonatory apparatus between genders, such as men's wider supraglottic space and women's shorter and stiffer vocal folds (e.g., Swartz, 1992; Whiteside and Irving, 1997, 1998; Koenig, 2000; Oh, 2011), others hypothesized that the pattern might stem from voicing contrast optimization in female speech (Whiteside and Irving, 1998). The physiological explanation is undermined by the fact that the same gender difference is not uniformly observed cross-linguistically (Oh, 2011; Lundeborg et al., 2012; Li, 2013; Reddy et al., 2013; Peng et al., 2014), further pointing to the potential socio-indexical relevance of this gender-based VOT difference in English. VOT is also reported to vary according to women's menstrual cycle; women who are at their reproductive peaks have longer VOTs than those at their lowest fertility levels (Whiteside et al., 2004a; Wadnerkar et al., 2006). Since women at the reproductive peaks of their menstral cycle are rated as more vocally attractive, Babel et al. (2014) reasoned that the increase in VOT, which could increase clarity in stop voicing contrasts, might also influence attractiveness judgments.

It should be noted that, in clear speech, a mode of speaking that is associated with increased articulatory efforts, VOT for voiceless stops in English has also been found to be lengthened while the VOT for voiced stops remain unchanged (Smiljanić and Bradlow, 2008).

F0 also carries a wealth of social information about a person. To begin with, pitch, one of the most perceptually salient feature of human voice (Banse and Scherer, 1996), is about half as high in men as it is in women (Titze, 2000). The pitch of voice is inversely correlated with perceived dominance; the lower the voice pitch, the greater the perceived dominance (Puts et al., 2006). Adjusted for the effects of sex and age, Stern et al. (2021) found that participants with lower voice pitch self-report as lower on neuroticism, but higher on dominance, extraversion, and openness to experience, as well as more unrestricted on sociosexual orientation, sociosexual behavior, sociosexual attitudes, and sociosexual desire. Paralinguistic intonational meanings have been argued to be grounded in terms of the Frequency Code (Ohala, 1983, 1984; Chen et al., 2004a), which exploits the link between larynx size and vibration rates of the vocal cords for the expression of power relations, and the Effort Code (Gussenhoven, 2002), which refers to the positive correlation between articulatory efforts and articulatory precision (de Jong, 1995). Specifically, higher pitch has more affective interpretations, which include "uncertain", "feminine", "submission", "friendly", "polite", and "vulnerable", while lower pitch has "certain", "masculine", "dominant", "confident", "protective", and "aggressive" interpretations (Gussenhoven, 2002; Chen et al., 2004a,b). Greater pitch excursion is also associated with informational interpretations such as "emphatic" and "significant" and affective interpretation of "surprised" and "agitated" and even "obliging" (Gussenhoven, 2002).

Perceived sexual orientation has also been associated with variation in VOT and F0. More-gay sounding men, for example, has been found to produce stop consonants with longer voice-onset times than less-gay sounding men (Smyth and Rogers, 2002). Gayness ratings were strongly correlated with independently made judgments of perceived intonational variability, even though mean F0 and F0 variability did not predict gayness ratings (Smyth et al., 2003). In particular, the voices that were rated as gay-sounding by one group of listeners were rated by an independent group of listeners as having greater F0 modulation; conversely, listeners were more likely to falsely judge a voice as having greater F0 modulation if that voice had been judged by an independent group to be gay-sounding.

As noted above, the difference in onset F0 after voiced and voiceless stops (onset F0 perturbations), is found to be greater in higher global F0 contexts than in lower ones (Hanson, 2009; Kirby et al., 2020), we hypothesize that listeners might make use of such an association when processing onset F0 perturbations produced by talkers of different genders or talkers associated with certain paralinguistic features given their different F0 profiles. There is some suggestive evidence to support this hypothesis. Zhang and Holt (2018), for example, found that global F0 differences can influence stop voicing categorization, but this F0 effect is more apparent when the talker is perceived to be female. Specifically, in a series of perceptual learning experiments, they recruited two groups of listeners, half presented with high vs. mid F0 global contours (the high F0 range group), while the other half with the mid and low F0 contours (the low F0 range group). They found significant differences in voicing responses depending on the global F0 height, with higher F0 contours associated with more voiceless response than lower F0 contours. Crucially, in two followup studies, they manipulated the perceived gender of the talker(s) acoustically (via changes in the formants of the stimuli) and visually. For the "female" voice stimuli (i.e., high F0 range stimuli with female-like formant values), listeners showed a difference in /p/ response according to the high or low global F0 profile of the stimuli within the "female" global pitch range, but no comparable global F0-dependent /p/ response difference was observed with the "male" stimuli (i.e., low F0 range stimuli with male-like formant values). These findings suggest that the perceived gender of the talker influences the effects of global F0 have on English stop voicing perception.

To be sure, Zhang and Holt's study did not address onset F0 perturbations specifically as the F0 differences are not localized to the onset of the vowel. Thus, it remains unclear if the gender of the talker would influence the effect of onset F0 perturbation on stop voicing perception. Also, since the participants' gender evaluation of the talkers was not examined, it is difficult to ascertain whether the participants' perception of the talker gender matched the expectation of the experimenter. Finally, their perceived gender findings were based on a within-subject design where listeners were presented with both "male" (i.e., low F0 range) and "female" (i.e., high F0 range) stimuli within the same block. This design raises the possibility that the different rates of /p/ responses across the perceived gender conditions might come about as a result of a contrast effect. That is, listeners only adjusted their expectation when they encountered both high and low F0 talkers, but not when they listened to a single talker with small variation in global F0.

The present study built on these earlier findings and examined whether the perceived gender and the listener's impression of the talker's facial and vocal features influence listeners' perception of word-initial voiced and voiceless stops in English using a matched-guise design (Lambert et al., 1960; Zahn and Hopper, 1985). In particular, three groups of listeners classified the same set of acoustic stimuli. Two groups were given a visual prompt of the talker: one group of participants in the visual prompt condition was presented with an image of a prototypical male and the other group with the image of a prototypical female. Given that previous studies have shown that rapid evaluative inferences based solely on facial and vocal information can exert a significant influence on the perceiver/listener behavior [e.g., sales (Jacob et al., 2011), stock market returns (Mayew and Venkatachalam, 2012), wage penalty (Grogger, 2011; Rickford et al., 2015), election outcomes (Todorov et al., 2005; Klofstad, 2017), housing market interactions (Purnell et al., 1999), likelihood of vowel imitation (Babel, 2012), and language processing speed (Staum Casasanto, 2010)], we hypothesize that listeners would adjust their perceptual cue weights if they are aware of the association between the VOT/onset F0 covariation on the one hand and the socio-indexical and personality characteristics on the other. We also aimed to examine whether facial and vocal impressions exert similar influences on the listener's cue weighting. Previous literature reported conflicting findings concerning the strength of facial and vocal impressions. While some studies reported stronger effects of facial impression over vocal impressions (e.g., Klofstad, 2017; Hou and Ye, 2019), others found the opposite tendency (Schroeder and Epley, 2015).

3. METHODS AND MATERIALS

3.1. Participants

304 native speakers of American English were recruited to participate in this study on Prolific (https://www.prolific.co/), a crowd-sourcing platform for online studies that, in addition to confirming the identity of each participant, gathers extensive self-reported demographic information from each participant for prescreening purposes. Participation in this study was limited to individuals who reported being 18–40 years old, native speakers of English, residents of the United States, right-hand dominant, with no history of hearing, language, neurological, or mental disorders. In the end, a total of 237 participants' responses were analyzed. Sixty-seven were excluded from the study due to failure to pass the headphone screen ($N = 23$) or failure to meet compliance checks (i.e., not a native speaker of English, participated in more than one prompt condition, and/or have a history of one or more of the following: speech/hearing/language disorders, dyslexia, autism, substance dependence, stroke, mental retardation, traumatic brain injury with greater than 1 h loss of consciousness, multiple sclerosis, Parkinson's disease, Alzheimer's disease, Huntington's disease, schizophrenia, bipolar, ADHD, or current major depression; $N = 44$). This attrition rate is consistent with other web-based studies (Thomas and Clifford, 2017; Woods et al., 2017; Brown et al., 2018; Giovannone and Theodore, 2021).

The cohort is roughly gender-balanced in each prompt condition. **Table 1** provides a detailed gender breakdown of the number of participants within each condition. The median age is 25 (Mean = 26.62, SD = 6.37). Additionally, 87 participants reported having some musical training and 128 reported speaking or having studied another language other than English. The participants were paid $2 for their participation in the study; the study lasted, on average, around 10 min.

3.2. Stimuli

In order to create a gender-neutral voice suitable for the study, a gender prototypicality rating task was conducted. The stimuli, based on recordings of /b/ "bear" and /p/ "pear" produced by a male native speaker of American English, were generated by modifying the recordings in terms of Formant Shift and Pitch Shift, using a custom-written script from Xu et al. (2013) that applied the "Change Gender" function in the Praat program (https://doi.org/10.1371/journal.pone.0062397.s002). In total, 25 stimuli were prepared, that is, 5 formant shift ratios (0.8, 0.9, 1, 1.1, 1.2) × 5 pitch shifts (−5, −4, 0, 4, 5). The "Change Gender" function in Praat shifts formant frequencies as a ratio of the original sound via manipulation of sampling frequency. The manipulation shifted the formant frequencies in the original speech token toward a more exaggerated female voice (formant shift ratios of 1.1 and 1.2) or toward a more male voice (formant shift ratios of 0.8 and 0.9). Prior to creating the different voices, the F0 of the original speech token was first resynthesized to have a flat F0 contour at 154 Hz. Ten participants, recruited on Prolific, listened to all 25 speech tokens in a randomized order to decide how male- or female-sounding a voice is by adjusting a sliding scale that ranges from prototypical female to prototypical male. The polarity of the scale was counter-balanced across participants. The voice with formant shift ratio of 1.1 and F0 at 154 Hz was chosen as the stimuli for the main experiment because it was rated most neutral (i.e., the midpoint of the gender prototypicality scale) most often and most consistently (mean = 49.2, sd = 5.5).

A 7-step /b/ to /p/ VOT continuum was created out of the selected gender-neutral voice "bear"/"pear" tokens by cross-splicing aspiration from the naturally produced voiceless bilabial /p/ in "pear" to the voiced bilabial /b/ in "bear" at 7 ms increments using the custom script described in Winn (2020). Each step on the continuum was given one of two F0 contours where F0 began at either 134 or 174 Hz and fell (or rose) linearly until 154 Hz at the 75 ms from vowel onset. The 7 (VOT) × 2 (F0 target) design yielded 14 distinct items. The intensity of all stimuli was normalized to the same level.

3.3. Procedure

Both the gender prototypicality rating task and the main experiment were hosted on Qualtrics. To ensure that participants were wearing headphones, all participants first passed a headphone screen developed by Woods et al. (2017). In this task, listeners judge which sound in a series of three pure tones is the quietest, with one sound presented out of phase on the stereo channels. This task is designed to be easy when the participant is wearing headphones or earbuds, but extremely challenging over loud speakers due to phase-cancellation. If participants did not correctly pass 5 out of 6 trials, they were reminded to wear headphones and asked to repeat the task. If they failed the headphone check twice, they were asked to return the task in order to receive partial compensation for their efforts.

After the headphone check, participants completed a short demographic survey to gather any information not made available through Prolific. This is followed by either one or two first impression rating task(s) depending on the prompt condition. Participants were randomly assigned to either a condition with visual prompt or one without. Those in the visual prompt conditions were shown either a prototypical male or prototypical female face selected from the Chicago Face Database (Ma et al., 2015). The specific faces can be found in the **Supplementary Data**. Participants in the visual prompt conditions completed two first impression rating tasks. The first rating task asked the participant to rate the talker faces in terms of their gender-prototypicality, the attractiveness, friendliness, confidence, trustworthiness and whether the individual looked gay. These personality attributes were selected in part based on previous research on listener's perceptual evaluations of linguistic variables (Eckert, 2008; Campbell-Kibler, 2009, 2010; McAleer et al., 2014) as well as their the associations between the specific

TABLE 1 | Mean ratings (and standard deviations in parentheses) for perceived vocal gender, attractiveness, friendliness, confidence, trustworthiness, and gayness in the two face conditions arranged by the gender of the participants.

Condition	Listener	*N*	Gender	Attractive	Friendly	Confident	Trustworthy	Gay
AudioOnly	Female	39	67 (16)	27 (17)	57 (22)	46 (20)	49 (17)	52 (23)
AudioOnly	Male	47	67 (20)	30 (22)	54 (23)	44 (23)	50 (22)	48 (27)
Female	Female	36	63 (15)	30 (18)	47 (21)	45 (18)	48 (19)	58 (22)
Female	Male	40	66 (18)	31 (20)	47 (18)	48 (19)	49 (19)	50 (24)
Male	Female	35	42 (16)	44 (20)	62 (16)	55 (16)	54 (14)	57 (11)
Male	Male	40	44 (22)	47 (20)	60 (17)	47 (18)	55 (15)	57 (22)

attributes and the two phonetic dimensions targeted in this study as reviewed in the Introduction.

The participants then listened to the voice of the talker and rated the voice on the same attributes as the visual impression survey. The stimulus heard was a recording of the word "bear" with zero VOT (i.e., step 1 of the VOT continuum) with a rising F0 onset. Participants in the "audio-only" condition completed only the vocal impression rating task.

Following the rating task(s), the participants were asked to listen to the target stimuli and determine whether they heard the word "bear" or "pear" by clicking on the corresponding picture. Each participant classified 112 stimuli (7 VOT steps × 2 F0 targets × 8 blocks). The trials were split into eight blocks, each consisted of the fourteen target stimuli randomly ordered within each block. The instructions (and the talker image in the visual prompt conditions) were repeated at the beginning of each block. The positions of the response pictures were counterbalanced across blocks. To encourage the participant to stay alert, the participant completed a ten-question Short Autism Spectrum Quotient (Allison et al., 2012) after four blocks of the categorization task. Following the completion of all eight blocks of the categorization task, participants completed the headphone screen again before exiting the task.

3.4. Predictions

Before diving into the results, it is worth laying out some *a priori* predictions based on the literature reviewed above. Concerning gender-based differences, we advance three potential hypotheses. As alluded to in Section 2, from the perspective of episodic/exemplar-based models of speech perception and production, when a talker is perceived to be of a particular gender or has certain paralinguistic features such as being attractive or friendly, the activation of the relevant socio-indexical/paralinguistic information will result in the activation of episodic traces that are consistent with, or linked to, the social category or paralinguistic feature (e.g., Sumner et al., 2014; Babel and Russell, 2015; McGowan, 2015). This means that the listener's perception will be primed to interpret the speech signal in ways that are consistent with the social expectation. Specifically, given that VOT is less distinct between voiced and voiceless stops in word-initial position in males compared to females, we expect listeners to be sensitive to this gender difference in VOT realization and exhibit less reliance on the VOT cue when listening to a talker who is perceived to be male than when the talker is perceived to be female. Assuming that there is a perception-production loop, where stored perceptual experiences are weighted by social and attentional factors and such perceptual exemplars are drawn upon to generate production targets (Pierrehumbert, 2002), we expect that male listeners may also rely less on the VOT cue than female listeners, if male listeners mirror the production tendencies of male speakers. Furthermore, to the extent that the perceptual cue weights for VOT and F0 are in a trading relation, we expect listeners who assign less weight to the VOT would rely more on F0 in stop voicing classification.

Turning to potential effects of socio-indexical and paralinguistic information on the relative cue weighting between VOT and F0, recall that, within a given talker's F0 range, onset F0 perturbations are larger when the global F0 environment is high and VOT for voiceless stops are shorter. To the extent that femininity, friendliness, trustworthiness are associated with higher overall F0 and more dynamic F0 excursion, we hypothesize that listeners may rely more on F0 information and less on VOT information for stop voicing perception when the talker is thought to be associated with those personality characteristics. To the extent that attractive, confident, or gay-sounding voices are associated with greater VOT differences between voiced and voiceless stops, we expect listeners to rely more on the VOT cue when listening to talkers who are rated as more attractive and confident.

4. RESULTS

We begin the presentation of the results of the study by first examining the effects of vocal impressions on the identification of stop voicing in English in Section 4.1 since visual information is only relevant in two of the three prompt conditions. Section 4.2 presents findings from the visual prompt conditions.

4.1. Results From All Prompt Conditions

Before introducing the first regression model, **Table 1** summarizes the vocal impression ratings. Several aspects of the rating data are noteworthy. Not only is there great variability in how the participants rated the talker's vocal gender prototypicality, as illustrated in **Figure 1**, there is also a great deal of variation in ratings for each dimension, as well as variation in how the attributes relate to each other. Specifically, there are strong positive correlations between

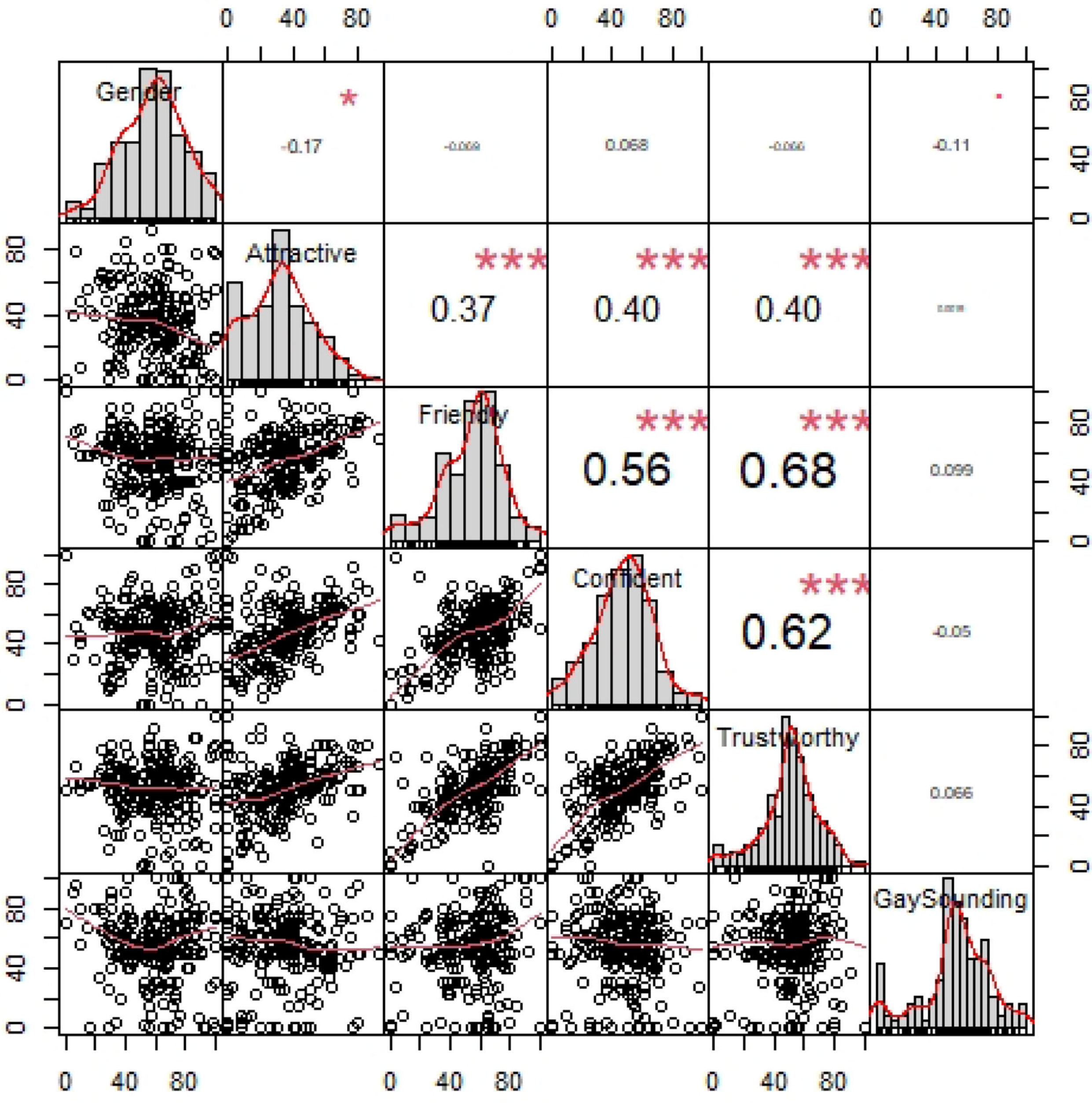

FIGURE 1 | Correlations between the ratings across different vocal attributes. Each point corresponds to the ratings of a participant. ***$p < 0.001$; *$p < 0.05$.

Attractiveness, Friendliness, Confidence, and Trustworthy and a negative correlation between Gender and Friendliness, as seen **Figure 1**.

4.1.1. Principal Component Analysis of the Vocal Impression Ratings

Given the highly correlated nature of some of the vocal impression attributes, in an effort to reduce the dimensionality of the mapping between vocal impressions and perceptual responses, rather than analyzing the vocal impression ratings individually, an integrated cue-combination approach was taken such that the vocal impression ratings were first submitted to a principal component analysis (PCA) to obtain linear combinations of these vocal impression ratings that would capture the maximum variation. The specifics of the PCA are as follows: the vocal impression ratings, which were z-scored, were analyzed using the `prcomp()` function in `R`, which performs a principal component analysis on a given data matrix; principal components with an eigenvalue greater than 1 were selected for the regression analysis (Kaiser, 1961).

TABLE 2 | The cumulative proportion of variance accounted for and loadings from the PCA of the vocal impression ratings.

	PC1	PC2	PC3	PC4	PC5	PC6
Vocal gender	−0.07	0.73	0.55	0.36	0.18	−0.05
Vocal attractiveness	0.40	−0.13	−0.36	0.82	0.14	−0.00
Vocal friendliness	0.53	−0.02	0.15	−0.29	0.55	0.56
Vocal confidence	0.51	0.25	0.05	−0.05	−0.78	0.26
Vocal trustworthiness	0.54	0.03	0.09	−0.25	0.12	−0.79
Vocal gay-sounding	0.05	−0.63	0.73	0.23	−0.14	0.01
Standard deviation	1.60	1.07	0.97	0.80	0.64	0.56
Proportion of variation	0.42	0.19	0.16	0.11	0.07	0.05
Cumulative proportion	0.42	0.62	0.77	0.88	0.95	1.00

The relative weighting and proportion of variance for each component for the vocal attributes are summarized in **Table 2**. The optimal linear combination (PC1), which accounts for 42% of the variance, and the 2nd component (PC2), which accounts for 19% of the variance, were selected as independent variables for the analysis below; the first two components collectively account for around 62% of the variance. PC1 has strong loadings for vocal attractiveness, friendliness, confidence, and trustworthiness, which can be characterized as "vocal appeal". PC2, on the other hand, is dominated by voice gender, confidence, and gay-sounding, which might be characterized as gender stereotypicality.

4.1.2. Model 1

Listeners' responses (/b/ = 0, /p/ = 1) were modeled with logistic mixed effects regressions using the `glmer()` function in the lme4 package (Bates et al., 2015) in R. The fixed effect predictors included in the model were trial block (BLOCK: 1–8), VOT continuum step (VOT: 1-7), onset F0 (F0: High or Low), prompt CONDITIONs (Helmert-coded: contrast 1 = audio only vs. visual prompt; contrast 2 = Male Face vs. Female Face), and the two PCs of the vocal impression ratings. The model also included the participant's GENDER (Male vs. Female) as a between-subject factor given that effects of facial and vocal impressions on listener behavior have been found to be gender-differentiated (Babel, 2012; Chen et al., 2016). All continuous variables (i.e., BLOCK, VOT, PC1, and PC2) were z-scored. Unless otherwise specified, categorical variables were sum-coded. The model also included all possible interactions between the fixed effects predictors other than BLOCK as well as by-subject random intercepts and by-subject random slopes for BLOCK, VOT, and F0, as well as the interaction between VOT and F0.

Model selection started with the maximal model with all possible interactions between fixed factors (the PCs of the vocal attributes did not interact with each other, however) as well as the random intercepts and slopes, and proceeded by comparing between models with and without the inclusion of a fixed/random factor and/or interaction. Predictors that do not improve model-likelihood significantly were dropped. In the end, neither PC1 nor PC2 of the vocal attributes was retained following this model selection procedure. The complete model in lme4 format is: `Response (pear = 1) ∼ BLOCK + VOT * F0 + VOT * GENDER + VOT * CONDITION + (1 + BLOCK + VOT * F0|PARTICIPANT)`.

TABLE 3 | Estimates for all predictors in Model 1.

	Model 1
Intercept	−0.35 (0.08)***
VOT	3.13 (0.13)***
F0	0.97 (0.05)***
Gender	−0.13 (0.07)
$\text{Condition}_{A/AV}$	−0.03 (0.15)
$\text{Condition}_{M/F}$	0.02 (0.18)
Block	0.09 (0.03)**
VOT:F0	−0.53 (0.04)***
VOT:Gender	0.23 (0.11)*
$\text{VOT:Condition}_{A/AV}$	0.45 (0.23)
$\text{VOT:Condition}_{M/F}$	−0.67 (0.28)*
AIC	18431.06
BIC	18643.91
Log Likelihood	−9189.53
Num. obs.	26544
Num. groups: Participant	237
Var: Participant Intercept	1.36
Var: Participant Block	0.08
Var: Participant F0	0.34
Var: Participant VOT	3.64
Var: Participant F0:VOT	0.11

Gender refers to the gender of the participant. $\text{Condition}_{A/AV}$, audio only vs. visual prompt; $\text{Condition}_{M/F}$, Male Face vs. Female Face.

****$p < 0.001$; **$p < 0.01$; *$p < 0.05$.*

A summary of the first regression model, Model 1, appears in **Table 3**. As expected, VOT is a significant predictor ($\beta = 3.13$, $z = 23.66$, $p < 0.001$) as well as onset F0 ($\beta = 0.97$, $z = 21.57$, $p < 0.001$), suggesting that /p/ responses are more likely when VOT is longer and when the onset F0 is higher. There is also a significant interaction between VOT and onset F0 ($\beta = -0.53$, $z = -14.26$, $p < 0.001$), suggesting that the likelihood of a /p/ response along the VOT continuum varies depending on the onset F0. Visual inspection of **Figure 2** shows that the F0 effect on /p/-response is strongest within the VOT range where VOT is not the most informative cue (i.e., the middle of the VOT continuum). There is also a significant effect of BLOCK ($\beta = 0.09$, $z = 3.21$, $p < 0.01$), suggesting that the participants are more likely to respond /p/ as the experiment progressed.

There is a significant interaction between VOT and $\text{CONDITION}_{M/F}$ ($\beta = -0.67$, $z = -2.43$, $p < 0.05$). As illustrated in **Figure 3**, the classification function along the VOT dimension in the male face condition is shallower than in the female face condition. Specifically, the listeners in the male face condition are less likely to hear /p/ toward the /p/ end of the VOT continuum than those in the female face condition, suggesting that listeners in the male talker condition are less reliant on VOT as a cue for determining stop voicing. A separate model with the CONDITION treatment-coded with the audio-only

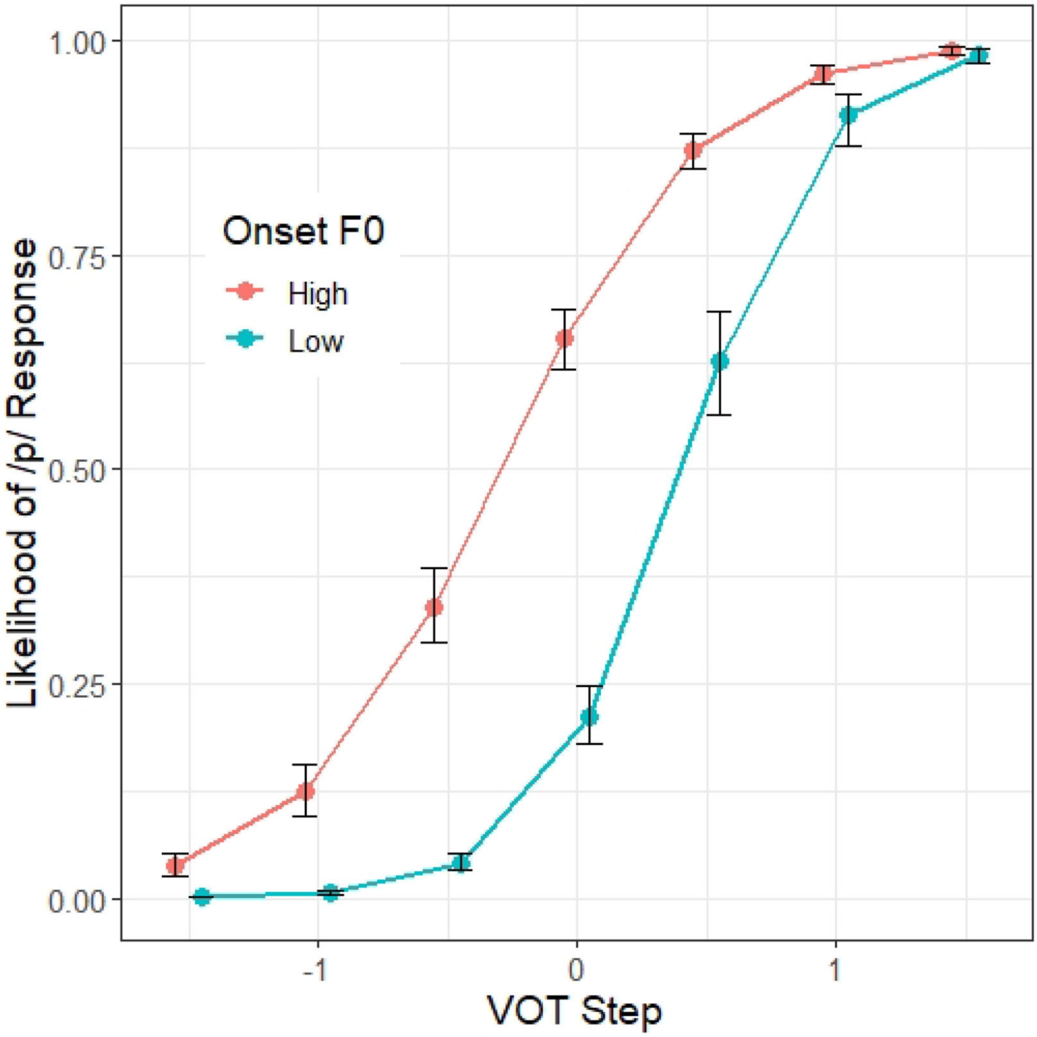

FIGURE 2 | Model 1 predictions of the probability of a /p/ response (y axis) at different steps on the VOT continuum (x axis) and F0 targets. The error bars indicate 95% confidence intervals.

condition as the baseline level showed that the response pattern from the audio-only condition differs significantly only from the male face condition, and not from the female face condition, suggesting that the VOT x CONDITION interaction is driven by the shallower VOT response pattern found in the male face condition.

There is also a significant interaction between VOT and participant GENDER ($\beta = 0.23, z = 2.04, p < 0.05$). Similar to the effect of CONDITION, as illustrated in **Figure 4**, male participants showed a shallower VOT slope than the female participants, suggesting that male listeners are less reliant on the VOT cue than the female listeners.

4.1.3. Interim Summary

The fact that the stop voicing categorization along the VOT dimension is affected by the prompt manipulation and the gender of the listener suggests that the listeners are not evaluating the speech signal in a vacuum. In accordance with our hypothesis, listeners are less reliant on VOT (as indexed by the coefficient of the VOT factor in the model) in classifying the stop voicing when the participant saw a prototypical male talker face. Also consistent is the finding that male *listeners* are less likely to rely on VOT as a cue for stop voicing. As noted earlier, VOT tends to be shorter in male than in female (e.g., Swartz, 1992; Robb et al., 2005), which means that the contrast between voiced and voiceless stops in males is more endangered in general. From the perspective of exemplar-based models that allow socio-indexical information to be encoded with each perceptual exemplar (e.g., Babel and Russell, 2015; McGowan, 2015), when the listeners were prompted to think that they were listening to a male talker, they might be activating perceptual exemplars that are consistent with male talkers and adjusting their expectations, making allowance for more ambiguities in their VOT classification (hence the shallower slope) to reflect

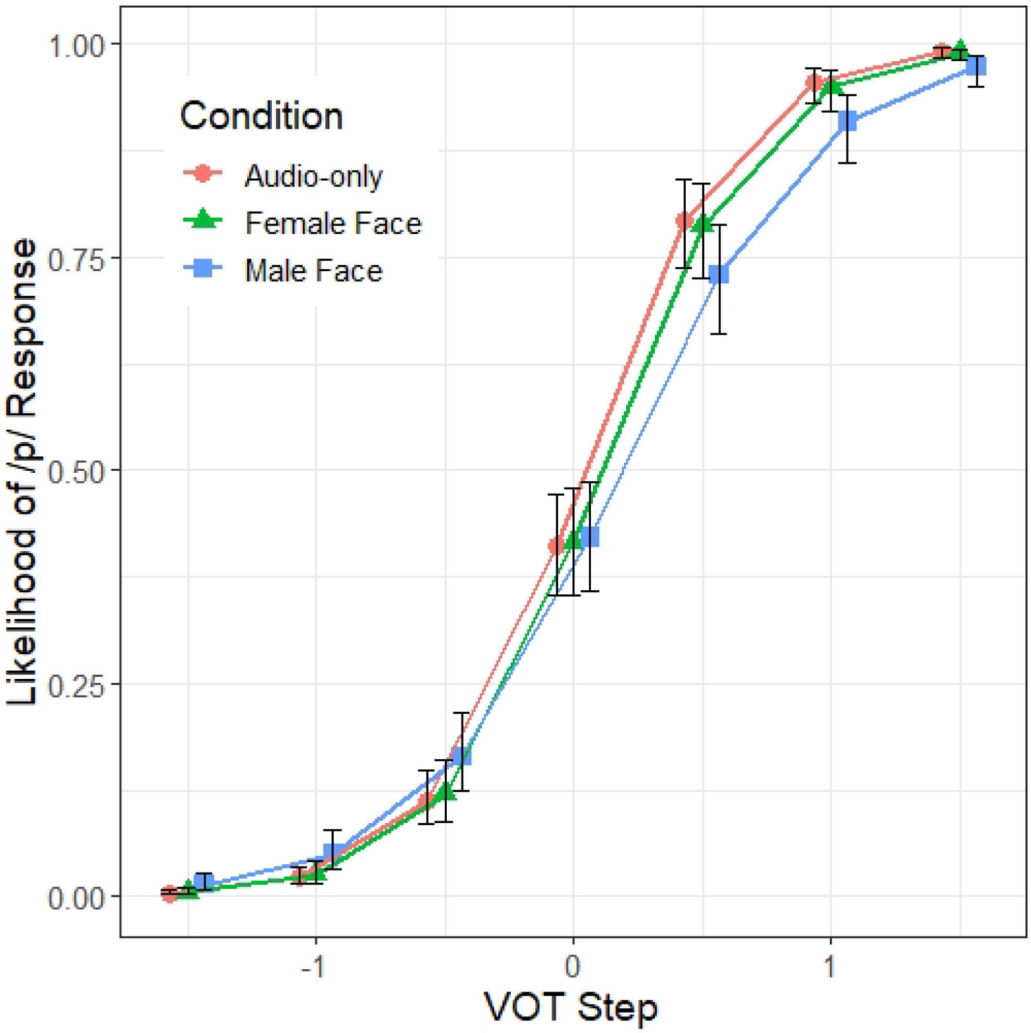

FIGURE 3 | Model 1 predictions of the probability of a /p/ response (y axis) at different steps along the VOT continuum (x axis) and across the three prompt conditions. The error bars indicate 95% confidence intervals.

their past perceptual experiences. Male listeners also rely less on VOT presumably because they are more attuned to the skewed VOT distribution in men as a result of the perception-production loop.

Our hypothesis about potential cue trading between VOT and F0 did not find support from the Model 1 results. The fact that VOT is modulated by the visual prompt manipulation but not F0 is surprising as the downweighting of the VOT cue by the listeners in the male face condition is expected to show a corresponding upweighting of F0 in the same face condition if VOT and F0 were in a trading relationship. Also unexpected is the lack of a significant vocal impression effect on cue weighting. One possible explanation for these findings might pertain to the stronger influence of visual impression over vocal ones on speech perception. Note though that the visual prompt effect is mainly driven by the male face condition, so the visual prompt manipulation alone is not likely to be sufficient to explain the mute presence of vocal impression. To this end, it is worth noting that the gender rating of the talker in the "audio-only" condition skewed toward the masculine-end of the gender prototypicality continuum (i.e., the average gender prototypicality score is 67 on a scale where 0 indexes most female-like and 100 indexes most male-like), suggesting that the talker voice might not be as gender-neutral as we had assumed based on the results of the stimulus selection task; recall that stimulus selection task showed that the chosen voice has an average gender prototypicality score of 49.2 with a standard variation of 5.5. The mute presence of vocal impression effects might have been influenced by the perceived gender-biased nature of the voice, which could have reduced the variance needed to detect any vocal impression effects.

To be sure, there is a marked difference in gender prototypicality across the two visual prompt conditions. That

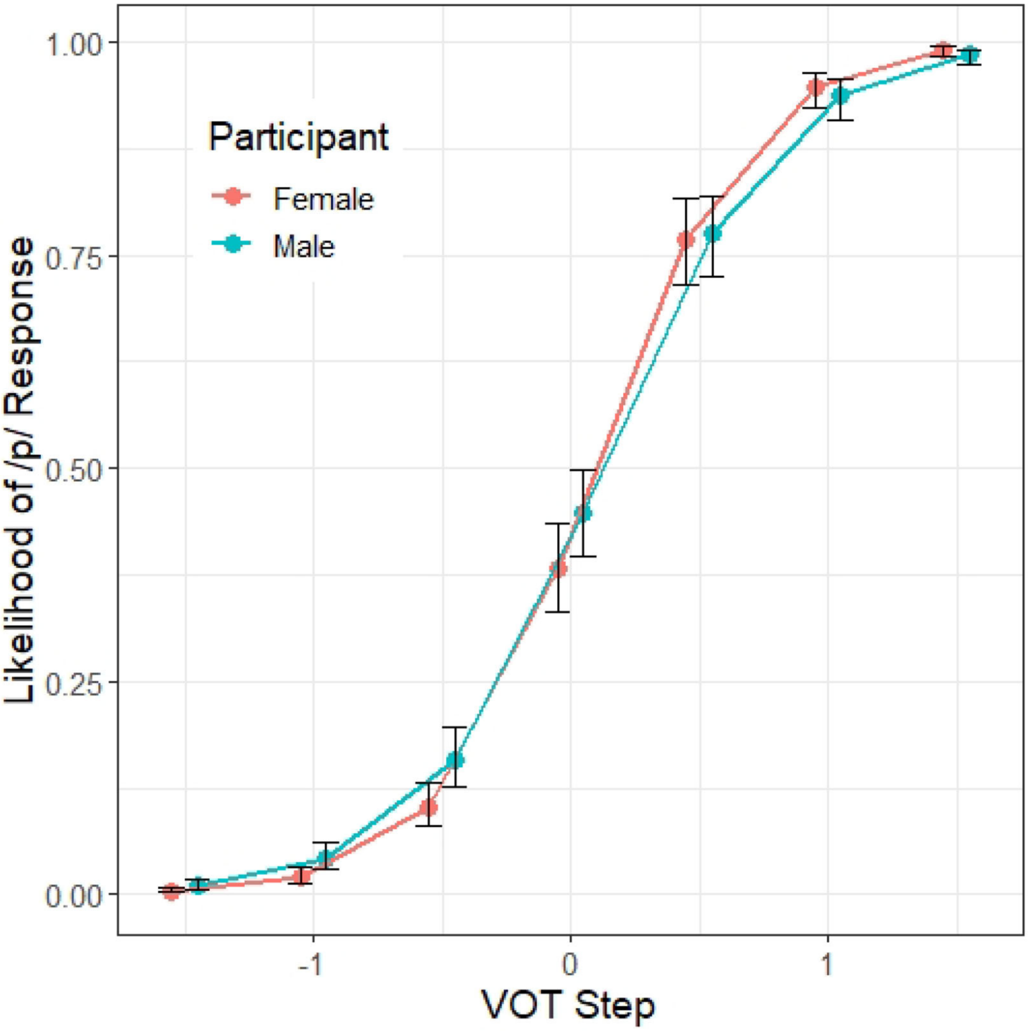

FIGURE 4 | Model 1 predictions of the probability of a /p/ response (y axis) at different steps along the VOT continuum (x axis) by the gender of the participants. The error bars indicate 95% confidence intervals.

is, the participants in the male face condition rated the talker as less masculine-sounding than in the female face condition (mean voice gender rating in the male face condition = 42.95 vs. female face = 64.55). These findings suggest that the visual prompts had an impact on how the listeners evaluated the voices; the voice was perceived to be more feminine when the participants were shown a male face and more masculine when the participants saw a female face. Listeners also did not process the visual information of the talker necessarily in the same way, particularly when it comes to perceived gender assumptions and visual first impression judgments. For example, there is quite a bit of variability in voice gender rating in both face conditions—male face: SD = 19.25, range = 0–100 vs. female face: SD = 16.27, range = 29–100. To examine in more depth the impact of the visual prompts on listeners' reliance on VOT and onset F0, the next section looks at whether and how the participants' visual impressions on the talker influence the participants' perceptual behavior.

4.2. Results From the Visual Prompt Conditions: Model 2

The last section demonstrated that the participants' reliance on VOT is impacted by the prompt condition and by the gender of the participants. No effects of vocal impressions were found. This section focuses on how the participants evaluated the talker based on the facial information presented and how the participants evaluated the talker influenced their perceptual responses.

Table 4 summarizes the visual impression ratings. As already noted above, there is quite a bit of variability in gender ratings in both face prompt conditions. This is noteworthy since the face images selected are deemed most gender-prototypical within the Chicago Face Database (Ma et al., 2015). As with the vocal attributes discussed above, there is a great deal of

TABLE 4 | Mean ratings (and standard deviations in parentheses) for perceived visual gender, attractiveness, friendliness, confidence, trustworthiness, and gayness in the two face conditions arranged by the gender of the participants.

Condition	Listener	*N*	Gender	Attractive	Friendly	Confident	Trustworthy	Gay
Female	Female	36	20 (15)	62 (20)	54 (17)	55 (18)	55 (16)	45 (22)
Female	Male	40	20 (13)	64 (20)	56 (18)	61 (20)	57 (16)	36 (19)
Male	Female	35	69 (13)	59 (21)	61 (15)	59 (15)	50 (18)	46 (16)
Male	Male	40	70 (17)	62 (20)	60 (19)	63 (17)	57 (17)	44 (14)

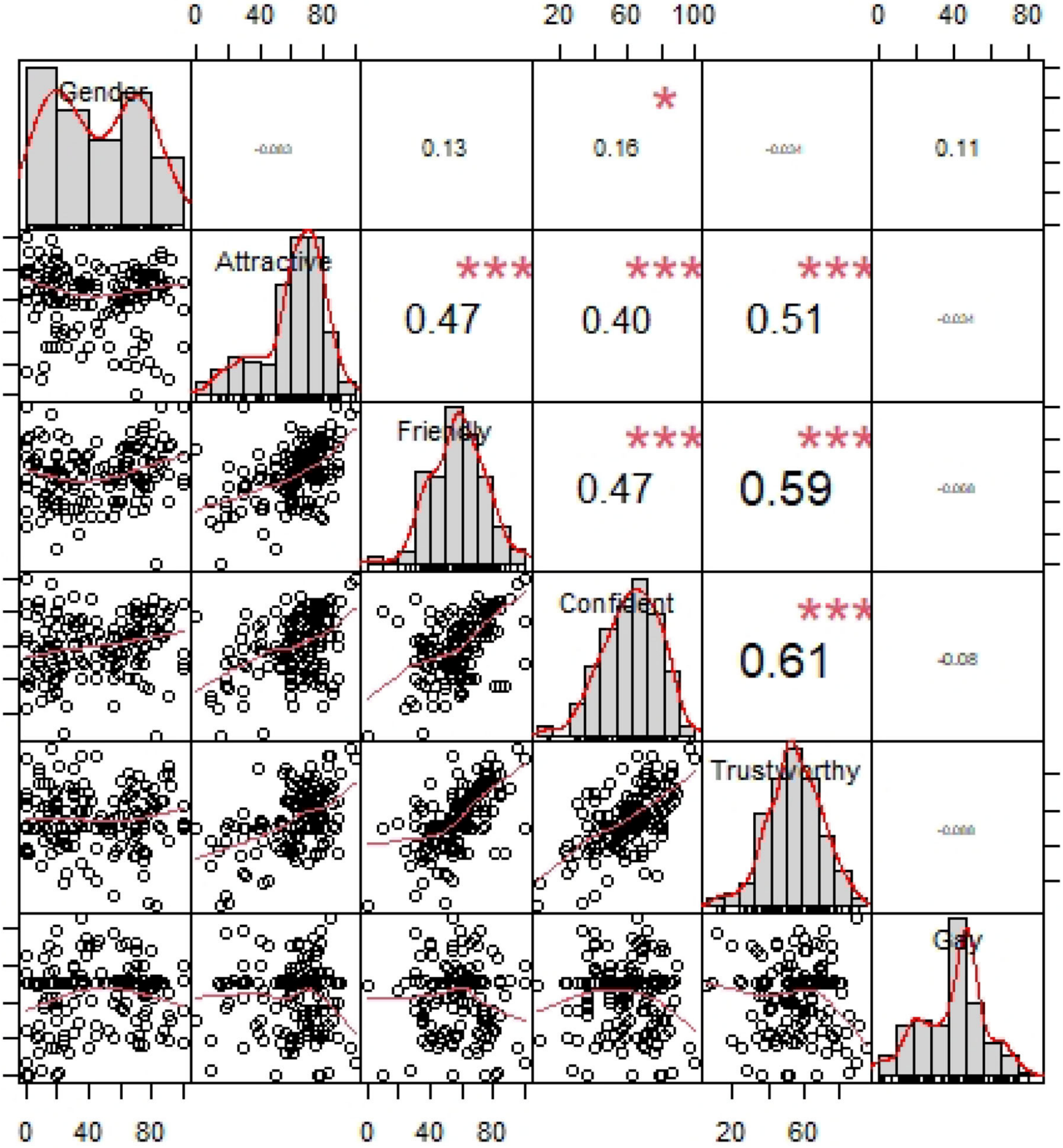

FIGURE 5 | Correlations between the ratings across different visual attributes. Each point corresponds to the ratings of a participant. $^{***}p < 0.001$; $^{*}p < 0.05$.

TABLE 5 | The cumulative proportion of variance accounted for and loadings from the PCA of the visual impression ratings.

	PC1	PC2	PC3	PC4	PC5	PC6
Visual gender	−0.05	0.79	−0.48	0.26	0.16	−0.23
Visual attractiveness	−0.45	−0.15	0.31	0.59	0.57	0.00
Visual friendliness	−0.50	0.10	−0.03	0.32	−0.70	0.38
Visual confidence	−0.49	0.14	−0.15	−0.61	0.36	0.47
Visual trustworthiness	−0.54	−0.08	0.08	−0.30	−0.18	−0.76
Gay-looking	0.08	0.57	0.80	−0.14	−0.07	0.03
Standard deviation	1.60	1.07	0.96	0.75	0.70	0.57
Proportion of variation	0.42	0.19	0.15	0.09	0.08	0.05
Cumulative proportion	0.42	0.62	0.77	0.86	0.95	1.00

TABLE 6 | The cumulative proportion of variance accounted for and loadings from the PCA of the vocal impression ratings.

	PC1	PC2	PC3	PC4	PC5	PC6
Gender	0.18	0.62	−0.69	0.08	−0.31	0.05
Vocal attractiveness	−0.42	0.05	0.06	0.90	−0.05	−0.02
Vocal friendliness	−0.52	−0.09	0.06	−0.27	−0.60	0.53
Vocal confidence	−0.46	0.40	−0.08	−0.19	0.70	0.31
Vocal trustworthiness	−0.54	0.03	−0.14	−0.27	−0.12	−0.78
Gay-sounding	−0.10	−0.67	−0.70	0.05	0.20	0.12
Standard deviation	1.60	1.07	0.96	0.75	0.70	0.57
Proportion of variation	0.43	0.19	0.15	0.11	0.07	0.05
Cumulative proportion	0.43	0.62	0.77	0.88	0.95	1.00

TABLE 7 | Estimates for all predictors in Model 2.

	Model 2
Intercept	−0.41 (0.10)***
VOT	3.02 (0.16)***
F0	1.05 (0.06)***
Gender	0.02 (0.10)
Appeal	0.12 (0.11)
Block	0.10 (0.04)**
VOT:F0	−0.59 (0.05)***
VOT:Gender	0.46 (0.16)**
F0:Gender	0.03 (0.06)
VOT:Appeal	−0.21 (0.17)
F0:Appeal	0.15 (0.06)*
Gender:Appeal	−0.14 (0.11)
VOT:F0:Gender	−0.10 (0.05)
VOT:F0:Appeal	−0.03 (0.05)
VOT:Gender:Appeal	0.34 (0.17)*
F0:Gender:Appeal	0.08 (0.06)
VOT:F0:Gender:Appeal	−0.10 (0.05)*
AIC	12098.96
BIC	12346.50
Log Likelihood	−6017.48
Num. obs.	16912
Num. groups: Participant	151
Var: Participant Intercept	1.47
Var: Participant Block	0.10
Var: Participant F0	0.39
Var: Participant VOT	3.45
Var: Participant F0:VOT	0.13

Gender refers to the gender of the participant; Appeal refers to the PC1 of the visual impression ratings.

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

variation in ratings for the other vocal impression dimensions as well as variation in how the attributes relate to each other (see **Figure 5**). Specifically, among the visual attributes, Attractiveness, Friendliness, Confidence, and Trustworthy are highly positively correlated with each other. There is also a weakly positive correlation between gender prototypicality and confidence. The distributions of the vocal attributes within the visual prompt sub-sample do not differ much from the full sample discussed above. There are strong correlations between Attractiveness, Friendliness, Confidence, and Trustworthy and between Gender and Friendliness.

Following the PCA procedure introduced above, we obtained linear combinations of the visual and vocal impression ratings that would capture the maximum variation. The relative weightings and proportion of variance for each component for the visual impression ratings are summarized in **Table 5**. The optimal linear combination (PC1), which accounts for about 42% of the variance, and the 2nd component (PC2), which accounts for approximately 19% of the variance, were selected as independent variables for the analysis below; the first two components collectively account for around 62% of the variance. PC1 has strong loadings for visual attractiveness, friendliness, confidence, and trustworthiness, which can be interpreted as indexing "visual appeal". PC2, on the other hand, is dominated by visual gender and gay-looking, which pertain to matters of gender and sexual orientation stereotypes.

Another PCA analysis of the vocal impression ratings was also conducted, focusing on just the vocal impression ratings from participants in the two visual prompt conditions only. The relative weightings and proportion of variance for each component for the vocal attributes are summarized in **Table 6**. Similar to the PCA of the vocal impression ratings of all three prompt conditions, PC1 has strong loadings for vocal attractiveness, friendliness, confidence, and trustworthiness, while PC2 is dominated by vocal gender, confidence, and gay-sounding.

A summary of the second regression model, Model 2, appears in **Table 7**. The second regression model is similar to the first model in all respects except that the CONDITION variable was not included; instead, we included the PC1 and PC2 of the visual and vocal impression ratings as discussed above. The signs of the principal components were reversed before entering the model for ease of interpretation (e.g., the higher the PC1 value of the visual impression ratings, the greater the visual appeal). Model selection started with the maximal model with all possible interactions between fixed factors (the impression rating attributes do not interact with each other, however), as

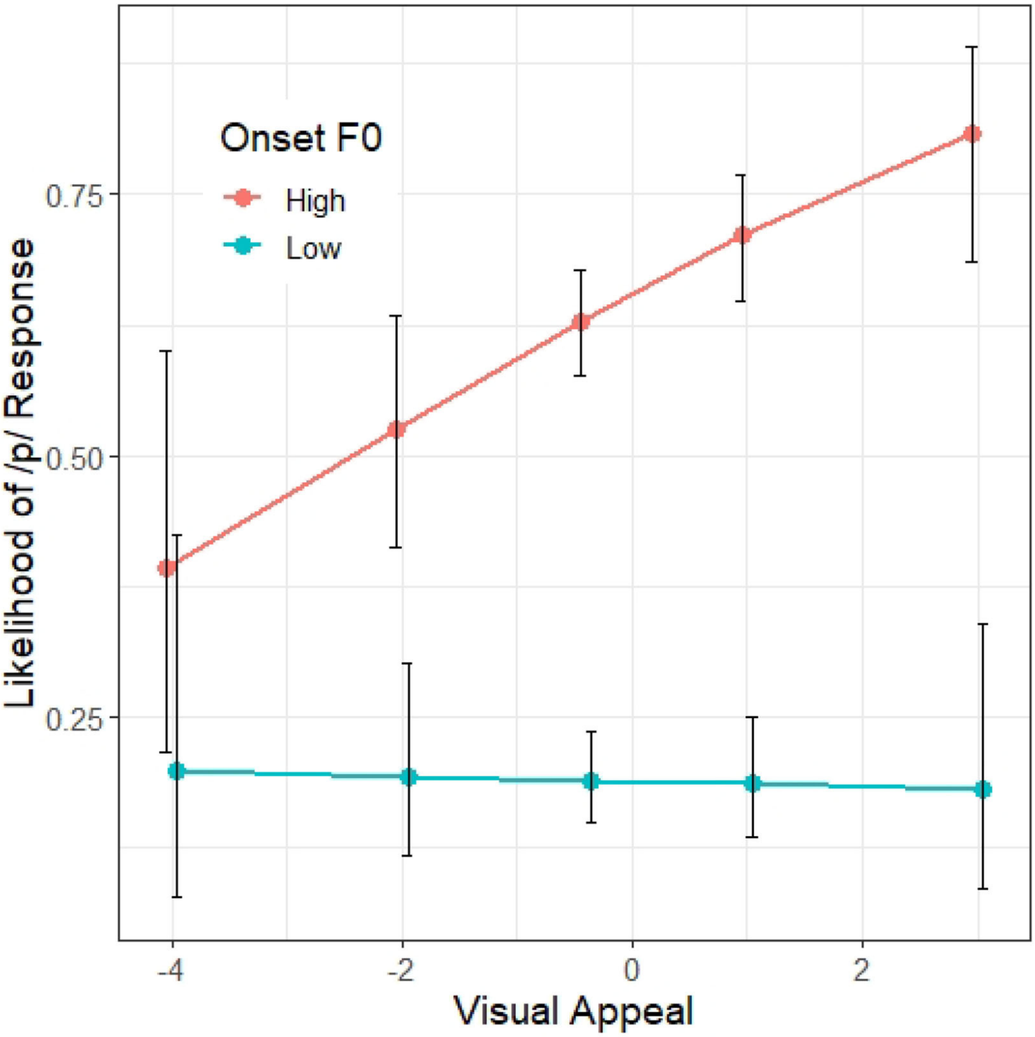

FIGURE 6 | Model 2 predictions of the probability of a /p/ response (y axis) in different F0 onset conditions according to the talker's visual appeal (x axis). The error bars indicate 95% confidence intervals.

well as the random intercepts and slopes, and proceeded by comparing between models with and without the inclusion of an impression attribute and its interaction with other factors. Visual and vocal impression attributes and their interactions that do not improve model likelihood significantly were dropped. In the end, out of the four impression attributes, only PC1 of the visual attributes was retained following this model selection procedure. For ease of reference, PC1 of the visual attributes will be referred to as "Visual Appeal" from hereon. The final model is as follows: `Response (pear = 1) ∼ BLOCK + F0 * VOT * GENDER * VISUAL APPEAL + (1+BLOCK + VOT * F0|PARTICIPANT).` In addition to the main effects of BLOCK, VOT, F0, and the interactions between the latter two, and between VOT and the gender of the participant, Model 2 also revealed several significant VISUAL APPEAL interactions. To begin with, there is a significant interaction between F0 and VISUAL APPEAL ($\beta = 0.15$, $z = 2.45$, $p = 0.01$), suggesting that the magnitude of the F0 effect on stop voicing perception is larger for listeners who found the talker visually more appealing (**Figure 6**). There is a significant interaction between VOT and participant GENDER ($\beta = 0.46$, $z = 2.81$, $p < 0.01$), but this interaction is mediated by VISUAL APPEAL ($\beta = 0.34$, $z = 2.01$, $p < 0.05$). As illustrated in **Figure 7**, the VISUAL APPEAL effect is driven by the behavior of the male participants. Specifically, the more the male participant found the talker visually appealing, the less reliant they are on VOT as a cue for stop voicing perception, as indicated by the shallower VOT slope.

Finally, there is also a significant four-way interaction between VOT, F0, participant GENDER and VISUAL APPEAL ($\beta = -0.10$, $z = -2.11$, $p < 0.05$). As illustrated in **Figure 8**, male and female listeners who rated the talker as having lower visual appeal do not differ very much in terms of their patterns of /p/ response

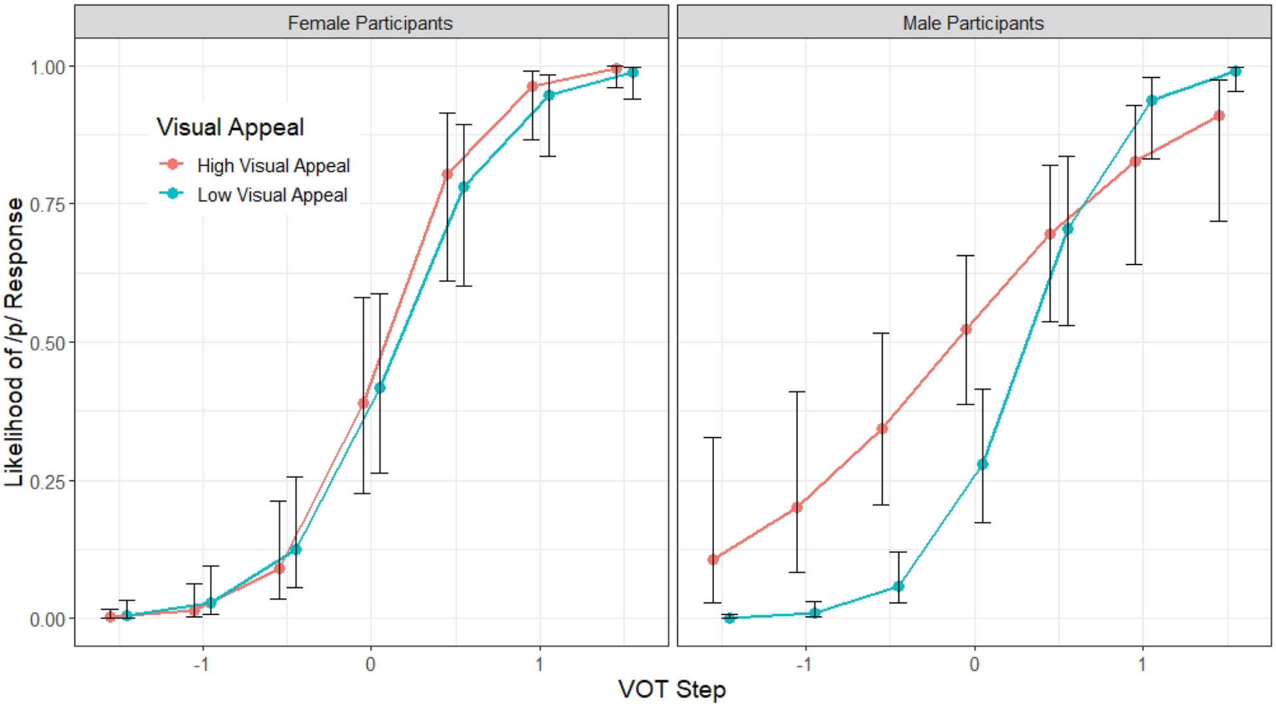

FIGURE 7 | Model 2 predictions of the probability of a /p/ response (y axis) at different steps along the VOT continuum (x axis) according to the participant's gender as well as the talker's visual appeal. While VISUAL APPEAL is continuous, for ease of presentation, only the patterns of talkers with high visual appeal (i.e., 2 standard deviation above the mean) vs. low visual appeal (2 standard deviation below the mean) are shown in the figure. The error bars indicate 95% confidence intervals.

across the VOT and F0 conditions. However, for the participants who rated the talker as having greater visual appeal, they are more likely to rely on the F0 cue (as indicated by the larger difference in /p/ response between the two onset F0 conditions) and less reliant on VOT information (as indicated by the shallower slope of the identification function along the VOT dimension). However, this visual appeal difference is more robust among the male listeners than the female listeners.

5. DISCUSSION

This study examined the effects of a listener's gender and his/her perception of a talker's gender and paralinguistic attributes on perceptual cue weighting using a matched-guise paradigm. Gender-neutral stimuli were presented to three groups of listeners, one group saw a prototypical male face, one saw a prototypical female face, and one without any visual prompt. Our regression analyses revealed that listeners who saw a male face showed less reliance on VOT compared to the listeners who saw a female face or were given no visual information. Male listeners are also less reliant on VOT in stop voicing classification. Listeners' visual impression of the talker also affected their weighting of the VOT and F0 cues. When visual information is available, listeners who had a favorable impression of the talker were less likely to rely on VOT and more likely to pay attention to the F0 cue in stop voicing classification. Male listeners who rated the talker as having more visual appeal showed the stronger reliance on F0 and the least reliance on the VOT cue.

While our findings show that perceptual cue weighting is influenced by the listener's gender and the subjective evaluations of the speaker by the listener, the mapping between the participant's interpretation of the talker's paralinguistic attributes does not always map onto the participants' perceptual responses to the VOT and onset F0 cues in the predicted manner. As noted above, to the extent that attractiveness and confidence are associated with greater VOT contrast realization, we had anticipated that listeners who rated the talker as more attractive and confident would rely more on the VOT cue than the onset F0 cue. Likewise, to the extent that femininity, friendliness, trustworthiness, and gayness are associated high mean F0 and more exaggerated F0 excursions, we had expected that listeners who rated the talker higher along these dimensions to rely more on the onset F0 cue than the VOT cue since onset F0 has been found to be more exaggerated when global F0 is high. Our results suggest that the influence of the participants' subjective impressions of the talker on VOT/F0 cue reliance is much more nuanced. To begin with, there are strong positive correlations between impressionistic judgments that were predicted to have opposite effects on cue weighting. That is, attractiveness, confidence, friendliness and trustworthiness are highly correlated even though the first two attributes were predicted to be positively associated with greater VOT reliance while the latter two attributes are associated with weaker VOT reliance. The

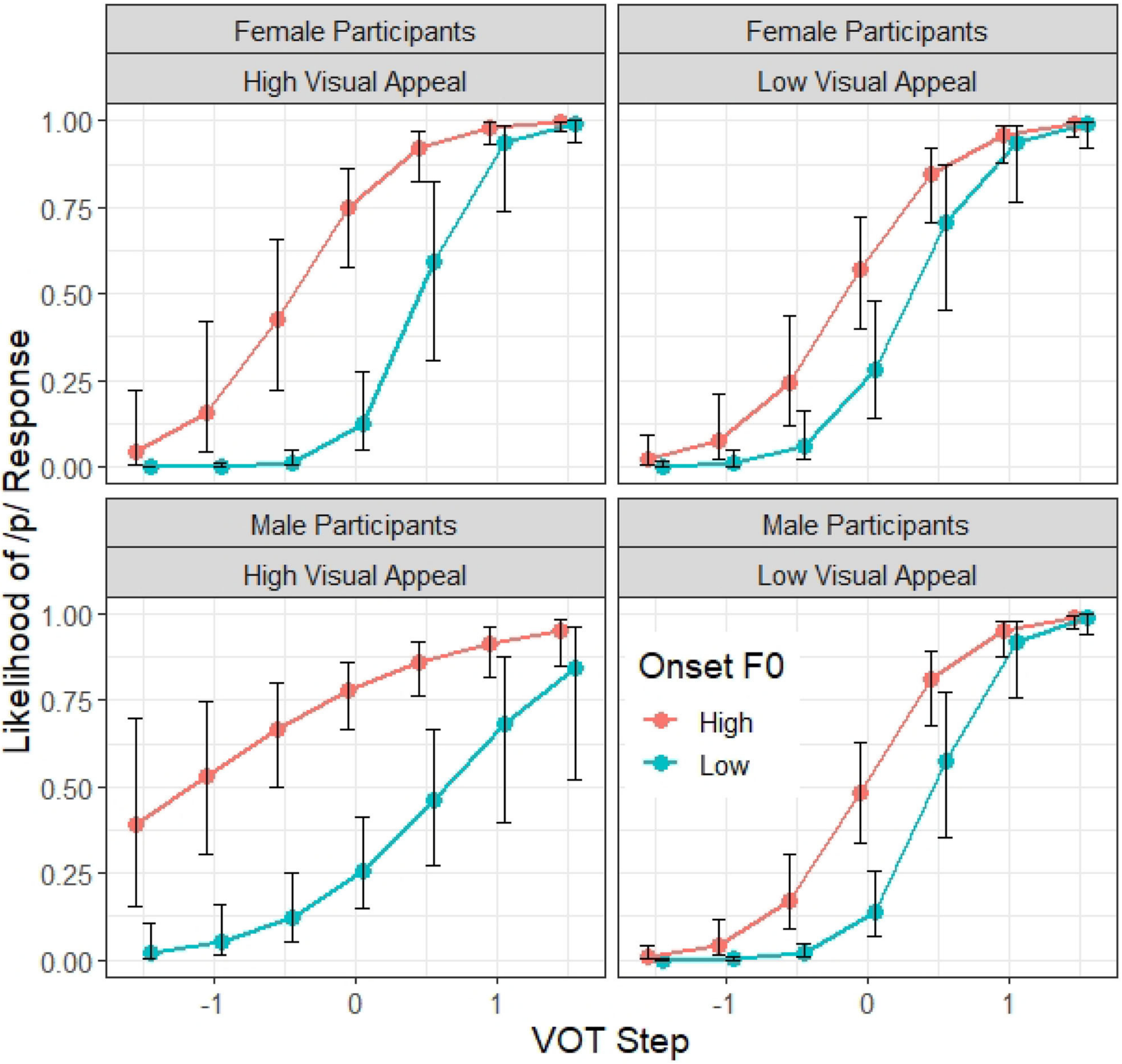

FIGURE 8 | Model 2 predictions of the probability of a /p/ response (y axis) at different steps on the VOT continuum (x axis) and F0 targets according to the talker's visual appeal and the gender of the participant. While VISUAL APPEAL is continuous, for ease of presentation, only the patterns of talkers who were rated by the participated as having high visual appeal (i.e., 2 standard deviation above the mean) vs. those with low visual appeal (2 standard deviation below the mean) are shown in the figure. The error bars indicate 95% confidence intervals.

cue-combination analytic approach adopted in the analysis (i.e., the use of Principal Component Analysis to reduce the number of highly correlated parameters prior to further modeling) prevents a direct mapping between impressionistic ratings and the participants' perceptual responses. In the end, we found that the participants would rely more heavily on the onset F0 cue than the VOT cue when the talker is rated as having greater visual appeal, a principal component involving strong loadings of attractiveness, friendliness, confidence, and trustworthiness. This state of affair points to the complexity in the way impressionistic judgments formed by the listeners interacted with the listeners' speech perceptual processes. While the Frequency Code and Effort Code hypotheses suggest potential universal associations between paralinguistic information and speech cues, it is unlikely that all associations between subjective evaluations and speech cues are fully translatable across individuals, speech communities, and cultures. Babel and McGuire (2013), for example, found that, even though perceived attractiveness ratings are highly correlated across three different varieties of North American English, listener populations nonetheless differed in the phonetic features used to make attractiveness judgments, suggesting that vocal attractiveness is dependent on community-specific preferences. Our findings suggest that more nuanced research is needed to elucidate the complex interplay between a listener's subject evaluation of his/her interlocutor and the way the listener perceives the speech outputs of that interlocutor.

Our findings are consistent with the idea that first impressions of a person can have subtle and often subjectively unrecognized effects on subsequent deliberate judgments, including perceptual cue weighting in a stop voicing classification task. The fact that visual appeal, rather than vocal appeal, exerts a stronger influence on perceptual cue weighting, as evidenced by the results of Model 2, is surprising *a priori* given the close connection between the speech cues and vocal impressions. Our findings suggest that listeners might, in general, rely more on visual impression than vocal ones to inform their perceptual judgments. Indeed, other studies have also reported stronger effects of facial impressions over vocal ones (e.g., Hou and Ye, 2019). In one study, the influence of visual impression is nearly triple that of vocal impression when evaluating competence (Klofstad, 2017). Recent models of social cognition and decision-making (Chaiken and Trope, 1999; Kahneman, 2003) posit a dual process where fast, unreflective, effortless "system 1" processes contrasts with slow, deliberate, effortful "system 2" processes. Inferences from facial appearance have been characterized as system 1 processes (Winston et al., 2002; Todorov et al., 2005). To be sure, the stronger effect of visual impression might also have stemmed from the particular design of this study. Participants in the visual prompt conditions were asked to evaluate the talker visually first prior to the talker's vocal information being introduced. Thus, the participant's earliest first impressions of the talker were formed entirely based on visual information alone. First impressions based on visual cues alone might have a stronger biasing effect on the subsequent behavior of the listeners than the vocal information which was introduced later. The gender-specificity of the effects of visual impressions on cue weighting is also noteworthy. The effects of visual appeal, as revealed in Model 2, is more strongly driven by the male participants. These findings are consistent with the observation that men and women may be affected by their own impressionistic judgments differently. For example, men evaluate female facial attractiveness as higher than male facial attractiveness while women do not show a similar tendency in evaluation male facial attractiveness higher than female facial attractiveness (Hou and Ye, 2019). Babel (2012) found that men and women exhibit different rates of vowel imitation depending on the race and attractiveness of the talker.

The fact that a listener's perception of the gender and personality features of a talker could affect the listener's cue weight raises question about the mechanism(s) behind such an influence. As noted earlier, exemplar-based models of speech perception and production that allow socio-indexical information to be encoded as part of the episodic traces in the mental lexicon provides a potential model for understanding how socio-indexical and paralinguistic information could modulate speech perception. We hypothesized that, when a listener judges a talker to be of a particular gender or has certain personality features, the listener's perceptual system might adjust its cue weight expectation in accordance to the specific socio-indexical and paralinguistic norms. Our findings are broadly consistent with these predictions. Specifically, the direction of the cue weight adjustments with respect to perceived gender is consistent with the idea that listeners are informed by the past experiences (i.e., VOT distinctions among oral stops are less distinct among males than among females). Male participants exhibited the strongest cue weight adjustments, presumably due to their familiarity with their own production tendencies relative to their female counterparts. As noted above, the influence of impressionistic judgments on the personality attributes of the talker on cue weights are more nuanced due to the complex mapping between VOT and F0 variations and personality traits. The final analysis suggests that when the participant found the talker to have high visual appeal, the participant is more likely to pay greater attention to onset F0 than VOT cues. This pattern is consistent with the observation that higher pitch is associated with more affective interpretations. That is, if individuals with greater visual appeal are seen as more affective people, great visual appeal might have primed the participants to activate perceptual experiences associated with affective individuals. Listeners might heighten their attention to onset F0 differences since affective individuals are associated with higher overall F0 and less distinct VOT contrast in their speech outputs.

The fact that native English-speaking listeners' perceptual weighting of VOT and onset F0 cues is impacted by the perceived socio-indexical and personality characteristics of the talker lends further support to the idea that the relationship between the VOT and onset F0 cues in English is part of the controlled phonetic knowledge of English speakers (Kingston and Diehl, 1994; Solé, 2007). According to the cue-reweighting approach to the development of tone split and tonogenesis (Hyman, 1976; Kang, 2014; Coetzee et al., 2018), one pathway to developing allophonic pitch variation is via the phonologization of consonantal perturbation of pitch on the neighboring vowel. The fact that the trading relation between VOT and onset F0 is part of the phonetic knowledge of English speakers raises the question of whether English might be undergoing a sound change in progress. That is, are English stops developing a tone split analogous to what has been documented in Afrikaans (Coetzee et al., 2018) recently? While this is not a question the present study can answer definitively, it is nonetheless important to note that, given the propagation of any sound change crucially depends on the innovative variation developing sociolinguistic significance, the fact that English-speaking listeners are sensitive to the social characteristics of the talker in their perceptual responses to VOT/F0 variation points to, at the minimum, the emergence of some form of sociolinguistic awareness of the VOT/F0 covariation. This interpretation is further supported by developmental studies that look at gender differentiation in VOT realization. Whiteside and Marshall (2001), for example, studied the developmental trajectory of VOT in English /p/ and /t/ for boys and girls aged 7, 9, and 11 years and found that mean VOT differences between voiced and voiceless stops were larger for girls than for boys at age 11 due to the boys' marked decrease in VOT difference from age 9 to 11. They argued that the gender differences might be the result of the amplification of an intrinsic variation due to anatomical differences between males and females. To be sure, it is not clear at this point if comparable onset F0 changes would accompany the gender-differentiated developmental changes in VOT, but our findings suggest that, at least among adult listeners, the trading relationship between VOT and onset F0 is gender-differentiated. These gender differences in the production and perception of VOT/onset F0 variation are prime materials (i.e., the first order indexicality association) for the speakers to recruit in their ideological projects (Eckert, 2019). What is observed in English today might be an analog to the precursor stage to the development of F0 distinctions in the Seoul Korean stop laryngeal system. Oh (2011), for example, examined the VOT of voiceless aspirated plosives in Seoul Korean and found that male speakers have significantly longer VOT than female speakers. She hypothesized a potential link between the gender difference to an ongoing change where the distinction between lenis and aspirated plosives are increasingly cued by differences in F0 rather than VOT. While no definitive historical evidence was provided, she did note that the gender difference appeared to have existed prior to the sound change commencing in Seoul Korean (see also Kang, 2014).

In sum, the present study offers crucial evidence for listeners' sensitivity to the talker's socio-indexical and personality characteristics in their perceptual responses to VOT and onset F0 variation. Our findings lend support for the type of cue reweighting model of sound change (Hyman, 1976; Kang, 2014; Coetzee et al., 2018), as they not only further cement the controlled phonetic knowledge interpration of VOT/onset F0 co-variation in English, but also reveal a sociolinguistic dimension to this co-variation. More investigation is needed to examine the possibility of a sound change in progress in North American English concerning the relation between stop voicing and F0. In particular, apparent time investigations or panel studies into the community patterning of the F0 perturbation effect in North American English across age groups and gender could be particularly revealing.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The supplementary materials only provide the images and sound files used. The dataset and the analysis scripts can be found at https://osf.io/fx8ay/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Social and Behavioral Sciences Institutional Review Board at the University of Chicago. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AY contributed to the design and implementation of the research topic, to the analysis of the results, and writing of the manuscript.

FUNDING

This work was supported in part by the National Science Foundation (BCS1827409).

ACKNOWLEDGMENTS

Many thanks to members of the Chicago Phonology Laboratory and the audience at the 6th GDUFS Forum on Linguistics and Applied Linguistics for their feedback and to Henry Mackall for his assistance in setting up the pilot versions of this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.840291/full#supplementary-material

REFERENCES

Allison, C., Auyeung, B., and Baron-Cohen, S. (2012). Toward brief "Red Flags" for autism screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist for Autism in toddlers in 1,000 cases and 3,000 controls. *J. Am. Acad. Child Adolesc. Psychiatry* 51, P202–P212. doi: 10.1037/t30469-000

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *J. Phonet.* 40, 177–189. doi: 10.1016/j.wocn.2011.09.001

Babel, M., and McGuire, G. (2013). "Perceived vocal attractiveness across dialects is similar but not uniform," in *Proceedings of Interspeech* (Lyon).

Babel, M., McGuire, G., and King, J. (2014). Towards a more nuanced view of vocal attractiveness. *PLoS ONE* 9, e88616. doi: 10.1371/journal.pone.0088616

Babel, M., and Russell, J. (2015). Expectations and speech intelligibility. *J. Acoust. Soc. Am.* 137, 2823–2833. doi: 10.1121/1.4919317

Banse, R., and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *J. Pers. Soc. Psychol.* 70, 614–636.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw*. 67, 1–48. doi: 10.18637/jss.v067.i01

Brown, V. A., Hedayati, M., Zanger, A., Mayn, S., Ray, L., Dillman-Hasso, N., et al. (2018). What accounts for individual differences in susceptibility to the McGurk effect? *PLoS ONE* 13, e0207160. doi: 10.1371/journal.pone.0207160

Campbell-Kibler, K. (2009). The nature of sociolinguistic perception. *Lang. Variat. Change* 21, 135–156. doi: 10.1017/S0954394509000052

Campbell-Kibler, K. (2010). Sociolinguistics and perception. *Lang. Lingusit. Compass* 4, 377–389. doi: 10.1111/j.1749-818X.2010.00201.x

Chaiken, S., and Trope, Y., editors (1999). *Dual-Process Theories in Social Psychology*. New York, NY: The Guilford Press.

Chen, A., Gussenhoven, C., and Rietveld, T. (2004a). Language-specificity in the perception of paralinguistic intonational meaning. *Lang. Speech* 47, 311–349. doi: 10.1177/00238309040470040101

Chen, D., Halberstam, Y., and Yu, A. C. L. (2016). Perceived masculinity predicts U.S. supreme court outcomes. *PLoS ONE* 11, e0164324. doi: 10.1371/journal.pone.0164324

Chen, F., Li, A., Wang, H., Wang, T., and Fang, Q. (2004b). "Acoustic analysis of friendly speech," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Montreal, QC), 1–569.

Chodroff, E., and Wilson, C. (2018). Predictability of stop consonant phonetics across talkers: between-category and within-category dependencies among cues for place and voice. *Linguist. Vanguard*. 4. doi: 10.1515/lingvan-2017-0047

Clayards, M. (2018a). Differences in cue weights for speech perception are correlated for individuals within and across contrasts. *J. Acoust. Soc. Am.* 144, EL172. doi: 10.1121/1.5052025

Clayards, M. (2018b). Individual talker and token variability in multiple cues to stop voicing. *Phonetica* 75, 1–23. doi: 10.1159/000448809

Coetzee, A. W., Beddor, P. S., Shedden, K., Styler, W., and Wissing, D. (2018). Plosive voicing in Afrikaans: differential cue weighting and sound change. *J. Phonet.* 66, 185–216. doi: 10.1016/j.wocn.2017.09.009

de Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *J. Acoust. Soc. Am.* 97, 491–504.

Dmitrieva, O., Llanos, F., Shultz, A. A., and Francis, A. L. (2015). Phonological status, not voice onset time, determines the acoustic realization of onset F0 as a secondary voicing cue in Spanish and English. *J. Phonet.* 49, 77–95. doi: 10.1016/j.wocn.2014.12.005

Eckert, P. (2008). Variation and the indexical field. *J. Sociolinguist.* 12, 453–476. doi: 10.1111/j.1467-9841.2008.00374.x

Eckert, P. (2019). The individual in the semiotic landscape. *Glossa J. Gen. Linguist.* 4, 1. doi: 10.5334/gjgl.640

Francis, A. L., Ciocca, V., Wong, V. K., and Chan, J. K. (2006). Is fundamental frequency a cue to aspiration in initial stops? *J. Acoust. Soc. Am.* 120, 2884–2895. doi: 10.1121/1.2346131

Francis, A. L., Kaganovich, N., and Discoll-Huber, C. (2008). Cue-specific effects of categorization training on therelative weighting of acoustic cues to consonant voicing in English. *J. Acoust. Soc. Am.* 124, 1234–1251. doi: 10.1121/1.2945161

Giovannone, N., and Theodore, R. M. (2021). Individual differences in lexical contributions to speech perception. *J. Speech Lang. Hear. Res.* 64, 707–724. doi: 10.1044/2020_JSLHR-20-00283

Gordon, P. C., Eberhardt, J. L., and Rueckl, J. G. (1993). Attentional modulation of the phonetic significance of acoustic cues. *Cogn. Psychol.* 25, 1–42.

Grogger, J. (2011). Speech patterns and racial wage inequality. *J. Hum. Resour.* 46, 1–25. doi: 10.1353/jhr.2011.0017

Gussenhoven, C. (2002). "Intonation and interpretation: phonetics and phonology," in *Proceedings of the Speech Prosody*, eds B. Bel and I. Arlien (Aix-en Provence: Université de Provence), 47–57.

Haggard, M., Summerfield, Q., and Roberts, M. (1981). Psychoacoustical and cultural determinants of phoneme boundaries: evidence from trading F0 cues in the voiced-voiceless distinction. *J. Phonet.* 9, 49–62.

Halle, M., and Stevens, K. (1971). "A note on laryngeal features," in *MIT Research Laboratory of Electronics Quarterly Progress Report*, 198–213.

Hanson, H. M. (2009). Effects of obstruent consonants on fundamental frequency at vowel onset in English. *J. Acoust. Soc. Am.* 125, 425–441. doi: 10.1121/1.3021306

Hay, J., Warren, P., and Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *J. Phonet.* 34, 458–484. doi: 10.1016/j.wocn.2005.10.001

Hombert, J.-M., Ohala, J. J., and Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language* 55, 37–58.

Hou, J., and Ye, Z. (2019). Sex differences in facial and vocal attractiveness among college students in China. *Front. Psychol.* 10, 1166. doi: 10.3389/fpsyg.2019.01166

Hyman, L. (1976). "Phonologization," in *Linguistic Studies Presented to Joseph H. Greenberg*, ed A. Juilland (Saratoga, CA: Anma Libri), 407–418.

Idemaru, K., Holt, L. L., and Seltman, H. (2012). Individual differences in cue weight are stable across time: the case of Japanese stop lengths. *J. Acoust. Soc. Am.* 132, 3950–3964. doi: 10.1121/1.4765076

Jacob, C., Guéguen, N., Martin, A., and Boulbry, G. (2011). Retail salespeople's mimicry of customers: effects on consumer behavior. *J. Retail. Cons. Serv.* 18, 381–388. doi: 10.1016/j.jretconser.2010.11.006

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *Am. Psychol.* 58, 697–720. doi: 10.1037/0003-066X.58.9.697

Kaiser, H. F. (1961). A note on Guttman's lower bound for the number of common factors. *Brit. J. Stat. Psychol.* 14, 1–2.

Kang, Y. (2014). Voice Onset Time merger and development of tonal contrast in Seoul Korean stops: a corpus study. *J. Phonet.* 45, 76–90. doi: 10.1016/j.wocn.2014.03.005

Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., and McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: an individual differences approach. *J. Exp. Psychol. Hum. Percept. Perform.* 43, 1594–1611. doi: 10.1037/xhp0000410

Kapnoula, E. E. (2016). *Individual differences in speech perception: sources, functions, and consequences of phoneme categorization gradiency* (Ph.D. thesis). University of Iowa, Iowa City, IA, United States.

Keyser, S. J., and Stevens, K. N. (2006). Enhancement and overlap in the speech chain. *Language* 82, 33–63. doi: 10.1353/lan.2006.0051

Kingston, J. (2007). "Segmental influences on F0: automatic or controlled?" in *Tones and Tunes, Volume 2: Experimental Studies Inword and Sentence Prosody*, eds C. Gussenhoven and T. Riad (Berlin: Mouton de Gruyter), 171–201.

Kingston, J., and Diehl, R. L. (1994). Phonetic knowledge. *Language* 70, 419–454.

Kirby, J., Kleber, F., Siddins, J., and Harrington, J. (2020). "Effects of prosodic prominence on obstruent-intrinsic F0 and VOT in German," in *Proc. 10th International Conference on Speech Prosody 2020* (Tokyo), 210–214.

Kleinschmidt, D. F., and Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol. Rev.* 122, 148–203. doi: 10.1037/a0038695

Kleinschmidt, D. F., Weatherholz, K., and Jaeger, T. F. (2018). Sociolinguistic perception as inference under uncertainty. *Top. Cogn. Sci.* 10, 818–834. doi: 10.1111/tops.12331

Klofstad, C. A. (2017). Look and sounds like a winner: perceptions of competence in candidates' faces and voices influences vote choice. *J. Exp. Polit. Sci.* 4, 229–240. doi: 10.1017/XPS.2017.19

Koenig, L. L. (2000). Laryngeal factors in voiceless consonant production in men, women,and 5-year-olds. *J. Speech Lang. Hear. Res.* 43, 1211–1228. doi: 10.1044/jslhr.4305.1211

Kong, E. J., and Edwards, J. (2016). Individual differences in categorical perception of speech: cue weighting and executive function. *J. Phonet.* 59, 40–57. doi: 10.1016/j.wocn.2016.08.006

Kronrod, Y., Coppess, E., and Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychon. Bull. Rev.* 23, 1681–1712. doi: 10.3758/s13423-016-1049-y

Ladefoged, P. (1967). *Three areas of Experimental Phonetics*. London: Oxford University Press.

Lambert, W. E., Hodgson, R. C., Gardner, R. C., and Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *J. Abnormal Soc. Psychol.* 60, 44–51.

Lehet, M., and Holt, L. L. (2017). Dimension-based statistical learning affects both speech perception and production. *Cogn. Sci.* 41, 885–912. doi: 10.1111/cogs.12413

Li, F. F. (2013). The effect of speakers' sex on voice onset time in Mandarin stops. *J. Acoust. Soc. Am.* 133, 142–147. doi: 10.1121/1.4778281

Lisker, L. (1986). "Voicing" in English: a catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Lang. Speech* 29, 3–11.

Löfqvist, A., Baer, T., McGarr, N. S., and Story, R. S. (1989). The cricothryoid muscle in voicing control. *J. Acoust. Soc. Am.* 85, 1314–1321.

Lundeborg, I., Larsson, M., Wiman, S., and McAllister, A. M. (2012). Voice onset time in Swedish children and adults. *Logoped. Phoniatr. Vocol.* 37, 117–122. doi: 10.3109/14015439.2012.664654

Ma, D. S., Correll, J., and Wittenbrink, B. (2015). The Chicago face database: a free stimulus set of faces and norming data. *Behav. Res. Methods* 47, 1122–1135. doi: 10.3758/s13428-014-0532-5

Mayew, W. J., and Venkatachalam, M. (2012). The power of voice: managerial affective states and future firm performance. *J. Fin.* 67, 1–43. doi: 10.1111/j.1540-6261.2011.01705.x

McAleer, P., Todorov, A., and Belin, P. (2014). How do you say "hello"? Personality impressions from brief novel voices. *PLoS ONE* 9, e90779. doi: 10.1371/journal.pone.0090779

McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Lang. Speech* 58, 502–521. doi: 10.1177/0023830914565191

Morris, R. J., McCrea, C. R., and Herring, K. D. (2008). Voice onset time differences between adult males and females: isolated syllables. *J. Phonet.* 36, 308–317. doi: 10.1016/j.wocn.2007.06.003

Myslin, M., and Levy, R. (2016). Comprehension priming as rational expectation for repetition: evidence from syntactic processing. *Cognition* 147, 29–56. doi: 10.1016/j.cognition.2015.10.021

Oh, E. (2011). Effects of speaker gender on voice onset time in Korean stops. *J. Phonet.* 39, 59–67. doi: 10.1016/j.wocn.2010.11.002

Ohala, J. J. (1973). *Explanations for the Intrinsic Pitch of Vowels.* Monthly Internal Memorandum, Phonology Laboratory Berkeley, Berkeley, CA.

Ohala, J. J. (1983). Cross-language use of pitch: an ethological view. *Phonetica* 40, 1–18.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica* 41, 1–16.

Ou, J., and Yu, A. C. L. (2021). Neural correlates of individual differences in speech categorization: evidence from subcortical, cortical, and behavioral measures. *Lang. Cogn. Neurosci*. doi: 10.31219/osf.io/u79hg

Ou, J., Yu, A. C. L., and Xiang, M. (2021). Individual differences in categorization gradience as predicted by online processing of phonetic cues during spoken word recognition: Evidence from eye movements. *Cogn. Sci*. 45, e12948.

Peng, J.-F., Chen, L., and Lee, C.-C. (2014). Voice onset time of initial stops in Mandarin and Hakka: effect of gender. *Taiwan J. Linguist.* 21, 63–80. doi: 10.6519/TJL.2014.12(1).3

Pierrehumbert, J. (2002). "Word specific phonetics," in *Laboratory Phonology VII*, eds C. Gussenhoven and N. Warner (Berlin: Mouton de Gruyter), 101–139.

Purnell, T., Idsardi, W., and Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *J. Lang. Soc. Psychol.* 18, 10–30.

Puts, D. A., Gaulin, S. J., and Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evol. Hum. Behav.* 27, 283–296. doi: 10.1016/j.evolhumbehav.2005.11.003

Reddy, B. M. S., Kumar, N. M., and Sreedevi, N. (2013). Voice onset time across gender and different vowel contexts in Telugu. *Lang. India* 14, 252–263. Available online at: http://www.languageinindia.com/dec2014/madhuvotfinal.pdf

Repp, B. H. (1982). Phonetic trading relations and context effects: new experimental evidence for a speech mode of perception. *Psychol. Bull.* 82, 81–110.

Rickford, J. R., Duncan, G. J., Gennetian, L. A., Gou, R. Y., Greene, R., Katz, L. F., et al. (2015). Neighborhood effects on use of African-American Vernacular English. *Proc. Natl. Acad. Sci. U.S.A.* 112, 11817–11822. doi: 10.1073/pnas.1500176112

Robb, M., Gilbert, H., and Lerman, J. (2005). Influence of gender and environmental setting on voice onset time. *Folia Phoniatr Logop.* 57, 125–133. doi: 10.1159/000084133

Ryalls, J., Zipprer, A., and Baldauff, P. (1997). A preliminary investigation of the effects of gender and race on voice onset time. *J. Speech Hear. Res.* 40, 642–645.

Schertz, J., Cho, T., Lotto, A., and Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *J. Phonet.* 52, 183–204. doi: 10.1016/j.wocn.2015.07.003

Schroeder, J., and Epley, N. (2015). The sound of intellect: speech reveals a thoughtful mind, increasing a job candidates appeal. *Psychol. Sci.* 26, 877–891. doi: 10.1177/0956797615572906

Shultz, A. A., Francis, A. L., and Llanos, F. (2012). Differential cue weighting in perception and production of consonanat voicing. *J. Acoust. Soc. Am. Express Lett.* 132, 95–101. doi: 10.1121/1.4736711

Smiljanić, R., and Bradlow, A. R. (2008). Stability of temporal contrastsacross speaking styles in English and Croatian. *J. Phonet.* 36, 91–113. doi: 10.1016/j.wocn.2007.02.002

Smyth, R., Jacobs, G., and Rogers, H. (2003). Male voices and perceived sexual orientation: an experimental and theoretical approach. *Lang. Soc.* 32, 329–350. doi: 10.1017/S0047404503323024

Smyth, R., and Rogers, H. (2002). "Phonetics, gender, and sexual orientation," in *Proceedings of the Annual Meeting of the Canadian Linguistics Association* (Montreal, QC: l'University du Quebec au Montreal), 299–301.

Solé, M.-J. (2007). "Controlled and mechanical properties in speech: a review of the literature," in *Experimental Approaches to Phonology*, eds M. J. Solé, P. S. Beddor, and M. Ohala (Oxford: Oxford University Press), 302–321.

Staum Casasanto, L. (2010). "What do listeners know about sociolinguistic variation?" in *University of Pennsylvania Working Papers in Linguistics* (Philadelphia, PA), 6.

Stern, J., Schild, C., Jones, B. C., DeBruine, L. M., Hahn, A., Puts, D. A., et al. (2021). Do voices carry valid information about a speaker's personality? *J. Res. Pers.* 92, 104092. doi: 10.1016/j.jrp.2021.104092

Strand, E. A. (2000). *Gender stereotype effects in speech processing* (Ph.D. thesis). The Ohio State University, Columbus, OH, United States.

Sumner, M., Kim, S. K., King, E., and McGowan, K. B. (2014). The socially weighted encoding of spoken words: a dual-route approach to speech perception. *Front. Psychol.* 4, 1015. doi: 10.3389/fpsyg.2013.01015

Swartz, B. L. (1992). Gender difference in voice onset time. *Percept. Motor Skills* 75, 983–992.

Thomas, K. A., and Clifford, S. (2017). Validity and MechanicalTurk: an assessment of exclusion methods and interactive experiments. *Comput. Hum. Behav.* 77, 184–197. doi: 10.1016/j.chb.2017.08.038

Titze, I. R. (2000). *Principles of Voice Production*. Iowa City, IA: National Center for Voice and Speech.

Todorov, A., Mandisodza, A. N., Goren, A., and Hall, C. C. (2005). Inferencces of competencce from faces predict election outcomes. *Science* 308, 1623–1626. doi: 10.1126/science.1110589

Toscano, J. C., and McMurray, B. (2010). Cue integration with categories: weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cogn. Sci.* 34, 434–464. doi: 10.1111/j.1551-6709.2009.01077.x

Wadnerkar, M. B., Cowell, P. E., and Whiteside, S. P. (2006). Speech across the menstrual cycle: a replication and extension study. *Neurosci. Lett.* 408, 21–24. doi: 10.1016/j.neulet.2006.07.032

Whiteside, S. P., Hanson, A., and Cowell, P. E. (2004a). Hormones and temporal components of speech: sex differences and effects of menstrual cyclicity on speech. *Neurosci. Lett.* 367, 44–47. doi: 10.1016/j.neulet.2004.05.076

Whiteside, S. P., Henry, L., and Dobbin, R. (2004b). Sex differences in voice onset time: a developmental study of phonetic context effects in British English. *J. Acoust. Soc. Am.* 116, 1179–1183. doi: 10.1121/1.1768256

Whiteside, S. P., and Irving, C. J. (1997). Speakers' sex differences in voice onset time: some preliminary findings. *Percept. Motor Skills* 85, 459–463.

Whiteside, S. P., and Irving, C. J. (1998). Speaker's sex differences in voice onset time: a study of isolated word production. *Percept. Motor Skills* 86, 651–654.

Whiteside, S. P., and Marshall, J. (2001). Developmental trends in voice onset time: some evidence for sex differences. *Phonetica* 58, 196–210. doi: 10.1159/000056199

Winn, M. B. (2020). Manipulation of voice onset time in speech stimuli: a tutorial and flexible Praat script. *J. Acoust. Soc. Am.* 147, 852–866. doi: 10.1121/10.0000692

Winston, J. S., Strange, B. A., O'Doherty, J., and Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nat. Neurosci.* 5, 277–283. doi: 10.1038/nn816

Woods, K. J., Siegel, M. H., Traer, J., and McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attent. Percept. Psychophys.* 79, 2064–2072. doi: 10.3758/s13414-017-1361-2

Xu, Y., Lee, A., Wu, W.-L., Liu, X., and Birkholz, P. (2013). Human vocal attractiveness as signaled by body size projection. *PLoS ONE* 8, e62397. doi: 10.1371/journal.pone.0062397

Yu, A. C. L., Abrego-Collier, C., and Sonderegger, M. (2013). Phonetic imitation from an individual-difference perspective: Subjective attitude, personality, and 'autistic' traits. *PLoS ONE* 8, e74746. doi: 10.1371/journal.pone.0074746

Zahn, C. J., and Hopper, R. (1985). Measuring language attitudes: the speech evaluation instrument. *J. Lang. Soc. Psychol.* 4, 113–23.

Zhang, X., and Holt, L. L. (2018). Simultaneous tracking of coevolving distributional regularities in speech. *J. Exp. Psychol. Hum. Percept. Perform.* 44, 1760–1779. doi: 10.1037/xhp0000569

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 26 April 2022
doi: 10.3389/fpsyg.2022.839087

Repetitive Exposure to Orofacial Somatosensory Inputs in Speech Perceptual Training Modulates Vowel Categorization in Speech Perception

Takayuki Ito[1,2]* and Rintaro Ogane[1,2]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, France, [2] Haskins Laboratories, New Haven, CT, United States

Orofacial somatosensory inputs may play a role in the link between speech perception and production. Given the fact that speech motor learning, which involves paired auditory and somatosensory inputs, results in changes to speech perceptual representations, somatosensory inputs may also be involved in learning or adaptive processes of speech perception. Here we show that repetitive pairing of somatosensory inputs and sounds, such as occurs during speech production and motor learning, can also induce a change of speech perception. We examined whether the category boundary between /ɛ/ and /a/ was changed as a result of perceptual training with orofacial somatosensory inputs. The experiment consisted of three phases: Baseline, Training, and Aftereffect. In all phases, a vowel identification test was used to identify the perceptual boundary between /ɛ/ and /a/. In the Baseline and the Aftereffect phase, an adaptive method based on the maximum-likelihood procedure was applied to detect the category boundary using a small number of trials. In the Training phase, we used the method of constant stimuli in order to expose participants to stimulus variants which covered the range between /ɛ/ and /a/ evenly. In this phase, to mimic the sensory input that accompanies speech production and learning in an experimental group, somatosensory stimulation was applied in the upward direction when the stimulus sound was presented. A control group (CTL) followed the same training procedure in the absence of somatosensory stimulation. When we compared category boundaries prior to and following paired auditory-somatosensory training, the boundary for participants in the experimental group reliably changed in the direction of /ɛ/, indicating that the participants perceived /a/ more than /ɛ/ as a consequence of training. In contrast, the CTL did not show any change. Although a limited number of participants were tested, the perceptual shift was reduced and almost eliminated 1 week later. Our data suggest that repetitive exposure of somatosensory inputs in a task that simulates the sensory pairing which occurs during speech production, changes perceptual system and supports the idea that somatosensory inputs play a role in speech perceptual adaptation, probably contributing to the formation of sound representations for speech perception.

Keywords: somatosensory stimulation, perceptual adaptation, multisensory integration, production-perception link, auditory representation

****Correspondence:***
Takayuki Ito
takayuki.ito@gipsa-lab.grenoble-inp.fr

Specialty section:
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

Received: *19 December 2021*
Accepted: *24 March 2022*
Published: *26 April 2022*

Citation:
Ito T and Ogane R (2022)
Repetitive Exposure to Orofacial
Somatosensory Inputs in Speech
Perceptual Training Modulates Vowel
Categorization in Speech Perception.
Front. Psychol. 13:839087.
doi: 10.3389/fpsyg.2022.839087

INTRODUCTION

Speech perception is auditory in nature but it is also an interactive process involving other sensory inputs. For example, visual information coming from a speaker's face helps in the identification of speech sounds in a noisy environment (Sumby and Pollack, 1954). Incongruent visual information from facial movements likewise affects speech perception (McGurk and MacDonald, 1976). Recent studies have demonstrated that somatosensory inputs also contribute to the perception of speech. When air-puffs, similar to those associated with a plosive speech sound (such as /p/), were presented to the skin, perception was biased in the direction of the corresponding sound (Gick and Derrick, 2009). When somatosensory stimulation using facial skin deformation was applied in conjunction with the speech sounds, vowel perception was systematically biased (Ito et al., 2009). In a vowel identification task on a "head/had" continuum, the presented vowels were perceived more as "head" when an upward skin stretch was applied, more as "had" when the skin stretch was downward, and there was no effect with backward skin stretch. A similar effect has been observed in both children and adults using a vowel continuum between /e/ and /ø/ (Trudeau-Fisette et al., 2019). When the skin stretch was backward, the presented sounds were perceived more as /e/, a vowel in which lip spreading is involved in production. A somatosensory influence on perception is not limited to vowel categorization, but is also observed in word segmentation in lexical processing (Ogane et al., 2020). The segmentation boundary changed depending on the placement of somatosensory stimulation in relation to the key vowel in a test phrase. While these studies suggest a potential role of somatosensory inputs in speech perception, the specific contribution of the somatosensory system is unknown.

Given that orofacial somatosensory inputs normally provide articulatory information in the context of speech production (Johansson et al., 1988; Ito and Gomi, 2007; Ito and Ostry, 2010), somatosensory effects on speech perception may be production related. This idea was initially proposed in the Motor Theory of Speech Perception (Liberman et al., 1967), and extended in the Direct Realist perspective (Fowler, 1986) and the Perception-for-Action-Control theory (Schwartz et al., 2012). The possible contribution of the sensorimotor system to perception has mostly focused on the motor system. For example, activity in brain motor areas has been observed during speech perception (Wilson et al., 2004; Skipper et al., 2005), and the perception of speech sounds can be modulated by applying transcranial magnetic stimulation to the premotor cortex (Meister et al., 2007; D'Ausilio et al., 2009). At a behavioral level, when speech articulation is simultaneous with listening, the perception of speech sounds is altered (Sams et al., 2005; Mochida et al., 2013; Sato et al., 2013). However, speech motor outflow always occurs in conjunction with correlated somatic input. While somatosensory function might be considered part of motor system, the somatosensory system may work independently in the perception of speech sounds since there is a direct influence and interaction between the somatosensory and auditory system in situations other than speech perception (Foxe et al., 2000; Beauchamp et al., 2008). Thus, investigating somatosensory function in speech perception may be important in clarifying the link between speech production and perception.

The contribution of somatosensory inputs to speech perception has been examined in the context of motor learning. Previous studies showed that adapting to different external environments during production changes the vowel category boundary (Nasir and Ostry, 2009). Similar perceptual changes have been reported in studies of adaptation to altered auditory feedback (Shiller et al., 2009; Lametti et al., 2014). Although both motor outputs and somatosensory inputs are involved in the speech motor learning tasks used in these previous studies, Ohashi and Ito (2019) specifically demonstrated that somatosensory inputs on their own can contribute to the recalibration of perception. That study applied additional somatosensory stimulation during adaptation to altered auditory feedback and assessed changes to the category boundary of fricative consonants. They observed perceptual recalibration in conjunction with somatosensory stimulation, suggesting that repetitive exposure to somatosensory inputs during learning can be a key to changing or recalibration of the speech perceptual representation.

In addition to motor learning, repetitive exposure to sensory stimuli also induces changes to sensory processing. In the case of speech, the phonetic boundary between two neighboring speech sounds can be biased away from the one that is repetitively presented as an adapter in training, which is known as selective adaptation (Eimas and Corbit, 1973). Similar effects can be seen in visual speech perception (Jones et al., 2010). This type of sensory adaptation has been frequently investigated in non-linguistic processing. In the visual domain, after looking at a high-contrast visual image, a low-contrast portion of a test image briefly appears invisible (e.g., Kohn, 2007). Similarly, after prolonged observation of a waterfall, an illusory upward motion can be induced when observing a static image (Mather et al., 1998). This effect has also been demonstrated in multisensory environments, including selective adaptation in audio-visual speech (Roberts and Summerfield, 1981; Saldaña and Rosenblum, 1994; Dias et al., 2016). In case of ambiguous speech sounds, visual information from speech movements changes auditory perception (Bertelson et al., 2003). Speech sounds also change visual speech perception (Baart and Vroomen, 2010). If somatosensory inputs contribute to the formulation or calibration of speech perceptual representations, repetitive exposure to orofacial somatosensory stimulation, such as occurs normally in conjunction with speech production and learning, may recalibrate the representation of speech sounds. If this adaptive change persists following perceptual training, then training with somatosensory stimulation may potentially be used as a tool for speech training and rehabilitation.

The present study examined whether repetitive exposure to orofacial somatosensory stimulation during a speech perception task changes the perception of speech sounds. To test this idea, we here focused on the category boundary between the vowels /ε/ and /a/ and applied orofacial somatosensory stimulation, specifically facial skin deformation, as used in previous studies (Ito et al., 2009; Trudeau-Fisette et al., 2019;

Ogane et al., 2020). The use of orofacial somatosensory stimulation is premised on the assumption that skin receptors provide kinesthetic information (Johansson et al., 1988; Ito and Gomi, 2007; Ito and Ostry, 2010). Given that somatosensory stimulation involving facial skin deformation changed the category boundary between "head" and "had" in on-line manner (Ito et al., 2009), training with the same auditory-somatosensory pairing may change or recalibrate the vowel category boundary in purely auditory perceptual tests. We carried out perceptual training paired with somatosensory stimulation and assessed changes to the category boundary. It might be expected, based on our prior work using a simple perceptual classification task (Ito et al., 2009), that upward skin stretch during vowel identification on a /ε/ to /a/ continuum would bias perception toward /ε/. However, if the effect of the training task on perception is similar to selective adaptation mentioned above, training might be accompanied by a perceptual shift toward /a/. Either perceptual change would suggest that the somatosensory system contributes to the link between speech production and perception, and that somatosensory inputs can help in the processing of speech sounds in ambiguous situations.

MATERIALS AND METHODS

Participants

Thirty native speakers of French participated in the experiment. The participants were all healthy young adults who reported normal hearing. All participants signed informed consent forms approved by the Local Ethical Committee of the University Grenoble Alpes (CERNI: Comité d'Ethique pour les Recherches non Interventionnelles: Avis-2015-03-03-62 or CERGA: Comité d'Ethique pour la Recherche, Grenoble Alpes: Avis-2018-12-11-4).

Auditory Stimulation

We focused on vowel categorization using an /ε/ to /a/ continuum, based on a previous study which showed a clear somatosensory effect on speech perception (Ito et al., 2009). These vowels were followed by the /f/ sound which is associated with a closing movement after the vowel production. The stimulus continuum was synthesized by using an iterative Burg algorithm for estimating spectral parameters. The procedure involved shifting the first (F1) and the second (F2) formant frequencies in equal steps from values observed for /ε/ to those associated with /a/. The stimulus sound was recorded by male speaker of French. The first and second formant values for the endpoint stimuli were 561 Hz and 1630 Hz for /ε/, and 712 Hz and 1203 Hz for /a/. A forty-six-step continuum was produced for the adaptive testing procedure used in Baseline and Aftereffect tests; a subset of these stimuli was selected for use in perceptual training (see below).

Somatosensory Stimulation

We used facial skin stretch applied by a robotic device to produce somatosensory stimulation (Phantom 1.0; SensAble Technologies). The experimental setup is shown in **Figure 1A**. Plastic tabs (2 cm × 3 cm) were attached to the skin lateral to the oral angle on each side of the face. These tabs were connected to the robotic device through thin wires. The wires were supported by wire supports to avoid contact with the facial skin. The skin was stretched when the robotic device applied force to the wires. The temporal profile of the applied force was a single cycle of a 3-Hz sinusoid with 2N peak force (see **Figure 1B**). Based on the previous finding that the upward skin stretch induced a relatively large change in vowel categorization judgments between "head" and "had" (Ito et al., 2009), we applied the skin stretch in an upward direction.

Perceptual Test and Adaptation Training

The main test was consisted of three phases: Baseline, Training, and Aftereffect (see **Figure 1C**). In all three phases, an identification test using the vowels /ε/ and /a/ was involved. The stimuli were presented through head-phones at a comfortable volume. On each trial, participants were asked to identify whether the sound was /εf/ or /af/ by pressing a key on a keyboard.

In the main perceptual training portion of the study, the method of constant stimuli (MCS) was used in order to expose participants to values between /ε/ and /a/ evenly during the training. We used 10 of 46 steps on the /ε/ to /a/ continuum (Nos. 1, 6, 11, 16, 21, 26, 31, 36, 41, and 46) and presented them 10 times each in pseudo-random order. Each training block consisted of 100 trials. This was repeated 5 times. In total, 500 stimuli were presented. For the experimental group which received somatosensory training (SOMA), somatosensory stimulation was applied on each trial. The temporal relationship between the sound stimulus and somatosensory stimulation is shown in **Figure 1B**. For the control group (CTL), we carried out the same training including the setup of the robot, but in absence of somatosensory stimulation.

In Baseline and Aftereffect tests, we used an adaptive method based on the maximum-likelihood (MLL) procedure to estimate the vowel category boundary (Shen and Richards, 2012). The benefit of this procedure is its ability to estimate the psychometric function and the associated category boundary with a relatively small number of responses in comparison to other conventional methods such as MCS. However, sounds near to the perceptual boundary are primarily tested. In this procedure, the auditory test stimulus on each trial is determined in an adaptive fashion based on the stimulus that provides the most information about the shape of the psychometric function. All stimuli on the forty-six-step continuum were used in this procedure. Each of the perceptual tests consisted of four 17-trial blocks. The first two blocks of the Baseline phase were removed from the analysis as familiarization trials for the identification task.

In order to examine if the effect of paired auditory-somatosensory training persisted 1 week later, we also repeated the Post-test using the same procedure as in the Aftereffect phase, based on MLL procedure. Five of 15 participants participated in the Post-test.

Data Analysis

We calculated the probability that the participant identified the presented vowel as /a/. We estimated the psychometric

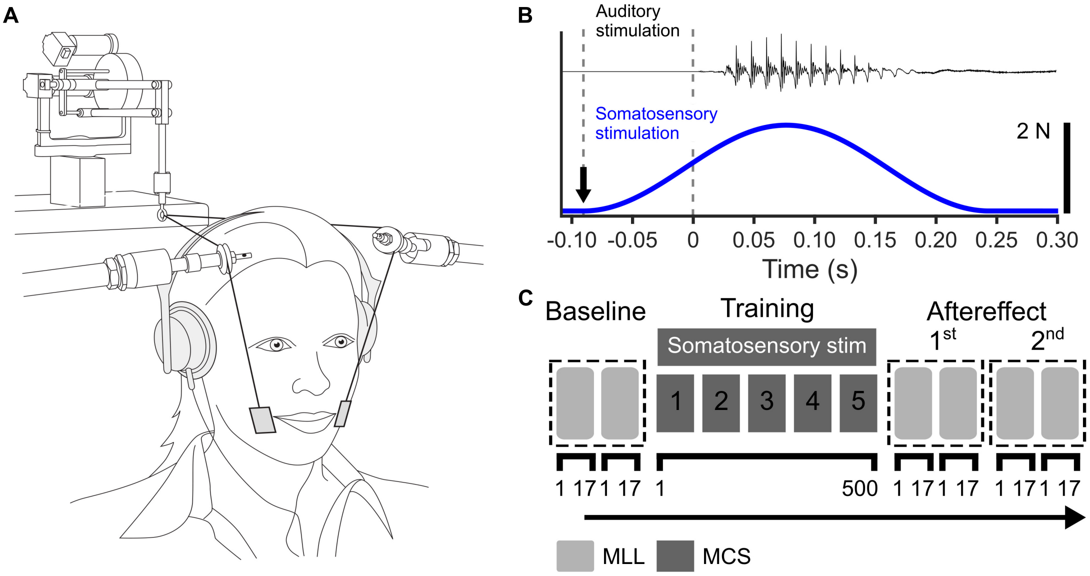

FIGURE 1 | (A) Experimental setup for somatosensory stimulation using facial skin deformation, reproduced from Ito et al. (2009). **(B)** Time course of auditory stimulus (top) and applied force during somatosensory stimulation (bottom). The black arrow represents the onset of somatosensory stimulation. **(C)** Experimental procedure in the auditory-somatosensory perceptual adaptation test. MLL represents the maximum likelihood procedure and MCS represents the perceptual test based on the method of constant stimuli.

function for each 17-trial block of the MLL procedure (Baseline, Aftereffect and Post-test) and for each 100-trial block of the MCS procedure (Training), and obtained estimates of the category boundary as the 50% value of the psychometric function. The baseline value for the category boundary was obtained by averaging the two blocks of the Baseline phase. In the Aftereffect and Post-test phases, we also averaged separately the first two (1st set) and the second two blocks (2nd set). The obtained category boundaries were normalized by dividing by the baseline boundary value.

To examine whether the category boundary changed following perceptual training, we applied a one sample *t*-test to the normalized perceptual boundary immediately following training (average of the first two blocks of the Aftereffect). This normalized perceptual boundary was also compared between control and somatosensory training groups using a Linear Mixed-Effects (LME) Models analysis with nlme package in R (Pinheiro et al., 2022). In the LME model including the following analyses, participants were always considered as a random effect.

We also applied a LME analysis to evaluate whether the perceptual boundary changed over the course of training (Training phase). Fixed factors were groups (CTL and SOMA) and blocks (1, 2, 3, 4, and 5). A separate one-sample *t*-test was also used to examine whether the category boundary averaged over the course of training was different than baseline.

The LME analysis was likewise used to evaluate the possible presence of persistent effect at a one-week delay. For this evaluation, we first compared changes between the 1st and 2nd sets in the Aftereffect phase, with groups (CTL and SOMA) and sets (1st and 2nd) as fixed factors. *Post hoc* tests with Bonferroni correction were carried out to compare all possible combinations using the multcomp package in R (Hothorn et al., 2008). Second, we compared the Aftereffect and Post-test phases in the five participants that completed both. In this analysis, we extracted the category boundaries for these participants in the Baseline, Aftereffect and Post-test measures and calculated separately the normalized boundary in the Aftereffect and Post-test, as described above. We used a LME analysis to assess whether the normalized boundary was different in the Aftereffect and Post-test measures. Fixed factors in this analysis were phases (Aftereffect and Post-test) and sets (1st and 2nd).

RESULTS

Shift of Category Boundary Due to the Training

Figure 2A shows representative results for the estimated psychometric function prior to and following training in the two conditions (CTL and SOMA). As shown here, the category boundary shifted in the direction of /ε/ following training with somatosensory stimulation (SOMA, solid blue line in the right panel of **Figure 2A**), indicating that the participants perceived /a/ more than /ε/ as an aftereffect. This shift was not observed following training in the control condition (CTL, solid gray line in the left panel of **Figure 2A**). Averaged perceptual changes with standard errors are shown in **Figure 2B**.

The amplitude of the shift was significantly different from zero [-0.163 ± 0.040, average ± s.e., t (14) = -4.08, $p < 0.005$] after the training with somatosensory stimulation

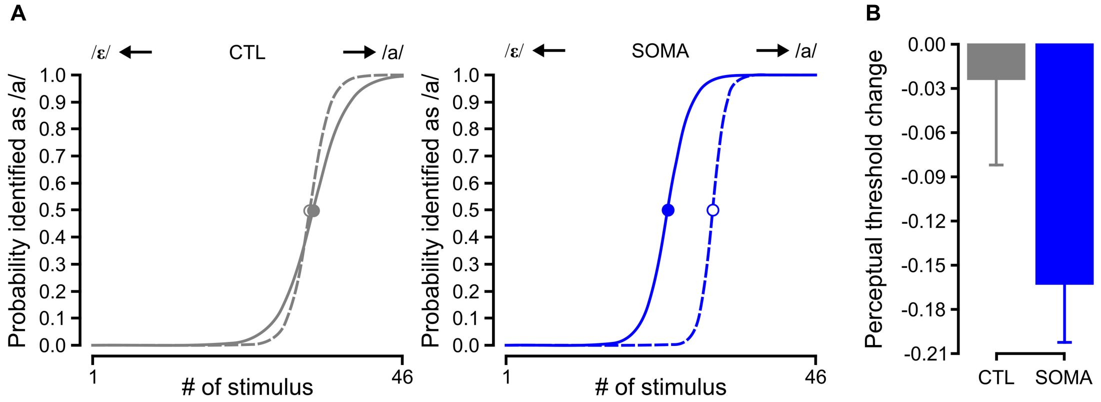

FIGURE 2 | (A) The estimated psychometric function in Baseline (dashed) and Aftereffect (solid) phases for control and somatosensory conditions in representative participants. Filled (Aftereffect) and open (Baseline) circles represent the 50% crossover value of the psychometric function. The left panel in gray shows the participant response in the control condition (CTL); the right panel in blue shows the response in the condition that received somatosensory stimulation (SOMA). **(B)** Averaged perceptual change of the 50% crossover values for the control (left, gray) and somatosensory condition (right, blue), respectively. Error bars represent standard errors across participants.

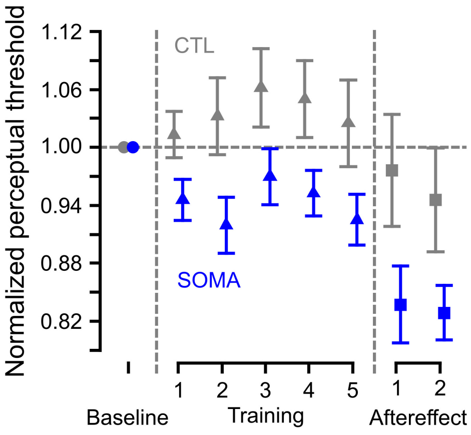

FIGURE 3 | Category boundary values normalized to the baseline category boundary over the course of the experimental procedures. Blue represents the somatosensory condition and gray represents the control condition. Error bars represent standard error across participants.

(SOMA). In the control condition (CTL), the magnitude of the shift was not different than zero [−0.023 ± 0.058, average ± s.e., t (14) = −0.41, $p > 0.6$]. A comparison between groups using a LME analysis also showed a significant effect [χ^2 (1) = 3.88, $p < 0.05$]. These results indicate that the repetitive exposure to somatosensory stimulation during auditory perceptual training can alter the perceptual category boundary as a consequence.

Perceptual Change During the Training

Figure 3 shows the averaged trajectory of the estimated category boundary over the course of training. In order to examine whether the category boundary changed during the training, we applied the LME analysis to the category boundary estimates obtained over the course of the Training phase. We found that there was no significant interaction between groups (CTL and SOMA) and blocks [χ^2 (4) = 1.93, $p > 0.7$], indicating that the pattern of change in the category boundary was similar for the two groups. In addition, we did not find a difference across blocks [χ^2 (4) = 8.70, $p > 0.06$], indicating that there was no change in category boundary over the course of the training. There was a significant overall difference between groups [χ^2 (1) = 5.19, $p < 0.03$], indicating that the mean value for the category boundary during training in the somatosensory condition was different than that in the control condition. A one-sample t-test using the data averaged across blocks showed that values were significantly different from zero in the SOMA condition [t (14) = −2.83, $p < 0.02$], but not in the CTL condition [t (14) = 1.03, $p > 0.3$]. This indicates that participants' perception in the SOMA condition shifted in the direction of /ɛ/ during the training phase. This change was not induced in the control condition. The results suggest that there were no temporal changes over the course of training in either group, while somatosensory stimulation induced an overall shift in perception in the experimental condition.

Persistence of Category Boundary Shift

Although we had limited data to evaluate, we assessed whether the perceptual aftereffect persists following training. We first compared category boundary estimates between the first two (1st set) and the last two blocks (2nd set) of the Aftereffect phase using the LME analysis. There were no significant differences between 1st and 2nd sets [χ^2 (1) = 0.34, $p > 0.5$]. *Post hoc* tests conducted for the individual conditions found no difference between sets for SOMA ($p > 0.7$) and CTL ($p > 0.6$), respectively. There was a significant interaction between groups and sets [χ^2 (2) = 11.62, $p < 0.01$]. *Post hoc* tests indicated a significant difference between SOMA and CTL in the 1st set ($p < 0.05$), and a marginal difference in the 2nd set ($p = 0.073$). There

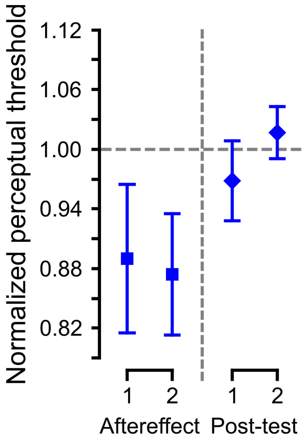

FIGURE 4 | Normalized category boundary in the Aftereffect phase and Post-test (1 week later). Error bars represent the standard error across participants.

is also a significant difference between groups [χ^2 (1) = 5.14, p < 0.05], such that the values for the SOMA group are different than those of the CTL group. Separate one-sample *t*-tests showed that the overall mean in the Aftereffect phase in the SOMA group was reliably different than zero [t (14) = −5.26, p < 0.01], whereas this was not the case for the CTL group [t (14) = −0.85, p > 0.4]. This indicates that the category boundary change following somatosensory stimulation persisted during Aftereffect trials.

We also evaluated if the perceptual change due to paired auditory-somatosensory simulation persisted one-week later. Since only five participants from SOMA group were tested following the one-week delay, we evaluated the effects using five datasets for these participants (one pre-training set, two following training and two after a 1 week delay). The averaged data with standard errors for each set of the Aftereffect and Post-test trials are shown in **Figure 4**. The LME analysis showed a significant difference between the Aftereffect and Post-test values [χ^2 (1) = 4.56, p < 0.05], but not between the 1st and 2nd sets [χ^2 (1) = 0.11, p > 0.7] nor in the interaction [χ^2 (2) = 4.95, p > 0.08], suggesting that the somatosensory effect was not present one-week later.

DISCUSSION

The present study examined whether repetitive exposure to somatosensory stimulation in a task which was designed to mirror the pairing of auditory and somatosensory stimulation that occurs during production and speech learning, changes the perceptual representation of speech sounds. We evaluated whether the category boundary between /ε/ and /a/ changed from before to after training with somatosensory stimulation. The somatosensory stimulation involved facial skin deformation in an upward direction. In previous work using a simple perceptual classification task (Ito et al., 2009), this manipulation was found to change the perception of speech sound toward /ε/ when presented with the speech stimuli during training. We found instead that the category boundary between /ε/ and /a/ was in fact shifted toward /ε/, that is, participants perceived /a/ more than /ε/ after training. Although a relatively small number of participants was available for a subsequent post-training test, the shift in the perceptual boundary did not appear to be present 1 week later. The results nevertheless suggest that repetitive exposure to somatosensory inputs associated with facial skin deformation is capable of changing the perceptual representation of speech sounds.

The results of the present study are in line with previous work showing that facial skin deformation changes the perception of speech sounds in on-line testing (Ito et al., 2009; Trudeau-Fisette et al., 2019; Ogane et al., 2020). Repetitive exposure to somatosensory stimulation during speech motor learning may account for the contribution of somatosensation to speech perception (Ohashi and Ito, 2019). The current results are consistent with this hypothesis. Paired auditory-somatosensory input during training, alters subsequent auditory perceptual judgments, suggesting a contribution of somatosensory exposure to speech perception and the presence of a link between speech production and perception.

As mentioned in the Introduction, the category boundary between vowels can be changed when we are repeatedly exposed to one of two vowels, a phenomenon in the speech perception literature known as selective adaptation. Eimas and Corbit (1973) originally showed that the category boundary between /ba/ and /pa/ was shifted toward /ba/, that is, the participants perceived /pa/ more than /ba/ after the training with repetitive exposure of /ba/. The pattern is similar to that of the current finding in which the category boundary shifted toward /ε/ when repetitive somatosensory stimulation, which has been previously shown to modify the perceived speech sound toward /ε/, was applied. A possible mechanism, originally proposed by Eimas and Corbit (1973) is fatigue of a linguistic feature detector as a result of repetitive exposure to the corresponding speech sounds. Kleinschmidt and Jaeger (2016) proposed another possible explanation associated with distributional learning. Although the current results cannot address this debate directly, the current somatosensory effect would fit with either account of selective adaptation. Specifically, in the control condition, we present all values on the speech-sound continuum an equal number of times. As a result, there is no effect on the category boundary, presumably because the entire speech sound representation is affected equally. Both linguistic feature detector and learned distribution accounts would predict a similar result under these conditions. Somatosensory stimulation in the present study serves to modify the perceived sound toward /ε/. Both feature detection fatigue for /ε/ and modification of the stimulation distribution would predict this effect which in turn, may be reflected as a change in the category boundary.

Selective adaptation in speech perception is considered to be an auditory phenomenon when the presented sounds are unambiguous. Previous studies using the McGurk effect (McGurk and MacDonald, 1976) showed that selective adaptation to auditory inputs was induced even when the sound was perceived differently as a result of incongruent visual stimulation (Roberts and Summerfield, 1981; Saldaña and Rosenblum, 1994; Dias et al., 2016). While selective adaptation is observed in visual speech perception (Baart and Vroomen, 2010), Dias et al. (2016) suggested that visual information may not contribute to selective adaptation in the McGurk effect. In the case of the present study, since training with somatosensory stimulation was found to induce a change in the auditory category boundary, the interaction mechanism may be different than in auditory-visual speech perception. This would be consistent with a previous study which found that simultaneous somatosensory and visual stimulation in speech perception did not interact with one other in terms of the behavioral response (Ito et al., 2021). Since somatosensory inputs to speech sounds affect the N1 peak in the auditory ERP (Ito et al., 2014), which is considered to be associated with the initial extraction of vowel related information (Näätänen and Picton, 1987), somatosensory inputs may affect the auditory processing of speech sounds at a lower level of vowel processing. However, somatosensory inputs also affect word segmentation in lexical decisions (Ogane et al., 2020). One future direction is a direct test of the idea that somatosensory stimulation may affect visual speech perception. Baart and Vroomen (2010) showed adaptation in visual speech perception of ambiguous lipread tokens after the exposure to an incongruent sound. Somatosensory stimulation may work in a similar fashion by providing information which disambiguates visual stimuli instead of sounds.

It is important to know how long the training effect lasts. The duration of training phase was limited and as a result, this type of sensory adaptation may not last for a long time. In the case of speech motor learning using altered auditory feedback, the post-training effects on adaptation gradually decrease over the course of the following 100 trials (Purcell and Munhall, 2006; Villacorta et al., 2007). The motion aftereffects described in the Introduction persist for several seconds to minutes. Although it is unknown yet how long selective adaptation lasts, this effect may only persist for a short period, as is the case with sensory adaptation in other modalities. Since the effect of somatosensory training was essentially absent one-week later in the limited number of participants that were tested, the current persistence of a somatosensory aftereffect on speech perception may be similar to other sensory aftereffects. In future investigations, it would be desirable to evaluate shorter periods after training, such as 1 h later, rather than one-week. These types of adaptation including selective adaptation are induced when transient stimulation is presented, and hence when the additional stimulation is removed, particularly after brief periods of training, it is difficult to maintain the adapted perception without receiving additional stimulation. Since this additional stimulation does not exist outside of the laboratory, it may limit the use of the current procedure for speech training or rehabilitation. Nevertheless, the current finding is in line with the more general idea that receiving specific paired of auditory-somatosensory inputs, such as occurs over long periods of time during speech motor training, may underlie a durable contribution of somatosensory inputs to the speech perceptual representation.

In previous work using the same speech sound continuum, in which skin stretch trials were interleaved with no-stretch trials, a change in perception of speech sound toward /ε/ was observed (Ito et al., 2009). In contrast, in the present study, multiple blocks of 100 trials with skin stretch were used. The repetitive pairing of auditory-somatosensory stimulation may have produced a quite different perceptual effect that favored "selective adaptation." As a result, participants might have perceived /a/ more than /ε/ even in the first 100 trial block.

A potential technical limitation of the present study is that perceptual boundary between speech sounds could not be estimated over a smaller number of trials. While the MCS provides a reasonable estimate of the perceptual boundary, it requires a relatively large number of trials. In the present study, we used 100 trials, and hence, only five estimates of the boundary value were obtained over the course of the current training. The procedure may thus lack the sensitivity needed to correctly capture any changes which might occur. We used a maximum likelihood procedure before and after training as an alternative to improve the possibility of detecting changes over a shorter period of time. However, this method still needs more than ten trials (17 trials in the current case) and requires that participants listen to sounds near to their perceptual boundary rather than over the entire sound continuum, which is the case with the MCS. Due to this technical limitation, it is difficult to characterize perceptual behavior over the course of training. Further investigation is required to better understand the time-course of the current adaptation mechanism.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Comité d'Ethique pour les Recherches non Interventionnelles, Comité d'Ethique pour la Recherche, Grenoble Alpes. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TI contributed to the conception and design of the study. TI and RO collected the data, wrote the first draft of the manuscript, involved in subsequent drafts of the manuscript. RO organized the database, performed the statistical analysis (all under TI's guidance), and produced the figures. Both authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

This work was supported by National Institute on Deafness and Other Communication Disorders Grant R01-DC017439.

ACKNOWLEDGMENTS

We thank Thomas Routier for assistance with data collection.

REFERENCES

Baart, M., and Vroomen, J. (2010). Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading. *Neurosci. Lett.* 471, 100-x-103. doi: 10.1016/j.neulet.2010.01.019

Beauchamp, M. S., Yasar, N. E., Frye, R. E., and Ro, T. (2008). Touch, sound and vision in human superior temporal sulcus. *Neuroimage* 41, 1011–1020. doi: 10.1016/j.neuroimage.2008.03.015

Bertelson, P., Vroomen, J., and De Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychol. Sci.* 14, 592–597. doi: 10.1046/j.0956-7976.2003.psci_1470.x

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., and Fadiga, L. (2009). The motor somatotopy of speech perception. *Curr. Biol.* 19, 381–385. doi: 10.1016/j.cub.2009.01.017

Dias, J. W., Cook, T. C., and Rosenblum, L. D. (2016). Influences of selective adaptation on perception of audiovisual speech. *J. Phon.* 56, 75–84. doi: 10.1016/j.wocn.2016.02.004

Eimas, P. D., and Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cogn. Psychol.* 4, 99–109. doi: 10.1016/0010-0285(73)90006-6

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct–realist perspective. *J. Phon.* 14, 3–28. doi: 10.1016/S0095-4470(19)30607-2

Foxe, J. J., Morocz, I. A., Murray, M. M., Higgins, B. A., Javitt, D. C., and Schroeder, C. E. (2000). Multisensory auditory-somatosensory interactions in early cortical processing revealed by high-density electrical mapping. *Brain Res. Cogn. Brain Res.* 10, 77–83. doi: 10.1016/s0926-6410(00)00024-0

Gick, B., and Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature* 462, 502–504. doi: 10.1038/nature08572

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biom. J.* 50, 346–363. doi: 10.1002/bimj.200810425

Ito, T., and Gomi, H. (2007). Cutaneous mechanoreceptors contribute to the generation of a cortical reflex in speech. *Neuroreport* 18, 907–910. doi: 10.1097/WNR.0b013e32810f2dfb

Ito, T., Gracco, V. L., and Ostry, D. J. (2014). Temporal factors affecting somatosensory-auditory interactions in speech processing. *Front. Psychol.* 5:1198. doi: 10.3389/fpsyg.2014.01198

Ito, T., Ohashi, H., and Gracco, V. L. (2021). Somatosensory contribution to audio-visual speech processing. *Cortex* 143, 195–204. doi: 10.1016/j.cortex.2021.07.013

Ito, T., and Ostry, D. J. (2010). Somatosensory contribution to motor learning due to facial skin deformation. *J. Neurophysiol.* 104, 1230–1238. doi: 10.1152/jn.00199.2010

Ito, T., Tiede, M., and Ostry, D. J. (2009). Somatosensory function in speech perception. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1245–1248. doi: 10.1073/pnas.0810063106

Johansson, R. S., Trulsson, M., Olsson, K. Â, and Abbs, J. H. (1988). Mechanoreceptive afferent activity in the infraorbital nerve in man during speech and chewing movements. *Exp. Brain Res.* 72, 209–214. doi: 10.1007/BF00248519

Jones, B. C., Feinberg, D. R., Bestelmeyer, P. E. G., DeBruine, L. M., and Little, A. C. (2010). Adaptation to different mouth shapes influences visual perception of ambiguous lip speech. *Psychon. Bull. Rev.* 17, 522–528. doi: 10.3758/PBR.17.4.522

Kleinschmidt, D. F., and Jaeger, T. F. (2016). Re-examining selective adaptation: fatiguing feature detectors, or distributional learning? *Psychon. Bull. Rev.* 23, 678–691. doi: 10.3758/s13423-015-0943-z

Kohn, A. (2007). Visual adaptation: physiology, mechanisms, and functional benefits. *J. Neurophysiol.* 97, 3155–3164. doi: 10.1152/jn.00086.2007

Lametti, D. R., Rochet-Capellan, A., Neufeld, E., Shiller, D. M., and Ostry, D. J. (2014). Plasticity in the human speech motor system drives changes in speech perception. *J. Neurosci.* 34, 10339–10346. doi: 10.1523/JNEUROSCI.0108-14.2014

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.* 74, 431–461.

Mather, G., Verstraten, F., and Anstis, S. (1998). *The Motion Aftereffect: A Modern Perspective*. Cambridge, MA: A Bradford Book.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., and Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Curr. Biol.* 17, 1692–1696. doi: 10.1016/j.cub.2007.08.064

Mochida, T., Kimura, T., Hiroya, S., Kitagawa, N., Gomi, H., and Kondo, T. (2013). Speech Misperception: Speaking and Seeing Interfere Differently with Hearing. *PLoS One* 8:e68619. doi: 10.1371/journal.pone.0068619

Näätänen, R., and Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology* 24, 375–425. doi: 10.1111/j.1469-8986.1987.tb00311.x

Nasir, S. M., and Ostry, D. J. (2009). Auditory plasticity and speech motor learning. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20470–20475. doi: 10.1073/pnas.0907032106

Ogane, R., Schwartz, J.-L., and Ito, T. (2020). Orofacial somatosensory inputs modulate word segmentation in lexical decision. *Cognition* 197:104163. doi: 10.1016/j.cognition.2019.104163

Ohashi, H., and Ito, T. (2019). Recalibration of auditory perception of speech due to orofacial somatosensory inputs during speech motor adaptation. *J. Neurophysiol.* 122, 2076–2084. doi: 10.1152/jn.00028.2019

Pinheiro, J., Bates, D., and R Core Team. (2022). *nlme: Linear and nonlinear mixed effects models*. Available online at: https://svn.r-project.org/R-packages/trunk/nlme/.

Purcell, D. W., and Munhall, K. G. (2006). Adaptive control of vowel formant frequency: evidence from real-time formant manipulation. *J. Acoust. Soc. Am.* 120, 966–977. doi: 10.1121/1.2217714

Roberts, M., and Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Percept Psychophys.* 30, 309–314. doi: 10.3758/bf03206144

Saldaña, H. M., and Rosenblum, L. D. (1994). Selective adaptation in speech perception using a compelling audiovisual adaptor. *J. Acoust. Soc. Am.* 95, 3658–3661. doi: 10.1121/1.409935

Sams, M., Möttönen, R., and Sihvonen, T. (2005). Seeing and hearing others and oneself talk. *Brain Res. Cogn. Brain Res.* 23, 429–435. doi: 10.1016/j.cogbrainres.2004.11.006

Sato, M., Troille, E., Ménard, L., Cathiard, M.-A., and Gracco, V. (2013). Silent articulation modulates auditory and audiovisual speech perception. *Exp. Brain Res.* 227, 275–288. doi: 10.1007/s00221-013-3510-8

Schwartz, J. L., Basirat, A., Menard, L., and Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): a perceptuo-motor theory of speech perception. *J. Neurolinguist* 25, 336–354. doi: 10.1016/J.Jneuroling.2009.12.004

Shen, Y., and Richards, V. M. (2012). A maximum-likelihood procedure for estimating psychometric functions: thresholds, slopes, and lapses of attention. *J. Acoust. Soc. Am.* 132, 957–967. doi: 10.1121/1.4733540

Shiller, D. M., Sato, M., Gracco, V. L., and Baum, S. R. (2009). Perceptual recalibration of speech sounds following speech motor learning. *J. Acoust. Soc. Am.* 125, 1103–1113. doi: 10.1121/1.3058638

Skipper, J. I., Nusbaum, H. C., and Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *Neuroimage* 25, 76–89. doi: 10.1016/j.neuroimage.2004.11.006

Sumby, W. H., and Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Trudeau-Fisette, P., Ito, T., and Ménard, L. (2019). Auditory and Somatosensory Interaction in Speech Perception in Children and

Adults. *Front. Hum. Neurosci.* 13:344. doi: 10.3389/fnhum.2019.00344

Villacorta, V. M., Perkell, J. S., and Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *J. Acoust. Soc. Am.* 122, 2306–2319. doi: 10.1121/1.2773966

Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702. doi: 10.1038/nn1263

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 24 May 2022
doi: 10.3389/fcomm.2022.858520

Immediate Integration of Coarticulatory Cues for /s/-Retraction in American English

*Jacob B. Phillips**

Department of Linguistics, University of Chicago, Chicago, IL, United States

Coarticulatory "noise" has long been presumed to benefit the speaker at the expense of the listener. However, recent work has found that listeners make use of that variation in real time to aid speech processing, immediately integrating coarticulatory cues as soon as they become available. Yet sibilants, sounds notable for their high degree of context-dependent variability, have been presumed to be unavailable for immediate integration, requiring that listeners hold all cues in a buffer until all relevant cues are available. The present study examines the cue integration strategies that listeners employ in the perception of prevocalic and pre-consonantal sibilants. In particular, this study examines the perception of /s/-retraction, an ongoing sound change whereby /s/ is realized approaching /ʃ/ as a result of long distance coarticulation from /r/. The study uses eye tracking in the Visual World Paradigm in order to determine precisely when listeners are able to utilize the spectral cues in sibilants in different phonological environments. Results demonstrate that while in most instances listeners wait until more cues are available before considering the correct candidate, fixation accuracy increases significantly throughout the sibilant interval alone. In the pre-consonantal environment, immediate integration strategies were strengthened when the coarticulatory cues of retraction were stronger and when they were more predictable. These findings provide further evidence that context-dependent variation can be helpful to listeners, even on the most variable of sounds.

Keywords: speech perception, sound change, cue integration, ambiguity, sibilants, coarticulation

Edited by:
Georgia Zellou,
University of California, Davis, United States

Reviewed by:
Jonathan Manker,
Rice University, United States
Navin Viswanathan,
The Pennsylvania State University (PSU), United States

Correspondence:
Jacob B. Phillips
jbphillips@uchicago.edu

Specialty section:
This article was submitted to Frontiers in Communication, a section of the journal Frontiers in Communication

Received: *20 January 2022*
Accepted: *28 April 2022*
Published: *24 May 2022*

Citation:
Phillips JB (2022) Immediate Integration of Coarticulatory Cues for /s/-Retraction in American English. Front. Commun. 7:858520. doi: 10.3389/fcomm.2022.858520

1. INTRODUCTION

Coarticulation has often been considered to be a process that primarily aids the speaker, as it decreases the articulatory distance between two adjacent gestures and may therefore decrease articulatory effort (Lindblom, 1990). Coarticulation can work in both directions, with preceding sounds affecting following sounds (carry-over coarticulation) and following sounds affecting preceding sounds (anticipatory coarticulation). For some researchers, coarticulation has been viewed as a process that not only aids the speaker, but also actively hinders the listener, as the increased degree of coarticulation between gestures may render the speech signal more ambiguous (Stevens and Keyser, 2010). Under this view, phonetic ambiguity arises because the coarticulated speech deviates substantially from the citation form, which in turn may diminish potential phonological contrasts between two sounds. Such accounts propose that in listener-directed speech, speakers will minimize coarticulation, thereby increasing articulatory effort and thus consequentially avoiding any potential ambiguity that could inhibit listener comprehension. However, research on elicited clear speech has found that speakers do not reduce anticipatory coarticulation in clear speech compared to normal conditions (Matthies et al., 2001). Other work has demonstrated that coarticulation is increased, rather than decreased, for more confusable words, suggesting that coarticulation itself, rather than the reduction of it, may be a form of hyperarticulation (Scarborough, 2004). This finding holds for different languages (English vs. French), different directions (anticipatory vs. carryover), and different types of coarticulation (vowel-to-vowel coarticulation, which could potentially reduce a phonemic contrast, and vowel nasalization, which would not imperil a phonemic contrast).

In this vein, many approaches to coarticulation propose that it is a process that mutually aids both speaker and listener. That is, while coarticulation may result in diminished phonological contrasts and greater deviation from citation forms, it provides listeners with helpful contextual information from adjacent phones, potentially easing the perception of the sounds in their relevant contexts. Treating coarticulation as a process that creates ambiguity disregards the role of context: What is ambiguous in isolation is not only clear, but beneficial, in context. This approach is built into varying, and often times, conflicting models of speech perception. Gesturalists posit that successful speech perception is accomplished by recovering the articulatory gestures that the speaker produced (Fowler, 1986, 1996, 2006) or intended (Liberman and Mattingly, 1985). In a gesturalist account, listeners make use of coarticulatory variation in order to better recover those gestures (e.g., Viswanathan et al., 2010). In contrast, auditorist approaches posit that listeners rely exclusively on their fine-tuned auditory systems and need not recruit their experiences as speakers (Lotto and Kluender, 1998; Diehl et al., 2004). In an auditorist account, our general auditory systems are sufficiently developedto account for and utilize context-dependent variation based off the acoustic signal alone (Lotto and Kluender, 1998; Holt and Kluender, 2000). Yet while these theories have much they disagree on, both approaches agree that coarticulation is more than something than can be overcome—it provides useful context-dependent information that aids, rather than hinders, speech perception[1]. Similarly, models of speech perception, like TRACE, also incorporate the perceptual benefit of coarticulation in word recognition (Elman and McClelland, 1986).

This perceptual benefit of coarticulation has been demonstrated robustly in the laboratory. Listeners are able to correctly identify the target word more quickly and accurately when more coarticulatory information is present (Martin and Bunnell, 1981; Whalen, 1991; Connine and Darnieder, 2009). Similarly, listeners are more accurate in identifying deleted segments when coarticulatory information is present than when it is missing (Ostreicher and Sharf, 1976). The development of eye-tracking has allowed researchers to examine the perceptual benefit of coarticulation in real-time, asking not only how contextual information improves task accuracy, but also how listeners use the cues of coarticulation to anticipate upcoming sounds. For example, Beddor et al. (2013) examined the perception of anticipatory nasal coarticulation, presenting listeners with two pictures that varied only on the presence or absence of the nasal consonant, e.g., *scent* /sɛnt/ and *set* /sɛt/. Beddor et al. (2013) found that listeners can anticipate the upcoming nasal, looking to an image like *scent* off coarticulation alone even before the nasal consonant is heard. However, the absence of nasality was not equally helpful; that is, oral vowels did not lead to faster or more accurate looks to words like *set*. These findings not only bolster earlier behavioral accounts that coarticulatory information is helpful to the listener, but also show that listeners can use that information as soon as it becomes available. This process by which listeners immediately use available information in lexical identification has been referred to as immediate integration or a "cascade" perception strategy. In addition to nasalization, immediate integration has been demonstrated for a variety of contrasts in which cues become available sequentially, like stop voicing (McMurray et al., 2008).

In contrast, a "buffer" strategy or delayed integration strategy describes the process by which listeners hold the unfolding information in a buffer until all relevant cues are available before beginning lexical identification. Galle et al. (2019) have suggested that, unlike for stops and nasalization, listeners use a buffer strategy for sibilant perception. That is, despite the potential for listeners to use spectral cues to immediately distinguish sibilants like /s/ and /ʃ/, the primary cues in contrasting the two places of articulation, listeners wait for the formant transitions, a secondary cue. Galle et al. (2019) explored a variety of possible explanations for this observation ranging from an auditory account that sibilants make contrasts at higher frequencies than other sounds to the possibility that spectral cues in sibilants are not reliable enough or simply too context-dependent and variable. The latter hypothesis is of particular interest as it contradicts findings of immediate integration for coarticulation like Beddor et al. (2013), which illustrate that context-dependent variation in vowels can be immediately integrated and help anticipate upcoming sounds due to the structured and predictable nature of coarticulation. The present study puts these different accounts in conversation through an examination of cue integration strategies for sibilant coarticulation. In particular, this study examines sibilant coarticulation in preconsonantal environments where coarticulation is predictable, but no formant transitions are available such that listeners could rely on those potential secondary cues.

The focus of the present study is /s/-retraction, a sound change in progress in many varieties of English by which /s/ approaches /ʃ/ in the context of /r/, most notably in /str/ clusters[2] So for a speaker exhibiting /s/-retraction, a word like *street* /strit/ may sound more like *shtreet* /ʃtrit/. This

[1] For a recent review the role of context-dependent perception in gesturalist and auditorist approaches (see Stilp, 2019).

[2] For a detailed discussion of the production, perception, and phonological accounts of /s/-retraction (see Phillips, 2020).

has been observed in various dialects of American English (Shapiro, 1995; Durian, 2007; Baker et al., 2011; Gylfadottir, 2015; Wilbanks, 2017; Smith et al., 2019; Phillips, 2020) as well as varieties of English across the Anglophone world (Lawrence, 2000 for New Zealand; Glain, 2013; Bailey et al., 2022 for the United Kingdom; Stevens and Harrington, 2016 for Australia). Additionally, corpus studies have demonstrated that /s/-retraction is advancing in apparent time in the United States (Gylfadottir, 2015; Wilbanks, 2017). At its core, /s/-retraction can be viewed as a coarticulatory process by which /s/ is produced with greater tongue body retraction and lip protrusion so as to minimize articulatory distance between /s/ and /r/ (Baker et al., 2011; Smith et al., 2019). These small articulatory changes can have outsized acoustic effects, resulting in a sibilant more characteristic of an /ʃ/ than /s/ (Baker et al., 2011). However, despite resulting in a sibilant that may surface between /s/ and /ʃ/, /s/-retraction need not necessarily create confusion due to the phonotactic restrictions of English: While /s/ and /ʃ/ are contrastive prevocalically, only /ʃ/ precedes /r/ and only /s/ precedes all other consonants. Thus, English phonotactic restrictions on preconsonantal sibilants create an environment in which extreme coarticulation is unfettered by potential lexical confusability.

In order to address these notions of ambiguity and confusability, the perception of /s/-retraction, not just its production, needs to be examined, and while a growing body of work has examined the production of the sound change, scant work has examined listeners' perception of it. In one perception study, Kraljic et al. (2008) found that exposure to sibilants ambiguous between /s/ and /ʃ in /str/ clusters, like *industry* /ɪndəʃtri/, where retraction is expected, does not alter an individual's /s/-/ʃ/ categorization as strongly as ambiguous sibilants in unpredictable prevocalic environments, like *dinosaur* /daɪnəʃɔr/. In another, Phillips and Resnick (2019) examined the perception of onset sibilants in nonce words, like *strimble* or *shtrimble*, where listeners may be less constrained by lexical/phonotactic restrictions. Phillips and Resnick (2019) found that individuals were less categorical, and less likely to perceive an /ʃ/ onset in /str/ clusters, where /s/-retraction is more expected, than in /spr/ and /skr/ clusters. Both studies demonstrate that listeners have detailed context-dependent knowledge about /s/-retraction based off their experiences. The present study asks how listeners use that information in real time. That is, can listeners use their knowledge of context-dependent spectral variation in sibilants in order to more quickly and accurately identify the target word? And crucially, by looking at perception in real time, we can examine how listeners deal with a case of ephemeral ambiguity: The ambiguity between the sibilants in these environments exists only for a short amount of time until disambiguating information, like the ultimate presence or absence of /r/, follows. Additionally, through an examination of a sound change in progress, rather than a potentially more stable coarticulatory pattern like vowel nasalization, the present study builds on previous work on cue integration to ask whether listeners are consistent and uniform in their use of a changing cue.

2. METHODS

2.1. Participants

A total of 52 participants were recruited from the University of Chicago undergraduate subject pool and received course credit or payment. All participants were between 18 and 22 years of age. Thirty-seven participants identified as female, 15 as male, and none as non-binary or transgender. Just over half of the participants (29) identified as straight/heterosexual. Similarly, 29 participants identified as white. Participants were geographically distributed across the United States, with more participants reporting growing up in suburban areas (34) compared to urban (15) or rural (3) environments. All participants were self-reported native speakers of North American English with no history of hearing loss, language and communication disorders, or any other medical conditions commonly associated with cognitive impairment. An additional nine individuals participated in this study but were excluded from analysis due to non-native status, language or neurological disorders, and/or non-attentive responses.

2.2. Stimuli

The target stimuli were designed to manipulate the degree of retraction in sibilant clusters to examine whether the anticipatory cues of /r/ presence can influence lexical processing. The stimuli thus included the relevant /sCr/ and /sC/ clusters as well as simplex prevocalic /s/ and /ʃ/. There were three sets of near minimal pair quadruplets, one for each place of articulation of the intervening stop: *sit-spit-spritz-shit* (bilabial), *sing-sting-string-shingle* (alveolar), and *sip-skip-script-ship* (velar). Stop initial quadruplets also varying in place of articulation and presence of /r/ were included as fillers: *pick-prick-brick-big* (bilabial), *tip-trip-drip-dip* (alveolar), and *kit-crypt-grip-gift* (velar).

The original auditory stimuli were produced by a college-aged male from Illinois. The speaker recorded five repetitions of each target word in the carrier phrase "Now select X." All stimuli materials were recorded at 48,000 Hz with a Shure SM10A head-mounted microphone in a sound-attenuated booth.

To provide control and consistency over the degree of retraction in the onset sibilants, all stimuli were cross-spliced. The onset sibilants from the target words were deleted and replaced with a sibilant digitally mixed from prevocalic /s/ (*sip*) and /ʃ/ (*ship*) at different scaling ratios, using a Praat script originally created by Darwin (2005). For each /sCr/ cluster, three degrees of retraction were used to test the hypothesis that listeners attend to coarticulation on the sibilant to anticipate the presence of absence of an upcoming /r/: minimal, moderate, and extreme retraction. The retraction conditions were designed in consultation with previous examinations of /s/-retraction (e.g., Baker et al., 2011), with the talker's natural production of /s/ in these environments, and with the researcher's perception. In all /sCr/ clusters, the minimal retraction condition was designed to exhibit less retraction than the speaker produces naturally and to be perceived clearly as an /s/; the stimuli was digitally mixed with 30% /ʃ/ and 70% [s] for /str/ clusters and 10% /ʃ/ and 90% [s] for

TABLE 1 | Scaling factors used in stimuli creation.

	Minimal retraction		Moderate retraction		Extreme retraction	
	/s/	**/ʃ/**	**/s/**	**/ʃ/**	**/s/**	**/ʃ/**
/spr/	0.90	0.10	0.60	0.40	0.30	0.70
/str/	0.70	0.30	0.40	0.60	0.10	0.90
/skr/	0.90	0.10	0.60	0.40	0.30	0.70
			Across conditions			
			/s/	**/ʃ/**		
		/s/	1.00	0.00		
		/sp/	0.90	0.10		
		/st/	0.90	0.10		
		/sk/	0.90	0.10		
		/ʃ/	0.00	1.00		

TABLE 2 | Pairing of visual images organized by place of articulation and onset environment.

	s–ʃ	s– sC	sC–sCr	sCr–ʃ
/p/	Sit–shit	Sit–spit	Spit–spritz	Spritz–shit
/t/	Sing–shingle	Sing–sting	Sting–string	String–shingle
/k/	Sip–ship	Sip–skip	Skip–script	Script–ship
	T–D	**T–Tr**	**Tr–Dr**	**Dr–D**
/p/	Pick–big	Pick–prick	Prick–brick	Brick–big
/t/	Tip–dip	Tip–trip	Trip–drip	Drip–dip
/k/	Kit–gift	Kit–crypt	Crypt–grip	Grip–gift

/spr/ and /skr/ clusters. The moderate retraction condition was designed to exhibit increased degrees of retraction to the model talker's natural production and to be perceived approaching the /s/-/ʃ/ boundary; the stimuli digitally mixed with 60% /ʃ/ and 40% [s] for /str/ clusters and 40% /ʃ/ and 60% [s] for /spr/ and /skr/ clusters. Finally, the extreme retraction condition was designed to contain twice again as much retraction as the speaker produced naturally and be perceived clearly as an /ʃ/; the onsets digitally mixed with 90% /ʃ/ and 10% [s] for /str/ cluster and 70% /ʃ/ and 30% [s] for /spr/ and /skr/ clusters. The /sC/ onsets did not differ between conditions, and digitally mixed with 10% /ʃ/and 90% [s] in the minimal, moderate, and extreme retraction conditions, consistent with the talker's natural production. So that all stimuli underwent similar manipulations, the onset sibilants in prevocalic environments were also cross-spliced; however, they were not digitally mixed since no retraction would be expected prevocalically. The scaling factors used for the creation of each onset environment can be seen in **Table 1**. Furthermore, to reduce the effects of stimuli manipulation, the stop-initial fillers were cross-spliced with onsets containing manipulated degrees of aspiration.

For each target word, four free and publicly available clipart images were selected, resized, and gray-scaled. Four naïve volunteers selected the image that best corresponded the intended word. In order to control for differences of style, darkness, or image resolution, all images were redrawn by hand, making adjustments to remove any text or distracting features. The hand-drawn images were then scanned, gray-scaled, and resized to 550 × 550 pixels.

2.3. Procedure

After informed consent, participants were first familiarized with the images and their associated lexical items. This was more straightforward for nouns and high frequency words than for adjectives, verbs, and low frequency items. Participants were first introduced to the images and their accompanying orthographic labels in a randomized order. Participants were asked to read the label aloud and explain to the researcher how the label relates to the image. To explain the task, the researcher provided two examples verbally: for a picture of a dog with the label "dog," the researcher would simply say "this is a dog," but for a picture of a cheetah with a label "fast," the researcher would say "cheetahs are fast." Following this connection-making task, participants were then shown images in a randomized order without the accompanying orthographic labels and asked to reproduce the corresponding label. All participants exhibited 100% accuracy in the label reproduction task, demonstrating that they had successfully associated the lexical items with the images. No subsequent effect of grammatical category or lexical frequency was observed.

For the identification task, participants were randomly assigned to one of three retraction conditions: minimal, moderate, or extreme retraction. Participants were seated in front of a Tobii T-60 eye-tracker, with a sampling rate of 60 Hz that was recalibrated for each participant. Two images, rather than the typical four, were presented in each trial in a modified Visual World Paradigm (Allopenna et al., 1998). This modification, in which only a single target and competitor image are presented without distractors, was also utilized by Beddor et al. (2013) for an examination of cue integration strategies for anticipatory nasalization. It should be noted that this modification may increase the sensitivity and likelihood that participants will exhibit looks to the target image sooner, centering the question of *can* listeners immediately use the spectral cues of sibilants rather than do they necessarily use them in normal conversations. The images were paired according to contrasts in **Table 2**, with the critical pair for the present study being /s/ vs. /ʃ/, e.g., *sing* vs. *shingle*, and /sC/ vs. /sCr/, e.g., *sting* vs. *string*. Thus, in each trial, participants were only considering one potential sibilant contrast, either a phonemic contrast between /s/ and /ʃ/ or context-dependent variation within a category. Participants were first asked to scan the screen and, after identifying the images, focus on a fixation cross in the center of the screen, equidistant between both images. Once a fixation on the cross was detected, a red box was displayed surrounding the cross. Participants were able to click on the box to play the auditory stimuli "Now select [word]," e.g., "Now select *sting*." Participants were directed to click on the corresponding image as quickly as they could, which signaled the end of the trial and automatically advanced to the next item. Each trial lasted roughly 5 s. Left and right eye movement was recorded throughout the experiment. A sample trial slide is provided in

FIGURE 1 | A sample trial illustrating the presentation of the images, *sting* **(left)** and *string* **(right)**, and fixation cross.

Figure 1 to illustrate how the visual stimuli and response options were presented.

2.4. Measurements

Both accuracy and gaze measurements were collected. Trial accuracy was defined by clicking on the correct image corresponding to the auditory stimuli. Although a trial may be ambiguous during the onset sibilant portion, the ultimate presence or absence of /r/ would disambiguate the stimulus. Thus, all participants exhibited >95% accuracy in image selection.

Participants' eye gaze was monitored from the initial display of the target and competitor images, through the cross fixation, until 2,000 ms following the onset on the target word or until they clicked on an image, whichever came first. Although eye gaze was tracked for both left and right eyes, analysis was conducted on the right eye exclusively. Unlike trial accuracy, which identifies whether the participant selected the correct image corresponding to the auditory stimuli, gaze measurements identify precisely when the target or competitor lexical item were considered, before the ultimate decision to click on the correct image was made. This not only provides a much more fine-grained temporal resolution than reaction time for mouse clicks, but also allows for an examination of alternative phonological candidates for the ultimately unambiguous stimuli.

The online measurement selected for analysis for the present experiment was the proportion of correct fixations over time, which is determined by examining the accuracy of each individual fixation. A fixation was determined to be a correct fixation if the right eye gaze fell within the 550 × 550 pixel region containing the image corresponding to the auditory stimuli. Fixations were binned into 20 ms windows. A proportion of 0 for a given bin means that there were no trials in the relevant condition during which eye gaze was detected within the 550 × 550 pixel region containing the target image. This means that all participants' gaze was directed at the fixation cross, the competitor image, or anywhere else on the screen other than the target image. Thus, it is not the proportion of target versus competitor fixations, but rather the proportion of target versus non-target fixations. Similarly, a proportion of 1 means that in all trials a target fixation was detected within the specified 20 ms window.

2.5. Predictions

The specific hypotheses for participants' eye gaze are as follows:

Hypothesis 1 states that listeners will make immediate use of spectral cues to distinguish /s/ and /ʃ/ in prevocalic environments. This hypothesis is formulated in direct response to the buffer strategy observed for prevocalic sibilants by Galle et al. (2019). Under this hypothesis, correct fixations on /s/ or /ʃ/ will emerge during the onset sibilant, when only spectral information can distinguish the two places of articulation. This hypothesis is tested by TimeWindow in /s/-/ʃ/ pairs. If listeners exhibit increased proportion of correct fixations over the sibilant interval, it suggests that they are using a cascade strategy for integrating the spectral cues of the onset sibilants, contra (Galle et al., 2019). If listeners wait until the onset of the vowel to increase their proportion of correct fixations, this suggests that a buffer strategy is used for sibilants. If such a buffer strategy is observed, Hypotheses 2–4 ask if this is true for pre-consonantal sibilants as well as prevocalic sibilants.

Hypothesis 2 states that listeners will make use of the coarticulatory cues in predicting the phonological context of the sibilant and do so as soon as those cues are available. Under this hypothesis, correct fixations on /sC/ or /sCr/ will emerge during the onset sibilant, before the ultimate absence or presence of /r/ disambiguates the stimuli. If such a pattern is observed, this demonstrates that like with vowel-nasal coarticulation observed by Beddor et al. (2013), long distance rhotic-sibilant coarticulation is immediately available and beneficial to the listener. Like in the prevocalic model, this hypothesis is again tested by TimeWindow, but in examination of /sC/-/sCr/ pairs. Additionally, this hypothesis is tested by RetractionCondition (minimal, moderate, or extreme) and its interaction with TimeWindow, examining if stronger cues of retraction, and thus stronger cues of coarticulation, increase the proportion of correct fixations over the course of the sibilant. If Hypothesis 2 is confirmed, then the following hypotheses stand to be tested:

Hypothesis 3 states that a retracted /s/ is a better indicator of rhotic presence than a non-retracted /s/ is for rhotic absence. That is, does a more retracted, i.e., more /ʃ/-like, onset predict an /sCr/ cluster better than a less retracted, i.e., more /s/-like, onset predicts an /sC/ cluster. A confirmation of this hypothesis would demonstrate that the cues of /s/-retraction are more useful in speech processing than the absence of such cues, much like the findings of Beddor et al. (2013) that a nasal vowel is a better cue of an upcoming nasal stop than an oral vowel is of an upcoming oral stop. This is tested by Cluster in examination of /sC/-/sCr/ pairs and its interaction with TimeWindow, where more correct fixations are predicted for /sCr/ clusters than /sC/ clusters over the course of the sibilant.

Hypothesis 4 states that the cues of /s/-retraction are a better indicator of rhotic presence in /str/ clusters compared to /spr/ and /skr/ clusters. A confirmation of this hypothesis would demonstrate that listeners have detailed phonological knowledge about /s/-retraction as a sound change in progress, with greater degrees of retraction observed in /str/ clusters (Baker et al., 2011), and adjust their expectations accordingly. This hypothesis is tested by PLACE of articulation (alveolar, bilabial, and velar) in examination of /sC/-/sCr/ pairs and its interaction with TIMEWINDOW and CLUSTER, where more correct fixations are predicted for alveolar clusters than bilabial and velar clusters, particularly in /str/ clusters, over the course of the sibilant. This hypothesis thus requires that listeners not only use phonological knowledge about the upcoming rhotic, but also about the upcoming stop before that stop is perceived.

3. RESULTS

The results of this experiment are presented in two sections. First, in Section 3.1, the results from the /s/-/ʃ/ pairs are presented, asking if listeners attend to the spectral cues of the onset sibilants immediately or whether they hold them in a buffer until vocalic information is available. This section tests Hypothesis 1. Secondly, in Section 3.2, the results from the /sC/-/sCr/ pairs are presented, which tests Hypotheses 2–4. These pairs ask whether listeners can use the coarticulatory cues of /s/-retraction immediately to anticipate the presence of an upcoming /r/.

3.1. Prevocalic Results

The prevocalic analysis asks if listeners can use spectral cues present over the course of the sibilant in order to correctly identify a prevocalic sibilant /s/ and /ʃ/, distinguishing words like *sip* /sɪp/ vs. *ship* /ʃɪp/. To test this, generalized linear mixed-effects models with a logit link function were fit to the accuracy of a given fixation (1,0) using the `glmer()` function in the `lme4` package (Bates et al., 2015) in R (R Core Team, 2015). As it takes ∼200 ms to plan and execute an eye movement and as the sibilant was 180 ms in duration, the model examined eye movements during the 180 ms window that began 200 ms following the onset of the stimulus sibilant. The prevocalic model includes trial ORDER (1–384, scaled), TIMEWINDOW of the sibilant (1–180, binned into 20 ms windows and scaled), and ONSET (/s/ and /ʃ/; treatment-coded with /s/ as base) as fixed effects. RETRACTIONCONDITION (minimal, moderate, and extreme) was not included as the prevocalic onsets were not manipulated between conditions. Self-reported responses for demographic categories like GENDER, SEXUALITY, AGE, and REGION did not reach a significance threshold of 0.05 and were pruned from the final models. Preliminary models for the different onset pairings included all two- and three-way interactions between the fixed effects predictors. All interactions that did not reach a significance threshold of 0.05 were pruned from the final models. Additionally, the preliminary models included maximally specified random effects structures, with by-subject random slopes and intercepts, which were progressively simplified until convergence was achieved. The results of the prevocalic logistic regression are presented in **Table 3**. The inclusion of by-subject random intercepts and by-subject random slopes for trial ORDER and ONSET suggests significant individual variability with respect to these predictors. By-item random slopes and intercepts are not included as there is only one item per onset cluster, given the training and time constraints of the current design.

TABLE 3 | Model predictions for all main effects and interactions in fixation accuracy for /s/ vs. /ʃ/ onsets, *N* = 26,750.

	Est.	SE	z	*p*
Intercept	−0.47	0.17	−2.79	**0.005****
Order	0.02	0.06	0.31	0.758
TimeWindow	0.39	0.01	27.36	**<0.001*****
Onset-SH	−0.17	0.09	−1.75	0.081

*A positive value indicates a greater prediction of fixations on the target word. *p < 0.05; **p < 0.01; ***p < 0.001. All p values less than 0.05 are in bold.*

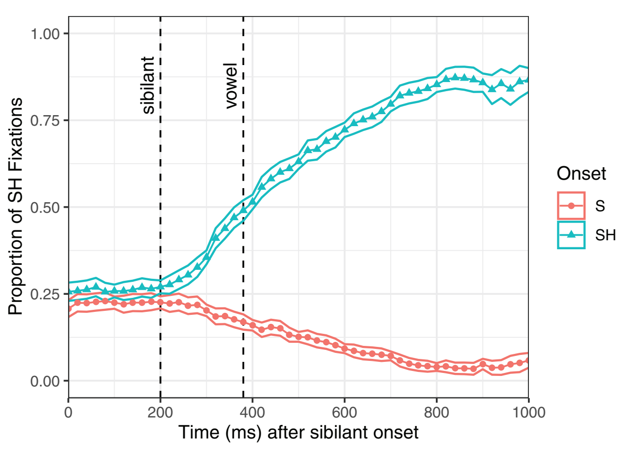

FIGURE 2 | Fixation proportion for /ʃ/ clusters (y-axis) by time following the sibilant onset (x-axis, binned into 20 ms windows) and onset category (color: /s/ = red circles, /ʃ/ = teal triangles). The vertical lines represent a 200 ms delay from the onset of the sibilant and vowel. A fixation executed during the sibilant interval would be observed between the black dashed vertical lines.

The negative intercept in the model ($z = -2.79, p = 0.005$) suggests that all else being equal, listeners are more likely to be looking anywhere other than the target image during the sibilant. However, the main effect of TIMEWINDOW ($z = 27.36, p < 0.001$) demonstrates that the proportion of correct fixations increases robustly over the course of the sibilant. The effect of TIMEWINDOW is visualized in **Figure 2**. Although the analysis is conducted on the proportion of correct fixations, I have chosen to visually present the proportion of /ʃ/ fixations. The primary choice in doing so is to allow the fixations for /s/ and /ʃ/ to visually diverge at the time at which the listener's eye gaze between the trials diverges. Unlike in the pre-consonantal stimuli, the prevocalic stimuli are cross-spliced but naturally produced, such that they potentially may be immediately disambiguated. Recall that while immediate disambiguation of sibilants has been demonstrated for /s/ and /ʃ/ in a gating task (Galle et al., 2019), immediate disambiguation has not been demonstrated in speech processing using eye tracking.

Figure 2 illustrates how the proportion of /ʃ/ fixations changes over the course of a trial. A trial with an /s/ onset is presented in red circles and a trial with an /ʃ/ onset is presented in teal triangles. For both /s/ and /ʃ/ onsets, participants begin with around one quarter of the fixations on the /ʃ/ image, which is supported by the intercept of the model. While the other fixations are not explicitly indicated in **Figure 2**, they may be to the cross equidistant between the images, where a participant's fixation is required to initiate the trial, or to the competing /s/ image. Since it takes ∼200 ms to plan and execute an eye movement, any fixations planned during the sibilant would be observed ∼200 ms later. Vertical lines are provided in **Figure 2** to indicate what sound was heard when a given eye movement was planned. Thus, if a look to the /ʃ/ image is planned during the sibilant it would be observed between the dashed lines. A look once the vowel has been heard and formant transitions, a secondary cue, are available would be observed following the second dashed line.

Preliminary inspection of **Figure 2** may first highlight that the most dramatic differences between the /s/ and /ʃ/ onsets is not observed until well after the vowel onset is heard. This suggests that in many trials, listeners wait until formant transitions are available to correctly identify the target word, keeping with Galle et al. (2019). However, I am primarily concerned with the fixation proportions *during* the sibilant interval, to ask specifically if listeners *can* use the spectral cues of sibilants even if they don't always do so. At the most basic level, this asks if accuracy of fixations increases over the course of the sibilant, which would be indicated by diverging predictions and steep slopes for /s/ and /ʃ/ onsets between the dashed lines. In **Figure 2**, a dramatic rise in proportion of /ʃ/ fixations is observed between the dashed lines for /ʃ/ onsets paralleled with a notable, but less dramatic, fall for /s/ onsets. Furthermore, the confidence intervals for /s/ and /ʃ/ diverge sharply and almost immediately during the sibilant interval. These visual findings are supported by the model with a significant main effect of TIMEWINDOW ($z = 27.36, p < 0.001$), which suggests that the proportion of correct fixations increases over the course of the sibilant. There is no significant effect of ONSET, either as a main effect or in interaction with any other effects, which suggests that listeners are equally accurate in their perception of /s/ and /ʃ/. However, as the inclusion of by-subject random slopes for ONSET improved model likelihood, there may be significant individual variation in the perception of the different sibilants.

3.2. Pre-consonantal Results

As the prevocalic analysis demonstrates that listeners are able to immediately use spectral cues to disambiguate two separate sibilants, the pre-consonantal analysis asks if listeners can use those same cues in order to predict the context of the sibilant. In these stimuli, the contrast is not between two phonemes but rather two phonological environments. The pre-consonantal model is fit on the same 180 ms window but for the /sC/–/sCr/ onsets and includes trial ORDER (1–384, scaled), TIMEWINDOW of the sibilant (1–180, binned into 20 ms windows and scaled), CLUSTER (/sC/ and /sCr/; treatment-coded with /sC/ as base), PLACE of articulation (alveolar, velar, and bilabial; Helmert-coded to first compare alveolar to the combined mean of velar and bilabial and then compare velar to bilabial), and RETRACTIONCONDITION (minimal, moderate, and extreme; treatment-coded with minimal as base) as fixed effects. Like with the prevocalic model, all non-significant interactions and predictors were pruned from the final model and random effects structure was progressively simplified until convergence was achieved. Results of the pre-consonantal model are presented in **Table 4**. The inclusion of by-subject random intercepts and by-subject random slopes for TRIALID, PLACE, and CLUSTER suggests significant individual variability with respect to these predictors.

Like in the prevocalic model, the significant negative intercept ($z = -2.37, p = 0.018$) suggests that participants are more likely to look away from the target image than toward it. And like in the prevocalic model, the main effect of TIMEWINDOW suggests that participants are more likely to look to the correct image over the course of the sibilant ($z = 2.48, p = 0.013$). This effect is noticeably smaller and less robust than in the prevocalic environment. While in the prevocalic environment the spectral cues are the primary cues in making the contrast between the two target items, in the pre-consonantal environment, the spectral cues are secondary coarticulatory cues present while the stimuli remain ambiguous until the ultimate presence or absence of /r/ disambiguates the candidates 77 ms after the end of the sibilant.

Fixations for the different retraction conditions, pooled across places of articulation, is illustrated in **Figure 3**. Although the model is fit on the accuracy of fixations, for the ease of visualization, I present the proportion of /sCr/ fixations. Again, vertical lines are provided as guideposts to what sound was heard when the eye movement was planned, including the following stop. In **Figure 3**, /sCr/ fixations rise noticeably in the moderate and extreme RETRACTIONCONDITION over the course of the sibilant, which is indicated by the positive slopes of the teal lines between the dashed vertical lines. Additionally, the proportions of /sCr/ fixations diverge for the moderate and extreme RETRACTIONCONDITION slightly at the end of sibilant period in both conditions, although the most noticeable divergence occurs after the sibilant ends during the stop period. These observations are supported by the interaction of TIMEWINDOW with RETRACTIONCONDITION in the regression, with more correct fixations predicted over the course of the sibilant in moderate and extreme retraction conditions (moderate: $z = 5.07, p < 0.001$; extreme $z = 3.43, p < 0.001$). These findings suggest that individuals are able to use the available coarticulatory cues of /s/-retraction in order to improve correct fixations, well before the onset of the disambiguating /r/. This interaction effect with RETRACTIONCONDITION also explains the relatively smaller main effect of TIMEWINDOW compared to the prevocalic model: While in the prevocalic /s/ and /ʃ/, helpful spectral cues are equally present in all stimuli, in the pre-consonantal stimuli, only few coarticulatory cues are available in the minimal retraction condition.

Figure 4 breaks down the findings by place of articulation. Visual inspection of the figure indicates a steeper teal line for /sCr/ clusters and divergence of the red /sC/ and teal /sCr/ confidence intervals in the alveolar onsets compared to the bilabial and velar onsets. This is supported by the model with the significant interaction of TIMEWINDOW, CLUSTER (SCR), and PLACE of articulation ($z = 2.82, p = 0.005$). Recall that place of articulation is Helmert-coded so the comparison made here is between alveolar onsets and the

TABLE 4 | Model predictions for all main effects and interactions in fixation accuracy for /sCr/ vs. /sC/ onsets, *N* = 27,067.

	Est.	SE	z	p
Intercept	−0.68	0.29	−2.37	**0.018***
Order	0.01	0.07	0.14	0.890
TimeWindow	0.07	0.03	2.48	**0.013***
Condition (Moderate)	−0.22	0.39	−0.57	0.569
Condition (Extreme)	−0.71	0.36	−1.98	0.053
Cluster (SCR)	−0.38	0.17	−2.19	**0.028***
Place (1)	−0.12	0.19	−0.64	0.522
Place (2)	−0.04	0.23	−0.19	0.851
TimeWindow × Condition (Moderate)	0.19	0.04	5.07	**<0.001*****
TimeWindow × Condition (Extreme)	0.11	0.03	3.43	**<0.001*****
TimeWindow × Cluster (SCR)	−0.03	0.03	−1.11	0.265
TimeWindow × Place (1)	−0.03	0.04	−0.81	0.418
TimeWindow × Place (2)	0.05	0.05	1.08	0.278
Cluster (SCR) × Place (1)	0.14	0.11	1.24	0.213
Cluster (SCR) × Place (2)	−0.25	0.13	−1.89	0.060
Cluster (SCR) × Condition (Moderate)	0.41	0.24	1.73	0.085
Cluster (SCR) × Condition (Extreme)	0.52	0.22	2.38	**0.017***
Place (1) × Condition (Moderate)	0.25	0.27	0.91	0.361
Place (1) × Condition (Extreme)	0.15	0.25	0.62	0.532
Place (2) × Condition (Moderate)	−0.07	0.33	−0.22	0.823
Place (2) × Condition (Extreme)	0.26	0.30	0.86	0.392
TimeWindow × Cluster (SCR) × Place (1)	0.17	0.06	2.82	**0.005**
TimeWindow × Cluster (SCR) × Place (2)	−0.04	0.07	−0.065	0.516
Cluster (SCR) × Place (1) × Condition (Moderate)	−0.64	0.16	−4.00	**<0.001*****
Cluster (SCR) × Place (1) × Condition (Extreme)	0.14	0.15	0.96	0.337
Cluster (SCR) × Place (2) × Condition (Moderate)	−0.03	0.18	−0.15	0.879
Cluster (SCR) × Place (2) × Condition (Extreme)	0.13	0.17	0.75	0.453

Place is Helmert-coded: Place (1) indicates alveolar compared to the mean of velar and bilabial; Place (2) indicates velar compared to bilabial. A positive value indicates a greater prediction of fixations on the target word.
**p* < 0.05; ***p* < 0.01; ****p* < 0.001. *All p values less than 0.05 are in bold.*

combined mean of velar and bilabial onsets. This suggests that listeners improve their consideration of the correct candidate most in /str/ clusters, precisely where /s/-retraction is both most expected and those cues are most available. No four-way interactions between TimeWindow, Cluster, Place, and RetractionCondition emerged as significant such that individuals were influenced most by greater levels of retraction in alveolar clusters. Rather, the results indicate that high degrees of retraction regardless of place of articulation are helpful to the listener and the spectral cues in /str/ clusters, which by nature of the stimuli always contain more cues of retraction than their bilabial and velar counterparts, aids the listener.

Additionally, the model suggests that other effects and interactions that do not have to do with the timing of sibilant can also influence the listener. Specifically, a main effect of Cluster emerged (SCR: $z = -2.19, p = 0.028$), such that individuals are less accurate in their consideration of /sCr/ clusters than /sC/ clusters across the board. This effect is counteracted in the extreme retraction condition by the interaction of Cluster (SCR) and RetractionCondition (moderate: $z = 1.73, p = 0.085$; extreme: $z = 2.38, p = 0.017$), which suggests that individuals are more accurate in their consideration of /sCr/ candidates when the highest degrees of retraction are available. Finally, a three-way interaction of interaction of Cluster (SCR), Place of articulation (alveolar compared to the mean of velar and bilabial), and RetractionCondition emerged as significant (moderate: $z = -4.00, p < 0.001$; extreme: $z = 0.96, p < 0.337$), such that the beneficial effects of the moderate retraction condition and the alveolar place of articulation are tempered in conjunction with one another.

The models and figures thus far pool data across 52 participants which can potentially obfuscate individual differences in processing styles. That is, we might ask do some participants use a buffer strategy while other participants use those cues more immediately indicative of a cascade strategy? In **Figure 5**, nine individual participants' fixation proportions are visualized, with three participants from each retraction condition. Fixations are pooled across places of articulation and confidence intervals are excluded due to the paucity of observations from a single individual. Participants are categorized into one of three patterns: delayed, buffer, and cascade integration. For participants who exhibit delayed looks,

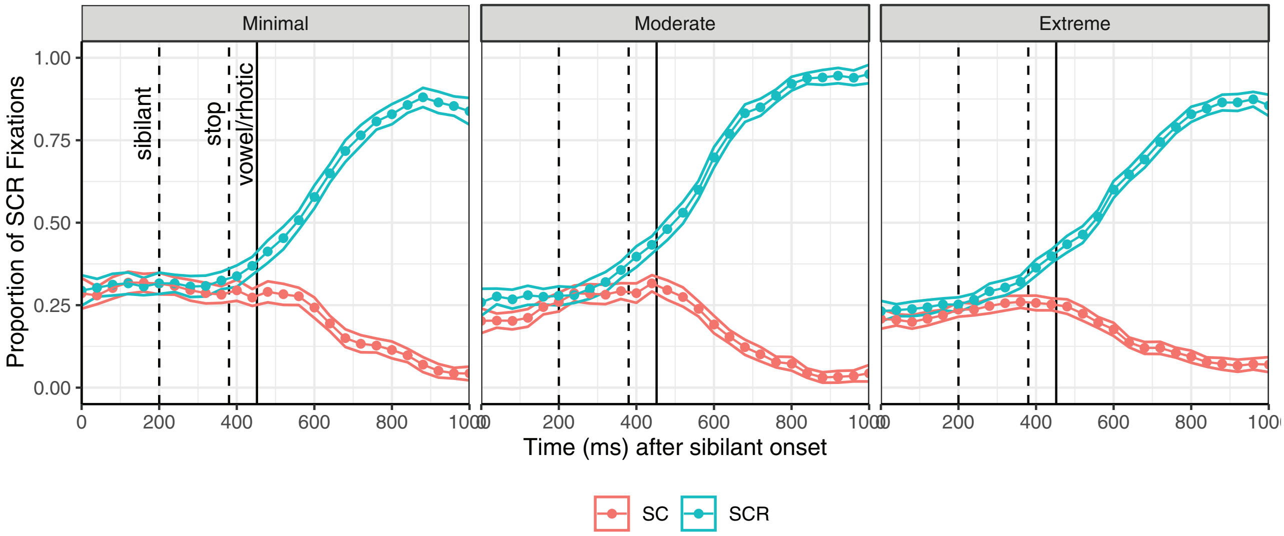

FIGURE 3 | Fixation proportion for /sCr/ clusters (y-axis) by time following the sibilant onset (x-axis, binned into 40 ms windows), cluster type (color: /sC/ = red circles, /sCr/ = teal triangles), and retraction condition (columns). The vertical lines represent a 200 ms delay from the onset of the sibilant, stop, and vowel/rhotic. A fixation executed during the sibilant interval would be observed between the black dashed vertical lines.

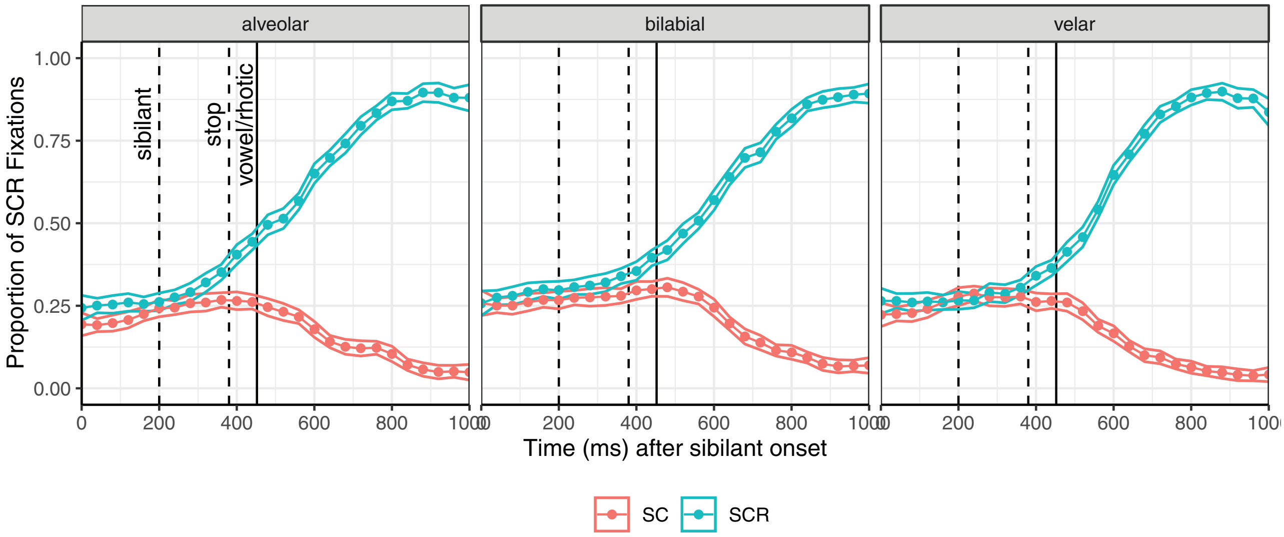

FIGURE 4 | Fixation proportion for /sCr/ clusters (y-axis) by time following the sibilant onset (x-axis, binned into 40 ms windows), cluster type (color: /sC/ = red circles, /sCr/ = teal triangles), and place of articulation (columns). The vertical lines represent a 200 ms delay from the onset of the sibilant, stop, and vowel/rhotic. A fixation executed during the sibilant interval would be observed between the black dashed vertical lines.

they begin with near 0% fixations on /sCr/ images, suggesting that they are often maintaining their gaze on the fixation cross, either because they are slower at directing their eye gaze or out of an effort to be a conscientious participant. Participants who exhibit delayed looks thus almost never exhibit clear indications of immediate integration such that their proportion of correct fixations increases during the sibilant interval. A second category is individuals who are looking to either the target or competitor image when the sibilant begins, but their consideration of /sCr/ and /sC/ images do not diverge until the stop or rhotic/vowel portion of the stimuli. These participants appear to exhibit a buffer strategy and wait to integrate the cues of retraction until additional information is available. Finally, the third pattern of participants is individuals who show evidence for increased consideration of the correct candidate during the sibilant portion alone, integrating the coarticulatory cues of /s/-retraction as

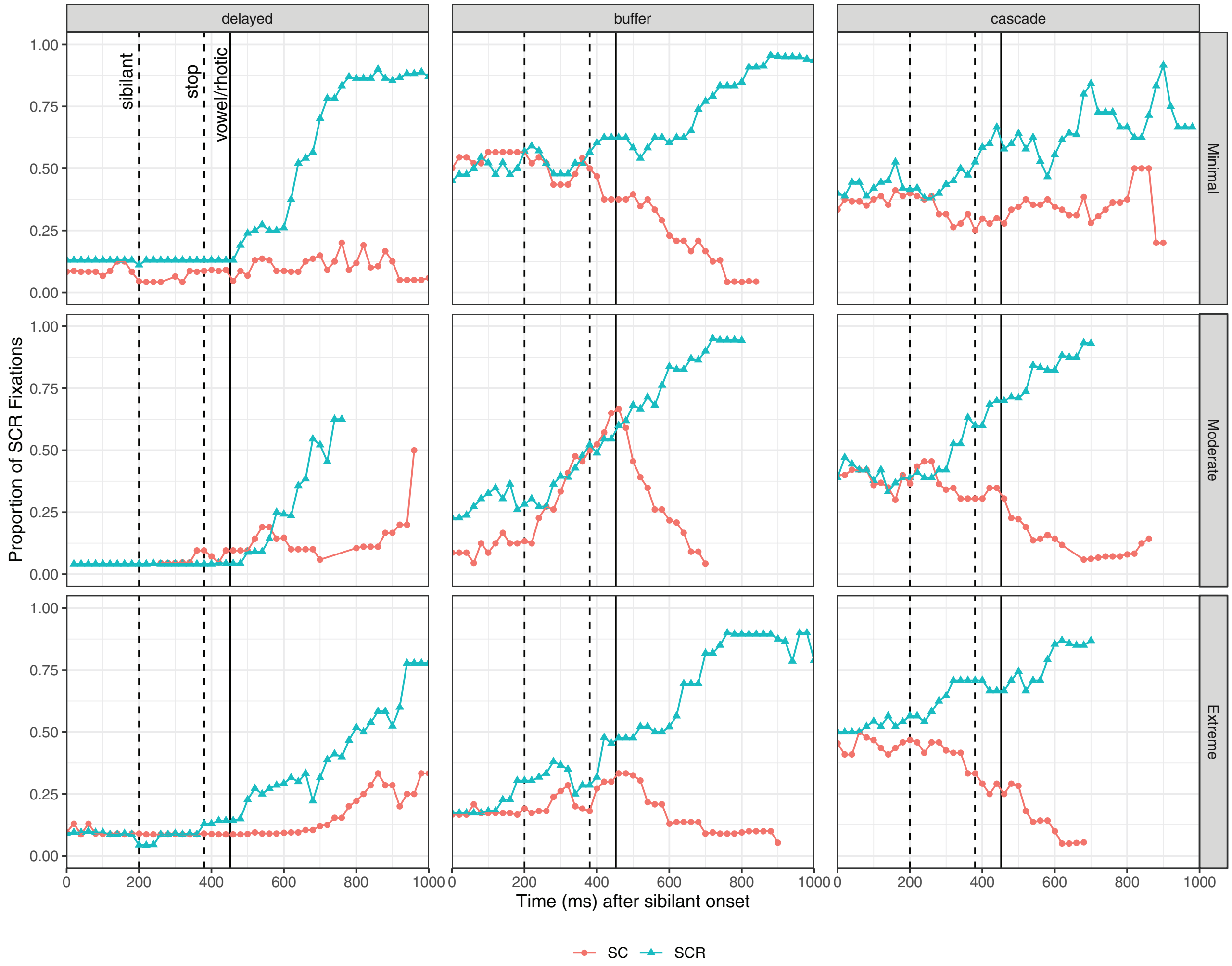

FIGURE 5 | Individual fixation proportion for /sCr/ clusters (y-axis) by time following the sibilant onset (x-axis, binned into 20 ms windows), cluster type (color: /sC/ = red circles, /sCr/ = teal triangles), retraction condition (rows), and pattern exhibited (columns). The vertical lines represent a 200 ms delay from the onset of the sibilant, stop, and vowel/rhotic. A fixation executed during the sibilant interval would be observed between the black dashed vertical lines.

soon as they are available to anticipate the upcoming /r/. This is not to say that all individual variation falls categorically into one of these three patterns, as intermediate strategies were observed by some participants. Rather, these nine individuals demonstrate that these three very different patterns in cue processing are utilized by participants in all three retraction conditions, suggesting that even with an abundance of cues of retraction, some individuals may still wait until the stimuli are disambiguated while other individuals will begin to inform their lexical identification with the smallest of coarticulatory cues.

4. DISCUSSION

The present study examined eye gaze movements to ask if listeners can use spectral information from sibilants immediately in speech processing. This study focused on two different phonological environments where spectral cues in the sibilants were doing different work: prevocalic environments, where spectral cues serve as the primary means of creating a phonemic contrast between /s/ and /ʃ/, and pre-consonantal environments, where spectral coarticulatory cues can foreshadow upcoming sounds without crossing any potential category boundaries. The results demonstrate that listeners can use spectral cues in both environments to immediately increase their consideration of the correct candidate, but more often than not listeners wait until all relevant cues have been heard.

Prevocalic /s/ and /ʃ/ are highly variable and context-dependent, meaning that no cut-and-dry category boundary can be used indiscriminately. The contrast between /s/ and /ʃ/ is made on a variety of different spectral cues and no one individual cue has been found to categorize sibilants between speakers (Jongman et al., 2000). Moreover, spectral cues on sibilants not only vary significantly in different phonological contexts, but also from speaker to speaker (Stuart-Smith, 2007). With Hypothesis

1, I asked if listeners can make immediate use of spectral information in such variable sounds in order to distinguish /s/ from /ʃ/. The results support this hypothesis and demonstrate that listeners can use the spectral cues of sibilants as they unfold in order to disambiguate phonemes, demonstrating that spectral information can be useful in even the most variable sounds.

While these findings add sibilants to a long list of sounds that listeners can begin to disambiguate before all relevant cues are available, they stand in contrast to previous work asking the same question. Galle et al. (2019) examined integration strategies for prevocalic sibilants and found that listeners appear to exhibit a buffer strategy of cue integration, waiting until the onset of the vowel before planning any gaze movements. Galle et al. (2019) explored a variety of different explanations for why sibilants appear to behave differently from other sounds, from acoustic explanations regarding the higher frequency bands occupied by fricatives to their sheer variability and unreliability. It is not immediately clear how to reconcile the present findings of a cascade strategy with the buffer strategy they observed. One possibility stems from differences in instructions: Participants in the present experiment were instructed to select the correct image as "quickly and accurately as possible," while Galle et al. (2019) "encouraged [participants] to take their time and perform accurately" (p. 12). It's possible that emphasizing speed may encourage participants to immediately integrate cues that would otherwise be stored in a buffer until additional cues become available. A second possibility comes from the experiment design: This study presents listeners with two potential candidates while Galle et al. (2019) provided four potential candidates. It's possible that when listeners know the nature of the phonological contrast between the candidates, they are more likely or more able to immediately integrate the spectral cues of that contrast, but as more candidates and contrasts are included, listeners may be more likely to hold spectral information in a buffer. Finally, and perhaps most likely, the difference may stem from differences in analysis: The present study asks whether the proportion of correct fixations improves over the course of the sibilant, while Galle et al. (2019) ask at what point the effect of the onset sibilant crosses a threshold in biasing /s/ consideration. So while Galle et al. (2019) find that listeners are relatively slower in categorizing a sibilant compared to a stop consonant, the present study finds that consideration of the correct candidate significantly improves during the sibilant itself.

With it established that listeners can immediately use the spectral cues of sibilants to discriminate phonological contrasts, the pre-consonantal analysis asks if they can use the same processing strategies for context-dependent variation in order to tease apart two lexical items that may initially be phonologically identical but phonetically distinct. With Hypothesis 2, I asked if listeners can use the coarticulatory cues of /s/-retraction as soon as they are available, such that a listener that hears a retracted /s/ may consider *string* to be a more viable candidate than *sting* even before the /r/ has been heard. The results of this study support this hypothesis, as individuals were shown to increase their consideration of the correct candidate over the course of the sibilant. Furthermore, the stronger the cues of retraction available, the greater the likelihood of considering the correct candidate. Thus, listeners not only are able immediately use the spectral cues of sibilants in order to make phonological contrasts, but also to make context-dependent predictions.

Building off Hypothesis 2, I asked in Hypothesis 3 if a retracted /s/ is a better indicator of rhotic presence than a non-retracted /s/ is of its absence. This was motivated in part by Beddor et al. (2013), who found that a nasalized vowel is a better indicator of an upcoming nasal stop than an oral vowel is for an upcoming oral stop. The results of the present study are inconclusive with respect to this hypothesis. That is, I show that participants are overall more accurate in their perception of /sC/ clusters than /sCr/ clusters, but participants are more likely to correctly look to an /sCr/ image when it is manipulated to have extreme coarticulatory cues. These findings demonstrate that listeners closely attend to different cues, but not all cues are equally helpful in every environment.

Finally, with Hypothesis 4, I again posed a follow-up to Hypothesis 2 to ask if the cues of retraction are more useful in the /str/ clusters where they are most expected than in /skr/ and /spr/ where they're less expected. While /s/-retraction has received increasing sociolinguistic and phonetic attention in recent years, little work has focused on the perception of the phenomenon in situ to ask if listeners attend to those cues. With Hypothesis 4, I ask if listeners have detailed phonological knowledge about the distribution of /s/-retraction and whether they use that knowledge in their consideration of lexical candidates in real time. The results of the present study appear to support this hypothesis as listeners exhibit increased accuracy in the consideration of /str/ clusters over the course of the sibilant compared to /spr/ and /skr/ clusters. However, it is worth noting that there is a potential confound here: not all /sCr/ clusters were manipulated to contain the same degree of retraction in the same conditions. Rather, each place series was manipulated independently relative to the model talker's baseline. Thus, alveolar /str/ clusters contain a greater proportion of /ʃ/ spectral energy than /spr/ and /skr/ clusters in each retraction condition. While this methodology maintains the natural inequalities in retraction that would be observed outside of the lab, it potentially obfuscates our understanding of the results. Is it the case that listeners show greater evidence for immediate integration of coarticulatory cues in /str/ clusters because they expect retraction in those clusters or because, like outside the lab, that is precisely where they are presented with the strongest cues of retraction?

The results of this study demonstrate that listeners can immediately integrate the spectral cues of sibilants in a laboratory setting when they know the nature of the contrast: In a *sip-ship* trial, listeners are expecting a phonological contrast between /s/ and /ʃ/ and, in a *sting-string* trial, they are anticipating or identifying whether the stimulus ultimately contains an /r/. However, it remains to be seen whether this effect can be observed outside of the lab or whether it persists in a more naturalistic task where multiple contrasts may be under consideration simultaneously. For example, if four potential candidates were provided in a trial, e.g., *sing-sting-string-shingle*, a listener is not only making a phonemic contrast between /s/ and /ʃ/ but also anticipating and identifying potential upcoming consonants. In such a scenario, a listener simply may be more likely to use a buffer strategy. However, there may also be a false impression of buffering, if, for example, consideration of *sing, sting*, and *string* may improve but consideration of *shingle* decreases. In this hypothetical trial, looks to the correct candidate may not diverge from other potential candidates, suggesting a buffer strategy, despite the fact that the listener is actively removing other potential candidates from consideration, indicating a cascade strategy. Furthermore, while the present study focused only on the time window during which the sibilant was heard, increasing the number of potential candidates and contrasts also changes the point at which those sounds are disambiguated: Prevocalic stimuli, like *sing* and *shingle*, are disambiguated at the end of the sibilant, but pre-consonantal stimuli, like *sting* and *string* remain temporarily ambiguous. One tool we could use to tease apart these temporal differences would be to consider the integration strategies for nonce words, like *stimble-shtimble-strimble-shtrimble*. While lacking in the temporal resolution that eye tracking allows, Phillips and Resnick (2019) examined categorization of such nonce words, demonstrating that listeners on the whole are reluctant to categorize pre-consonantal onsets as /ʃ/. As listeners uphold the phonotactic restrictions of English even in the perception of nonce words, it is unlikely that nonce words would provide novel or informative evidence for cue integration strategies of pre-consonantal sibilants. Moreover, it is this phonotactic restriction on pre-consonantal sibilants that creates the space for coarticulation to vary so dramatically, giving rise to sound change emergence without endangering a phonemic contrast.

More than asking whether listeners can immediately integrate the coarticulatory cues on sibilants to aid in speech processing, this study asks whether listeners can use the variable cues of a sound change in progress. If a change is underway, it may be the case that listeners are highly variable not just in whether they attend to the cues of retraction, but also in what their acoustic expectations for /str/ clusters are or even in what their phonological representations are, i.e., /str/ vs. /ʃtr/. The present study assumes that listeners retain an underlying /s/, in part due to the phonotactic restrictions that allow even the most extreme [ʃ] to be categorized as /s/ pre-consonantally and in part due to the orthographic biases that may favor a retained /s/. Regardless of its underlying representation, /s/-retraction can help distinguish /str/ clusters from not only /st/ clusters, as examined through the present study, such that *string* and *sting* are readily disambiguated, but also /str/ clusters from /s/ onsets, such that *string* and *sing* are also disambiguated before the end of the sibilant. In its current state, where /s/ is generally intermediate between a canonical /s/ and /ʃ/, /s/-retraction is unlikely to create temporary ambiguity between /str/ clusters and /ʃ/ onsets, such that *street* and *sheet* would be initially confusable. However, it is possible, should phonological reanalysis occur or should /s/ be allophonically produced as [ʃ] in /str/ clusters, that /s/-retraction introduces a new temporary ambiguity between /str/ (or /ʃtr/) clusters and /ʃ/ onsets. This is not tested in the present experiment and the current state of /s/-retraction outside of the laboratory does not predict such a categorical [ʃ] realization, yet it remains a possibility that increased coarticulatory cues do not always disambiguate all phonological environments.

The examination of the individual listeners' results suggests that a range of different patterns were observed in each experimental condition, which demonstrate that individuals can use the cues of /s/-retraction even when they are weak. However, they need not always, as many participants show no such evidence of immediate integration. Given the nature of /s/-retraction as a change in progress, it's not clear in the present design whether listeners' unequal experiences with the change in progress can influence the robust individual variability observed. There was no effect of listener age as all participants were college-aged. Additionally, there was no effect of geographic region, which may initially be unexpected. However, /s/-retraction is a sound change noted for not being associated with any single region or demographic and has instead been referred to as a "general American innovation" (Shapiro, 1995). It is possible that regardless of how geographic region was treated, including using a rural/urban divide, geographic generalizations about the state of /s/-retraction could not capture the distribution of the change in progress. Additionally, it's possible that the geographic variation has been neutralized or diminished since all participants were members of the same community in Chicago at the time of the study and may have had similar exposure to the sound change following their formative years apart.

Furthermore, two other factors that may explain the individual variation were not included in the present design: listeners' categorical judgments and production. It is possible that if we had a means of discerning listeners' underlying representations or if we had examined at what point listeners will categorize an /str/ cluster as an /ʃtr/ cluster, that these would help predict listener variability. For instance, a listener who has an underlying /ʃtr/ cluster may immediately attend to the spectral cues as there's a phonemic contrast in play, rather than a question of coarticulation foreshadowing upcoming sounds. Speakers' own production, that is whether they produce significant retraction in /str/ clusters, may also predict their reliance on the spectral cues of retraction. We might predict that a speaker who produces more retraction may be more likely to immediately integrate the relevant cues. Alternatively, it is possible that a speaker who produces more retraction will attend only to the cues of extreme retraction (to the exclusion of moderate and minimal retraction) while a speaker who produces less retraction will attend only to the cues of minimal retraction (to the exclusion of moderate and maximal retraction). This would mirror findings from an imitation task by which only extreme retractors exhibited convergence in extreme retraction conditions, even if that meant reducing their relative degree of /s/-retraction, and only minimal retractors exhibited convergence in minimal retraction conditions, even if that meant increasing their relative degree of /s/-retraction (Phillips, 2020). At stake here is whether experience with the sound change makes a listener more sensitive to the cues across the board or whether a listener is more sensitive to cues that better align with their own speech. I leave these questions to future work and recognize that the individual variability observed here is robust even if it is not predictable.

5. CONCLUSION

The boundary between /s/ and /ʃ/ is anything but a clear and reliable line, clouded by mountains of ambiguity and variability. Listeners attend to the vast amount of information at their disposal to constantly shift the boundaries, whether that be because of phonological contexts, some facet of the speaker's identity, or simply as a result of the sounds they were recently exposed to Kraljic et al. (2008). This means that there is a lot of potentially conflicting information that listeners have to deal with in a short span of time. It was perhaps unsurprising that Galle et al. (2019) suggested that sibilants are possibly too variable and unreliable to be immediately integrated. Rather, listeners were thought to sit through a few milliseconds of ambiguity and wait until they have all the relevant information they need to start processing. Yet the present study finds the opposite: Despite the notable variability, or perhaps because of it, listeners are able to immediately use the cues available to them to begin lexical identification. It's worth noting that just because they can, does not mean that they must, as fixation accuracy does not cross 50% until after the vowel onset.

Moreover, the present study finds that listeners not only immediately use cues in contrasting different sibilants like /s/ and /ʃ/, but also in pre-consonantal environments where no phonological contrast between /s/ and /ʃ/ exists. In these environments, unconstrained by phonological contrasts, /s/ shows extreme coarticulatory variability, approaching the /s/-/ʃ/ boundary. This study demonstrates that listeners are astutely aware of this coarticulatory variability and use it in real-time to disambiguate words like *string* and *sting* that should be ambiguous at that point in time. Beyond demonstrating that listeners have detailed knowledge of the sound change and use that knowledge in perception, these results make interesting implications for the future of /s/-retraction as a sound change. Firstly, the results of this study demonstrate that listeners are attending to coarticulatory cues in /spr/ and /skr/ clusters despite the fact that retraction is currently much more advanced in /str/ clusters. This suggests that these environments may be the next loci for the sound change, following many other Germanic languages (Bukmaier et al., 2014). Secondly, it demonstrates that the spectral cues on /s/ serve an important role in contrasting /sC/ and /sCr/ sequences. While the results still clearly suggest that the presence or absence of /r/ is the primary disambiguating force in words like *string* and *sting*, the fact remains that listeners are carefully attending to the sibilant, in part because it temporally precedes the /r/. If the sound change continues to advance and if listeners begin to reanalyze the onset as /ʃ/, it is possible that listeners will begin to shift cue weight onto the onset sibilant until it is the primary cue in contrasting these clusters. In this scenario, the rhotic itself would eventually become redundant, which may lead to it being reduced or deleted entirely. In a possible distant future, the contrast would not be between *string* and *sting*, but *shting* and *sting*.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Chicago Social & Behavioral Sciences IRB. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JP conceived, designed, and conducted the experiment, analyzed the data, and wrote the article.

FUNDING

This research was supported by the National Science Foundation Doctoral Dissertation Research Improvement Grant No. BCS-1749342 and the University of Chicago-Mellon Foundation Dissertation Completion Fellowship.

ACKNOWLEDGMENTS

Thank you to Alan C. L. Yu for his mentorship and guidance and to Diane Brentari, Susan Lin, and Jane Stuart-Smith for their valuable feedback. Thank you to Ming Xiang for the generous use of the Language Processing Laboratory facilities and equipment. Thank you to Alex Kramer for technological assistance and to Sasha Elenko and Max Fennell-Chametzky for their assistance in data collection. This article is based on Chapter 3 of the author's dissertation (Phillips, 2020).

REFERENCES

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *J. Mem. Lang.* 38, 419–439. doi: 10.1006/jmla.1997.2558

Bailey, G., Nichols, S., Turton, D., and Baranowski, M. (2022). Affrication as the cause of /s/-retraction: evidence from Manchester English. *Glossa* 7. doi: 10.16995/glossa.8026

Baker, A., Archangeli, D., and Mielke, J. (2011). Variability in American English s-retraction suggests a solution to the actuation problem. *Lang. Variat. Change* 23, 347–374. doi: 10.1017/S0954394511000135

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., and Brasher, A. (2013). The time course of perception of coarticulation. *J. Acoust. Soc. Am.* 133, 2350–2366. doi: 10.1121/1.4794366

Bukmaier, V., Harrington, J., and Kleber, F. (2014). An analysis of post-vocalic /s-ʃ/ neutralization in Augsburg German: evidence for a gradient sound change. *Front. Psychol.* 5:828. doi: 10.3389/fpsyg.2014.00828

Connine, C. M., and Darnieder, L. M. (2009). Perceptual learning of co-articulation in speech. *J. Mem. Lang.* 61, 368–378. doi: 10.1016/j.jml.2009.07.003

Darwin, C. (2005). *Digital Mixing Script*. Brighton: University of Sussex.

Diehl, R. L., Lotto, A. J., and Holt, L. L. (2004). Speech perception. *Annu. Rev. Psychol*. 55, 149–179. doi: 10.1146/annurev.psych.55.090902.142028

Durian, D. (2007). "Getting stronger every day?: more on urbanization and the socio-geographic diffusion of (STR)," in *University of Pennsylvania Working Papers in Linguistics, Vol. 13*, eds S. Brody, M. Friesner, L. Mackenzie, and J. Tauberer (Philadelphia, PA), 65–79.

Elman, J., and McClelland, J. (1986). "Exploiting the lawful variability in the speech wave," in *Invariance and Variability of Speech Processes*, eds J. S. Perkell and D. Klatt (Hillsdale, NJ: Lawrence Erlbaum Associates).

Fowler, C. (1996). Listeners do hear sounds, not tongues. *J. Acoust. Soc. Am*. 99, 1730–1741. doi: 10.1121/1.415237

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *J. Phonet*. 14, 3–28. doi: 10.1016/S0095-4470(19)30607-2

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Percept. Psychophys*. 68, 161–177. doi: 10.3758/BF03193666

Galle, M. E., Klein-Packard, J., Schreiber, K., and McMurray, B. (2019). What are you waiting for? Real-time integration of cues for fricatives suggests encapsulated auditory memory. *Cogn. Sci*. 43:e12700. doi: 10.1111/cogs.12700

Glain, O. (2013). *Les cas de palatalisation contemporaine (CPC) dans le monde anglophone* (Ph.D. thesis). Université Jean Moulin, Lyon, France.

Gylfadottir, D. (2015). "Shtreets of Philadelphia: an acoustic study of /str/-retraction in a naturalistic speech corpus," in *University of Pennsylvania Working Papers in Linguistics, Vol. 21* (Philadelphia, PA), 2–11.

Holt, L. L., and Kluender, K. R. (2000). General auditory processes contribute to perceptual accommodation of coarticulation. *Phonetica* 57, 170–180. doi: 10.1159/000028470

Jongman, A., Wayland, R., and Wong, S. (2000). Acoustic characteristics of English fricatives. *J. Acoust. Soc. Am*. 108, 1252–1263. doi: 10.1121/1.1288413

Kraljic, T., Brennan, S. E., and Samuel, A. G. (2008). Accommodating variation: dialects, idiolects, and speech processing. *Cognition* 107, 54–81. doi: 10.1016/j.cognition.2007.07.013

Lawrence, W. P. (2000). /str/ → /textipaStr/: assimilaton at a distance? *Am. Speech* 75, 82–87. doi: 10.1215/00031283-75-1-82

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi: 10.1016/0010-0277(85)90021-6

Lindblom, B. (1990). "Explaining phonetic variation: a sketch of the H&H theory," in *Speech Production and Speech Modelling*, eds W. J. Hardcastle and A. Marchal (Dordrecht: Kluwer Academic Publishers), 403–439. doi: 10.1007/978-94-009-2037-8_16

Lotto, A. J., and Kluender, K. R. (1998). General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. *Percept. Psychophys*. 60, 602–619. doi: 10.3758/BF03206049

Martin, J. G., and Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects. *J. Acoust. Soc. Am*. 69, 559–567. doi: 10.1121/1.385484

Matthies, M. L., Perrier, P., Perkell, J. S., and Zandipour, M. (2001). Variation in anticipatory coarticulation with changes in clarity and rate. *J. Speech Lang. Hear. Res*. 44, 340–353. doi: 10.1044/1092-4388(2001/028)

McMurray, B., Clayards, M., Tanenhaus, M. K., and Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychon. Bull. Rev*. 15, 1064–1071. doi: 10.3758/PBR.15.6.1064

Ostreicher, H., and Sharf, D. (1976). Effects of coarticulation on the identification of deleted consonant and vowel sounds. *J. Phonet*. 4, 285–301. doi: 10.1016/S0095-4470(19)31256-2

Phillips, J. B. (2020). *Sibilant categorization, convergence, and change: the case of /s/-retraction in American English* (Ph.D. thesis). University of Chicago, Chicago, IL, United States.

Phillips, J. B., and Resnick, P. (2019). Masculine toughness and the categorical perception of onset sibilant clusters. *J. Acoust. Soc. Am*. 145:EL574. doi: 10.1121/1.5113566

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Scarborough, R. (2004). *Coarticulation and the structure of the lexicon* (Ph.D. thesis). University of California, Los Angeles, Los Angeles, CA, United States.

Shapiro, M. (1995). A case of distant assimilation: /str/ → /ʃtr/. *Am. Speech* 70, 101–107. doi: 10.2307/455876

Smith, B. J., Mielke, J., Magloughlin, L., and Wilbanks, E. (2019). Sound change and coarticulatory variability involving English /ɹ/. *Glossa* 4:63. doi: 10.5334/gjgl.650

Stevens, K., and Keyser, S. J. (2010). Quantal theory, enhancement and overlap. *J. Phonet*. 38, 10–19. doi: 10.1016/j.wocn.2008.10.004

Stevens, M., and Harrington, J. (2016). The phonetic origins of /s/-retraction: acoustic and perceptual evidence from Australian English. *J. Phonet*. 58, 118–134. doi: 10.1016/j.wocn.2016.08.003

Stilp, C. (2019). Acoustic context effects in speech perception. *WIREs Cogn. Sci*. 11:e1517. doi: 10.1002/wcs.1517

Stuart-Smith, J. (2007). "Empirical evidence for gendered speech production: /s/ in Glaswegian," in *Laboratory Phonology, Vol. 9*, eds J. Cole and J. I. Hualde (New York, NY: Mouton de Gruyter), 65–86.

Viswanathan, N., Magnuson, J. S., and Fowler, C. A. (2010). Compensation for coarticulation: disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *J. Exp. Psychol*. 36, 1005–1015. doi: 10.1037/a0018391

Whalen, D. H. (1991). Subcategorical phonetic mismatches and lexical access. *Percept. Psychophys*. 50, 351–360. doi: 10.3758/BF03212227

Wilbanks, E. (2017). "Social and structural constraints on a phonetically motivated change in progress: (STR) retraction," in *Working Papers in Linguistics, Vol. 23* (Philadelphia, PA). doi: 10.5070/P7121040720

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 17 June 2022
doi: 10.3389/fcomm.2022.896013

Production and Perception of Mandarin Laryngeal Contrast: The Role of Post-plosive F0

Yuting Guo[†] *and Harim Kwon*[*†]

Linguistics Program, Department of English, George Mason University, Fairfax, VA, United States

This study examines the relation between plosive aspiration and post-plosive f0 (fundamental frequency) in the production and perception of the laryngeal contrast in Mandarin. Production data from 25 Mandarin speakers showed that, in word onsets, VOTs (voice onset time) of aspirated and unaspirated plosives were different, as expected. At the same time, the speakers produced different post-plosive f0 between aspirated and unaspirated plosives, but the difference varied according to the lexical tones – post-aspirated f0 was higher than post-unaspirated f0 in high-initial tones (i.e., lexical tones with high onset f0), but the pattern was the opposite and less robust in low-initial tones. In the perception of the same participants, VOT was the primary cue to aspiration but, when VOT was ambiguous, high post-plosive f0 yielded more aspirated responses in general. We claim that the asymmetry in f0 perturbation between high-initial and low-initial tones in production arises from different laryngeal maneuvers for different tonal targets. In low-initial tones, in which the vocal folds are slack and the glottal opening is wider, aspirated plosives have a lower subglottal air pressure than unaspirated plosives at the voicing onset, resulting in lower post-aspirated f0 than post-unaspirated f0. But in high-initial tones, the vocal folds are tense, which requires a higher trans-glottal pressure threshold to initiate phonation at the onset of voicing. As a result, the subglottal pressure does not decrease as much. Instead, the faster airflow in aspirated than unaspirated plosives gives rise to the pattern that post-aspirated f0 is higher than post-unaspirated f0. Regardless of this variation in production, our perception data suggest that Mandarin listeners generalize the f0 perturbation patterns from high-initial tones and associate high post-plosive f0 with aspirated plosives even in low-initial tone contexts. We cautiously claim that the observed perceptual pattern is consistent with the robustly represented production pattern, as high-initial tones are more prevalent and salient in the language and exhibit stronger f0 perturbation in the speakers' productions.

Keywords: Mandarin Chinese, laryngeal contrast, aspiration, fundamental frequency (f0), production-perception relation, secondary cue

OPEN ACCESS

Edited by:
*Georgia Zellou,
University of California, Davis,
United States*

Reviewed by:
*Jiayin Gao,
Ludwig Maximilian University of
Munich, Germany
James Kirby,
Ludwig Maximilian University of
Munich, Germany*

***Correspondence:**
*Harim Kwon
hkwon20@gmu.edu*

[†] *These authors share first authorship*

Specialty section:
*This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Communication*

Received: *14 March 2022*
Accepted: *13 May 2022*
Published: *17 June 2022*

Citation:
*Guo Y and Kwon H (2022) Production
and Perception of Mandarin Laryngeal
Contrast: The Role of Post-plosive F0.
Front. Commun. 7:896013.
doi: 10.3389/fcomm.2022.896013*

INTRODUCTION

F0 Perturbation

Laryngeal properties (such as voicing or aspiration) of onset plosives influence the fundamental frequency, or f0, at the onset of the following vowels. This phenomenon, commonly referred to as f0 perturbation, has been widely attested across languages, such as Cantonese (Francis et al., 2006; Luo, 2018), Dutch (Löfqvist et al., 1989), English

(House and Fairbanks, 1953; Lehiste and Peterson, 1961; Hombert et al., 1979; Ohde, 1984; Löfqvist et al., 1989; Hanson, 2009), French (Kirby and Ladd, 2016), German (Kohler, 1982; Hoole and Honda, 2011), Italian (Kirby and Ladd, 2016), Japanese (Gao and Arai, 2019), Khmer (Kirby, 2018), Mandarin (Xu and Xu, 2003; Luo, 2018), Russian (Mohr, 1971), Spanish (Dmitrieva et al., 2015), Thai (Gandour, 1974; Kirby, 2018), Vietnamese (Kirby, 2018), Xhosa (Jessen and Roux, 2002), Yoruba (Hombert et al., 1979), among others. The most commonly reported pattern shows that a (phonologically) voiced plosive has a lower post-plosive f0 than a (phonologically) voiceless one, although there are some notable patterns.

First, f0 perturbation occurs in so-called true voicing languages and in aspirating languages alike. That is, it seems less relevant whether the language contrasts prevoiced vs. voiceless unaspirated categories or unaspirated vs. aspirated categories. For example, both Spanish and English show similar f0 perturbation (Dmitrieva et al., 2015). This might be because English unaspirated plosives are phonologically voiced (Kingston and Diehl, 1994; Hanson, 2009). However, findings on languages with a three-way laryngeal contrast (prevoiced vs. unaspirated vs. aspirated) suggest that the difference between unaspirated and aspirated categories cannot entirely be reduced to phonological voicing. For example, Kirby (2018) examines Khmer, Vietnamese, and Thai, all with the three-way contrast, and finds that aspirated plosives are followed by a higher f0 than the unaspirated ones, at least for some speakers in all three languages. This provides evidence for the bona fide effects of consonantal aspiration (or the lack thereof) on the following f0.

Although the commonly reported pattern of f0 perturbation is voiceless (or aspirated) plosives having higher post-plosive f0 than voiced (or unaspirated) ones, this is not always the case. For example, Xu and Xu (2003) report that, in Mandarin, f0 is lower after aspirated plosives than after unaspirated plosives because aspiration causes the sub-glottal air pressure to decrease sharply, lowering f0 at the release of the plosives. However, Luo (2018) provides contradicting findings such that aspiration in Mandarin raises f0 quite robustly. The cause of this discrepancy is unclear (see more in the section: F0 Perturbation in Mandarin).

Second, f0 perturbation is attested in both tonal languages and non-tonal languages although the effects are less robust in tonal languages. For instance, the f0 differences between English unaspirated and aspirated series can last more than 100 ms after the voicing onset whereas they last 40∼60 ms in a tonal language, Yoruba (Hombert et al., 1979). Other studies on tonal languages (e.g., Chen, 2011, on Shanghainese; Gandour, 1974, on Thai; Francis et al., 2006, on Cantonese; Xu and Xu, 2003, on Mandarin) also suggest that f0 perturbation is limited to the very onset of the vowel and its exact duration is determined by the tonal contexts. Furthermore, Kirby (2018) reports that in Thai and Vietnamese, the perturbation effect is clearly observed in citation forms, but not in connected speech. This indicates that the effects of f0 perturbation may interact not only with tonal contexts but also with sentence-level prosody. See also Hanson (2009), Chen (2011), and Xu and Xu (2003), for similar effects in English, Shanghainese, and Mandarin, respectively.

Third, though the magnitude of the f0 perturbation is quite small (ranging 8–16 Hz in different languages, Table 1 in Coetzee et al., 2018), listeners use the f0 at the vowel onset to determine the preceding consonant's laryngeal category across different languages. English listeners, for instance, use f0 as a cue to consonant's laryngeal category not only when VOT, the phonetic property that is primarily responsible for the laryngeal contrast, is ambiguous (e.g., Whalen et al., 1990), but also when it is not ambiguous (e.g., Whalen et al., 1993). Even in tonal languages, in which f0 is primarily responsible to carry tonal information, and the perturbation, if any, is less consistent and temporally limited, post-plosive f0 influences listeners' perceptual judgments on onset plosive's laryngeal category. For example, Francis et al. (2006) report that falling f0 contours at the onset of a high-level tone signal aspirated plosives to Cantonese listeners and this perceptual pattern does not match the f0 patterns in Cantonese plosive productions. They claim that the use of post-plosive f0 as a consonantal cue, therefore, does not originate from the experience of hearing the covarying VOT and f0. Rather, Cantonese listeners' perception shows the influence of the language-independent, general auditory enhancing effects among different phonetic properties (Kingston and Diehl, 1994; Francis et al., 2006).

Despite the universality of the phenomenon, the source of f0 perturbation is controversial. Some have argued that f0 perturbation is a physiological or physical epiphenomenon of consonantal voicing or aspiration (e.g., Hombert et al., 1979; Löfqvist et al., 1989). Several different hypotheses have been offered on the exact mechanism of f0 perturbation. First, the aerodynamic hypothesis claims that voiced plosives differ from voiceless ones in how air pressure changes during and after their oral closure, leading to differing f0 after the release. In the case of voiced plosives, supraglottal air pressure gradually builds up during the closure because voicing requires a continuous airflow through the glottis. This results in a decrease in the trans-glottal air pressure difference, which in turn leads to a decrease in f0. On the other hand, voiceless plosives have a greater volume of airflow from subglottal to supraglottal cavities upon the release, resulting in faster vocal fold vibration (but see also Xu and Xu, 2003). Another hypothesis claims that f0 perturbation arises from the states of vocal folds during plosive voicing (e.g., Halle and Stevens, 1971; Löfqvist et al., 1989). During the plosive closure, the vocal folds remain slack for voiced plosives whereas they are stiff for voiceless plosives to halt the vibration. The tension of the vocal folds influences the f0 of the flanking vowels, such that slack vocal folds lower, and stiff vocal folds raise, the rate of their vibration. Still another hypothesis claims that f0 perturbation is due to the larynx height difference between the voiced and voiceless plosives (e.g., Honda, 2004). To allow for vocal fold vibration during the closure, the larynx is lower for voiced plosives than for voiceless ones. As the larynx height is usually positively correlated with f0, voiced plosives have lower post-plosive f0 than voiceless ones.

Despite the differences in their exact mechanisms, these hypotheses commonly suggest that the effects of plosive voicing (or voicelessness) on the following f0 are automatic and determined by the biomechanics of the larynx. In contrast, it has also been claimed that speakers actively induce the f0 differences to enhance the phonological contrast (e.g., Kingston and Diehl, 1994; Kingston, 2007). Under this phonological hypothesis, post-plosive f0 is not a mere by-product of sustaining voicing during the plosive closure or aspiration after the plosive release. Rather, speakers enhance the phonological laryngeal contrast by enhancing covarying phonetic properties. This results in the plosives of different laryngeal categories having distinct post-plosive f0, prolonged beyond the very beginning of the vowel. Therefore, this hypothesis can readily explain why the languages that contrast prevoiced and voiceless plosives (e.g., Spanish) and those contrasting aspirated and unaspirated plosives (e.g., English) show similar f0 perturbation patterns. In addition, in tonal languages, speakers would not enhance consonantal contrast using post-plosive f0 because f0 plays a central role in conveying lexical (or grammatical) information (Francis et al., 2006).

As pointed out in previous research (e.g., Chen, 2011; Hoole and Honda, 2011; Dmitrieva et al., 2015), these two views, automatic vs. phonological, are not incompatible with each other. In fact, it is possible that the biomechanical factors determine the connection between the voicing and f0, which serves as the resource for speakers to use as an enhancement strategy for plosive laryngeal contrast. Building on this previous conversation on f0 perturbation, this study asks how speakers of a tonal language use post-plosive f0 as a consonantal cue. Focusing on the relation between plosive aspiration and post-plosive f0, we investigate the production and perception of Mandarin word-initial plosives in different tonal contexts. The rest of the introduction will briefly review the relevant background on Mandarin and present the main questions for the two experiments.

F0 in Mandarin

Lexical Tones

Mandarin has four lexical tones, typically described as high-level (Tone 1), rising (Tone 2), low-dipping (Tone 3), and falling (Tone 4) (e.g., Xu, 1997; Duanmu, 2007). In this paper, tones are abbreviated as T1, T2, T3, and T4, and syllables produced with a specific tone are noted with a number added to the syllable. For example, /t^{h}a1/ refers to the syllable /t^{h}a/ with T1.

Xu (1997) describes the f0 contours of the four lexical tones as the following. T1 begins with a high f0 and maintains the same level through the entire vowel; T2 starts with a low f0, and then falls slightly until 20% into the vowel before rising throughout the rest of the vowel; T3, in citation form, begins with a low f0, falls to the lowest f0 at the midpoint of the vowel, and then rises sharply to the end of the syllable although the final rise is usually absent in non-prepausal positions; and T4 starts with a high f0, and then drops sharply from the 20% of the vowel until the end of the syllable. As f0 perturbation due to onset consonant is expected to be most distinct in the beginning of the vowel (adjacent to the onset consonant), two important aspects of these tones should be noted. First, T1 and T4 begin with a high f0 while T2 and T3 with a low f0. Second, T1 has the most static f0 contour and, in connected speech, T2 and T4 have more dynamic f0 contours than T3 during the first half of the vowel.

As for the physiological properties of Mandarin tones, studies have shown that larynx height is in general positively correlated with f0 (e.g., Hallé, 1994; Moisik et al., 2014). Specifically, the larynx is higher at the syllable onsets in T1 and T4 than in T2 and T3. However, Moisik et al. (2014) claim that the role of larynx height may be only facilitatory and, thus, the relation between larynx height and tones is not necessarily straightforward. This suggests that speakers may utilize different laryngeal settings (including larynx height, and vocal fold tension, among other things) to produce different tonal targets in Mandarin.

F0 Perturbation in Mandarin

Mandarin plosives are typically classified as voiceless unaspirated and voiceless aspirated, with aspiration as the primary distinction (Mandarin plosives are henceforth referred to as **unaspirated** and **aspirated** plosives). The language does not have voiced obstruents and, thus, the voiced consonants that can occur as a word onset are sonorants, such as /m n l w j/.

Inconsistent results have been reported on f0 perturbation in Mandarin (e.g., Xu and Xu, 2003; Luo, 2018; Chi et al., 2019). Xu and Xu (2003) suggest that aspiration is associated with low f0 although the specific pattern can be influenced by the tonal contexts. The lowering of aspiration is more robust in tones beginning with a low f0 (T2/T3, henceforth low-initial tones) than in those with a high f0 (T1/T4, high-initial tones). They attribute this pattern to the aerodynamics of the aspiration, which is characterized by a rapid outward flow of a large volume of air at the release of a plosive. This airflow, occurring between the release of oral closure and the glottal pulsing, lowers the subglottal air pressure for the aspirated plosives more than for the unaspirated ones, decreasing post-aspirated f0. The effects of aerodynamic force become even stronger when the intended pitch is low which is realized with slack vocal folds. Therefore, at the onset of low-initial tones, the vocal folds are slack and the f0 difference between aspirated and unaspirated series is enlarged.

By contrast, Luo (2018) reports that aspiration raises the f0, which extends longer in high-initial tones than in low-initial tones. In T2, they did not find a clear pattern of f0 perturbation. As for the source of this pattern (higher f0 after aspirated than unaspirated plosives), Luo mentions that aspiration is typically associated with high transglottal air pressure, elevated larynx, and stiff vocal folds, all of which raise the f0. On the other hand, she attributes the longer f0 perturbation in the high-initial tones than in low-initial tones to speakers' control (e.g., Kingston and Diehl, 1994). According to Luo (2018), in Mandarin, high-initial tones are more salient than low-initial ones both phonologically and perceptually. Phonologically, high-initial tones are more likely to be preserved in phonological processes, and perceptually, listeners are more accurate in perceiving high-initial tones. Assuming that tonal language speakers actively suppress the biomechanically-motivated automatic f0 perturbation to enhance the tonal contrast (e.g., Hombert et al., 1979; Francis et al., 2006), there is less need for this suppression when the tones are salient. Therefore, in Mandarin, high-initial tones allow for more f0 variability than low-initial tones.

The cause of the divergent findings in Xu and Xu (2003) and Luo (2018) is unclear. However, it is worth mentioning that the participants in both studies are all female speakers, who produce the target syllables embedded in different carrier phrases. In Luo's (2018) carrier phrase, the target syllables are always preceded by T1 whereas Xu and Xu (2003) use two different types of carrier phrases differing in the preceding syllable tones, T1 and T3. The two studies also differ in how they use f0 measurements in their analyses. While Xu and Xu's (2003) analyses are based on the raw f0 measured in Hz, Luo (2018) uses z-scored f0 normalized by speaker. The different patterns are possibly due to a great inter-speaker variation, as well.

Chi et al. (2019) compare two male speakers' glottal opening and oral airflow in aspirated and unaspirated plosives in T1. Their findings corroborate the possibility of the inter-speaker variation. One of the two tested speakers does not show the f0-aspiration covariation but shows faster oral airflow in aspirated than unaspirated plosives, especially when preceding a low vowel /a/ (Figure 5 in Chi et al., 2019). This speaker shows a negative relationship between the post-plosive f0 and oral airflow rate, suggesting that the post-plosive f0 decreases as the oral airflow rate increases, presumably for the consonant aspiration and a low vowel. This is consistent with the aerodynamic interpretation in Xu and Xu (2003). However, the other speaker does not show this airflow rate difference between aspirated and unaspirated plosives. And only this speaker tends to produce higher f0 for aspirated than unaspirated plosives, consistent with Luo's (2018) findings, although the f0 difference is not large enough to distinguish the aspiration contrast.

Despite the diverging patterns and potential individual variation, the previous findings commonly suggest that the f0 perturbation in Mandarin is fairly limited to the vowel onset. This is consistent with previous findings in other tonal languages (e.g., Hombert et al., 1979; Francis et al., 2006; Kirby, 2018).

Current Study

This study examines the role of post-plosive f0 as a secondary cue for Mandarin plosive laryngeal contrast in two experiments. We ask how the lexical tone mediates the f0 patterns in production, as well as the listeners' perceptual responses. The f0 at the vowel onset is expected to be influenced, interactively, by the lexical tone and the perturbation effects due to the onset consonants.

Experiment 1 examines the plosive production of Mandarin speakers to investigate the f0 patterns at vowel onset, influenced by the laryngeal category of the onset consonant, in CV syllables. The central questions for Experiment 1 are (1) how the aspiration (or the lack thereof) of the onset consonant changes the f0 at the onset of voicing following the onset consonant, and (2) how the tonal contexts influence the relation between consonant aspiration and f0 at voicing onset, if any. As mentioned above, the existing findings on the f0 perturbation in Mandarin are divergent and inconclusive (e.g., Xu and Xu, 2003; Luo, 2018; Chi et al., 2019). We aim to provide an additional set of empirical data, including both female and male speakers, on the f0 perturbation in Mandarin.

Experiment 2 examines Mandarin plosive perception. In Experiment 2, we specifically ask (1) whether the f0 differences between different laryngeal categories, if any, are used by Mandarin listeners as a cue to the onset aspiration, and (2) how the tonal contexts influence the listeners' use of f0 as a consonantal cue, if at all. It is still unknown whether Mandarin listeners use f0 as a secondary cue to the laryngeal contrast, to the best of our knowledge. Since f0 is the primary cue for lexical tones in the language, Mandarin listeners might not rely on the post-plosive f0 to determine the laryngeal category of the onset plosives. If Mandarin listeners do use the post-plosive f0 as a cue for the onset plosive, such an outcome may have different interpretations depending on the findings in Experiment 1. If the production patterns provide evidence for the perceptual patterns (i.e., if the listeners' behaviors reflect the robust patterns present in the speakers' production), the listeners' behaviors can be attributed to their native language experience. On the other hand, if the listeners associate post-plosive f0 with consonant aspiration in the absence of systematic f0 perturbation patterns in Mandarin productions, their perceptual behaviors could be attributed to the general auditory enhancing effects (Kingston and Diehl, 1994; Francis et al., 2006).

EXPERIMENT 1: PRODUCTION

Experiment 1 examines Mandarin speakers' plosive productions in CV syllables, asking how f0 at the vowel onset changes as a function of the laryngeal category of the onset consonant, in different tonal contexts.

Methods

Participants

Twenty-five native speakers of Mandarin Chinese (15 female and 10 male, mean age = 26, range = 19∼46) were recruited from the George Mason University community, in Virginia, USA. They were self-identified as native speakers of Mandarin, born and raised in the North China. All participants learned and spoke English as their second language, but they reported to be dominant in Mandarin. The participants moved to the US at the mean age of 22 (range 19∼35) and had lived in the US for 1∼48 months (mean = 13) at the time of testing, except for one participant (F05), who had been in the U.S. for 20 years. After confirming the data from this participant were not distinct from the rest of the group, we decided to include her in the analysis. The individual data are provided in the **Supplementary Materials**. No participants reported any history of speech or hearing disorders. The participants received monetary compensation for their participation.

Stimuli

The stimuli were 24 monosyllabic Mandarin words, with 3 onset consonants (aspirated, unaspirated, sonorant) * 2 vowel contexts (low [a], high [u]) * 4 lexical tones. We were mainly interested in comparing aspirated and unaspirated plosives, and sonorant onsets were also included as fillers. For the onset consonants, we used /t/, /t^h/, and /w/, as they yielded the least number of lexical gaps when combined with the vowels /a/ and /u/. However, /t^{h}a2/ is still lexically missing in Mandarin and, thus, was substituted with /p^{h}a2/, as f0 patterns for /t^{h}a2/ and /p^{h}a2/ are known to be similar (Ohde, 1984; Xu and Xu, 2003). In order to avoid directly comparing syllables with different onsets, we also substituted /ta2/ with /pa2/.

Written Mandarin words corresponding to each of the 24 syllables were selected based on the word frequency data from the Modern Chinese Balanced Corpus (Xiao, 2010, corpus size = 100 million words). Only the words labeled as "most common" were selected. None of the selected words was a bound morpheme in Mandarin. For the complete list of stimuli, see **Appendix A**.

The selected words were embedded in a carrier phrase 请说___一次 (/t͡ɕʰiŋ3 ʂwɔ1 ____ ji2 t͡sʰi4/, 'Please say ____ one time.')[1], and visually presented to the participants. The visual prompts included the entire carrier phrase in Chinese characters, with the stimulus word both in Chinese characters and Pinyin[2].

Procedure

The experiment took place in a sound-attenuated booth at the Phonetics and Phonology Lab at George Mason University. Participants were seated in front of a Macbook computer that presented the stimuli. Their productions were digitally recorded onto a separate Macbook Pro, using a lapel microphone (Røde smartLav+) and an external Focusrite Scarlette Solo 2nd Generation audio-interface, with a sampling rate of 44.1 kHz via the Praat program (Boersma and Weenink, 2020). The microphone was attached to the participants' shirt on the upper chest, ∼6 inches away from the speakers' mouth.

The visual prompts for stimuli were presented to the participants one at a time in the middle of the laptop screen using PsychoPy (Peirce, 2007). In order to elicit a comparatively stable speaking rate across participants, the sentences were presented with a fixed inter-stimulus interval of 3.5 seconds. Participants were instructed to read aloud each sentence on the laptop screen as naturally as possible. All written and oral instructions were provided in Mandarin.

Each stimulus (24 words) was repeated 6 times in randomized orders, resulting in a total of 144 trials per speaker. The 144 trials were presented in two blocks of 3 repetitions, with a self-paced break between the blocks. Beforehand, a short practice block with 2 trials was included to familiarize the participants with the task. The recording session took approximately 10 minutes.

Measurements and Data Preparation

All measurements were taken using Praat (Boersma and Weenink, 2020) by one of the authors (YG). Before taking the measurements, 23 of 3,600 (144 tokens * 25 speakers, 0.6%) tokens were removed due to production errors (e.g., not producing the target word, hesitation, self-correction, unintended noise such as coughing or clearing throat, etc.). For the remaining tokens, three different acoustic landmarks were labeled for each target token with the stop onset: (1) the onset of the stop burst, (2) the onset of the periodicity of the vowel following the stop consonant, and (3) the offset of the vowel second formant. VOT was calculated by subtracting (1) from (2), and the vowel duration by subtracting (2) from (3). For the fillers with the sonorant onset, the segmentation between the approximant onset /w/ and the following vowel was determined by visual inspection of the spectral patterns. The boundary was located at the point where the second formant (F2) moved up from the steady-state (Peterson and Lehiste, 1960), as well as the amplitude increased suddenly. The higher formants were used when F2 was not useful.

The f0 values from 20 equidistant points of the post-onset vowel, and then the first 8 (out of 20) f0 values (from the first 35% of the vowel) were used in the subsequent statistical analyses. As the duration of Mandarin sentence-medial vowels varies according to the lexical tones (e.g., Deng et al., 2006), the absolute duration of the 35% of the vowel used in this time-normalized method differs across the tones (mean duration for T1 75 ms; T2 80 ms; T3 75 ms; and T4 71 ms)[3]. The f0 values were extracted using a Praat script, with a 600 Hz pitch ceiling, a 75 Hz pitch floor, and a 10 ms time step. Any tracking errors were hand-corrected. In this process, an additional 5.3% of the data were removed due to unreliable f0 tracking when the vowel was not modal-voiced. A large portion of these excluded data was due to creaky voice, mostly in T3, but to a smaller extent in the other tones, when the f0 was low (see Kuang, 2017, for the discussion on creaky voice in different Mandarin tones).

Results

All statistical analyses in this study were conducted in R (R Core Team., 2021). To investigate the f0 perturbation in different tonal contexts, we built a series of linear mixed-effects models using the *lme4* package (Bates et al., 2014) on the normalized f0 (z-score). Z-scores were used instead of the raw f0 values (Hz), to facilitate comparisons across different speakers. In the initial model, we included the following factors as the fixed effects: ONSET (aspirated, unaspirated, sonorant), lexical TONE (T1, T2, T3, T4), VOWEL height (low, high), TIME points (eight categories from 0 to 7), and their interactions. TIME was coded using the orthogonal polynomial coding scheme and the rest of the fixed factors were Helmert-coded. The random effects structure of the model was determined using a forward best path algorithm (Barr et al., 2013), and the final model included by-SUBJECT random intercept, as well as by-SUBJECT slopes for ONSET, TONE, and VOWEL. The best fitting model was selected by comparing models using the likelihood ratio tests. The interactions ONSET * TONE * VOWEL * TIME and TONE * TIME * VOWEL did not improve the model fit [$\chi^2 = 10.60$, $p = 0.99$; $\chi^2 = 15.90$, $p = 0.78$, respectively] and, thus, they

[1] Note that the post-plosive f0 is likely affected by the preceding T1 in the carrier sentence (see, for example, Xu and Xu, 2003, for the discussion on this carryover effects). According to Xu and Xu (2003), both f0-ASP and f0-UNASP are higher after T1/T4 than after T2/T3, but f0-UNASP shows greater carryover effects than f0-ASP. If this is the case, it is possible that the preceding T1 elevated f0-UNASP more than f0-ASP and, consequently, the difference between f0-ASP and f0-UNASP in the current outcome is overplayed in low-initial tones but underplayed high-initial tones.

[2] Pinyin was included because this experiment was designed in parallel with a separate study testing L2 learners of Mandarin. Native Mandarin speakers would not need Pinyin to read common words in Chinese.

[3] We also tried a different method, in which we extracted the f0 values every 8 ms for the first 64 ms of the post-onset vowel, but the results were consistent with those obtained from the time-normalized method reported here.

TABLE 1 | F0 difference (z-score): aspirated–unaspirated (Tukey HSD *post-hoc* pairwise comparisons).

Time points		0 (0%)	1 (5%)	2 (10%)	3 (15%)	4 (20%)	5 (25%)	6 (30%)	7 (35%)
Tone	**Vowel**								
T1	Low	0.11***	0.15***	0.18***	0.19***	0.18***	0.15***	0.15***	0.12***
	High	0.27***	0.23***	0.20***	0.17***	0.16***	0.15***	0.13***	0.13***
T2	Low	−0.20***	−0.13***	−0.06$^{(*)}$	−0.02	0.00	0.01	0.02	0.03
	High	0.02	0.01	0.02	0.03	0.04	0.07*	0.07*	0.07*
T3	Low	−0.32***	−0.26***	−0.17***	−0.14***	−0.11***	−0.12***	−0.09**	−0.08*
	High	−0.14***	−0.16***	−0.14***	−0.13***	−0.11***	−0.10**	−0.09**	−0.07*
T4	Low	0.08**	0.10***	0.09**	0.06$^{(*)}$	0.03	−0.04	−0.06	−0.09**
	High	0.29***	0.23***	0.15***	0.09**	0.05	0.00	−0.03	−0.06$^{(*)}$

Significance codes: *** *for* $p < 0.001$, ** *for* $p < 0.01$, * *for* $p < 0.05$, *and* $^{(*)}$ *for* $p < 0.1$. *Shaded cells indicate significant f0 differences that are unidirectional starting from time point 0 and continuing without a break.*

were discarded. Consequently, the best model included four predictors ONSET, TONE, VOWEL, and TIME with the three-way interactions ONSET * TONE * TIME, ONSET * TIME * VOWEL, and ONSET * TONE * VOWEL. The outcome of this final model is in **Appendix B (Table B1)**.

Here, we present *p*-values for each significant factor and interaction obtained from the likelihood ratio tests comparing the best model and the model without the factor/interaction under consideration. Significant interactions were followed by *post-hoc* analyses using Tukey's HSD tests using the *emmeans* package (Lenth, 2020). If a predictor is significant in multiple interactions (or a main effect and interactions), only the highest-level interaction is reported along with the results of *post-hoc* testing.

We found the following significant interactions: ONSET: TONE: VOWEL [$\chi^2 = 2843.1$, $p < 0.0001$], ONSET: TONE: TIME [$\chi^2 = 1667.5$, $p < 0.0001$], ONSET: TIME: VOWEL [$\chi^2 = 250.8$, $p < 0.0001$]. As the predictor of our main interest, ONSET, was involved in multiple three-way interactions, we conducted the *post-hoc* Tukey pairwise comparisons on ONSET * TONE * VOWEL * TIME. The results of the pairwise comparisons are summarized in **Tables 1**, **3**, **4**, using the differences between the β coefficient values of different onset consonants. Shaded in **Tables 1**, **3**, **4** are the cells with significant f0 differences presumably attributable to onset consonants – that is, the cells with unidirectional f0 differences starting from time point 0 (closest to the onset consonant) and continuing without a break.

Figure 1 presents the mean f0 contours of the post-onset vowels. The contours are smoothed with loess and the shading displays a 95% confidence interval. To facilitate the visual interpretation of the figure, the z-normalized f0 is converted back to the Hz scale using the group mean (Brunelle et al., 2020), and the f0 contours of the entire duration of the post-onset vowels are plotted instead of the first 35% used in the statistical analysis. The vertical dotted line is added to indicate the 35% threshold included in the statistical analysis. The f0 contours are time-normalized, aligned from the voicing onset to the vowel offset (see Xu and Xu, 2021, for the comparison between different alignments). Individual speakers' production data are presented in the **Supplementary Materials**.

Aspirated and Unaspirated Stops

The f0 contours following an aspirated plosive (f0-ASP) and those following an unaspirated plosive (f0-UNASP) showed distinct patterns, but both the direction and the duration of the f0 differences varied according to the tonal contexts (**Table 1**). As for the direction of the f0 differences, f0-ASP was higher than f0-UNASP (indicated by positive numbers in **Table 1**) in T1 and T4, while the pattern showed the opposite direction in T2 and T3 (with the exception for /t^{h}u2/∼/tu2/ pair which showed no significant difference). The perturbation duration was also mediated by the tonal contexts. Specifically, the longest perturbation duration was observed in T1 and T3. In T1, the f0-ASP differed significantly from the f0-UNASP throughout the selected 35% of the vowel in T1 and T3 (corresponding to the mean duration of 75 ms in both tones), followed by T4 (10∼15% or 20∼30 ms). The perturbation due to aspiration (or lack thereof) was fairly limited in T2, either to the vowel onset (5% or 11 ms) in the /a/ context or not significant in the /u/ context.

As for the effects of VOWEL, the syllables with the high vowel /u/ had higher f0 than those with the low /a/, showing the expected vowel-intrinsic f0 patterns (e.g., Whalen and Levitt, 1995). This effect of vowel-intrinsic f0 was greater in high-initial tones than in low-initial tones (see **Figure 1**). In addition, the difference between f0-ASP and f0-UNASP in high-initial tones was greater in the /u/-contexts than in the /a/-contexts, but the same difference in low-initial tones was greater in the /a/-contexts than in the /u/-contexts.

In addition, aspirated plosives had longer VOT than unaspirated plosives, as expected (**Figure 2**). The influence of plosive ASPIRATION (aspirated, **unaspirated**), lexical TONE (**T1**, T2, T3, T4), and VOWEL height (**low**, high) on VOT (ms) was examined in a linear mixed effect model (Bates et al., 2014). The reference levels are bold-faced. The model included the interactions among the fixed factors, and by-SUBJECT random intercept. The model output is presented in **Appendix B (Table B2)**. The results revealed a significant three-way interaction ASPIRATION * TONE * VOWEL, and the follow-up Tukey's HSD tests (Lenth, 2020) confirmed that aspirated and unaspirated stops were significantly different [$\beta = -97.7$, $p <$

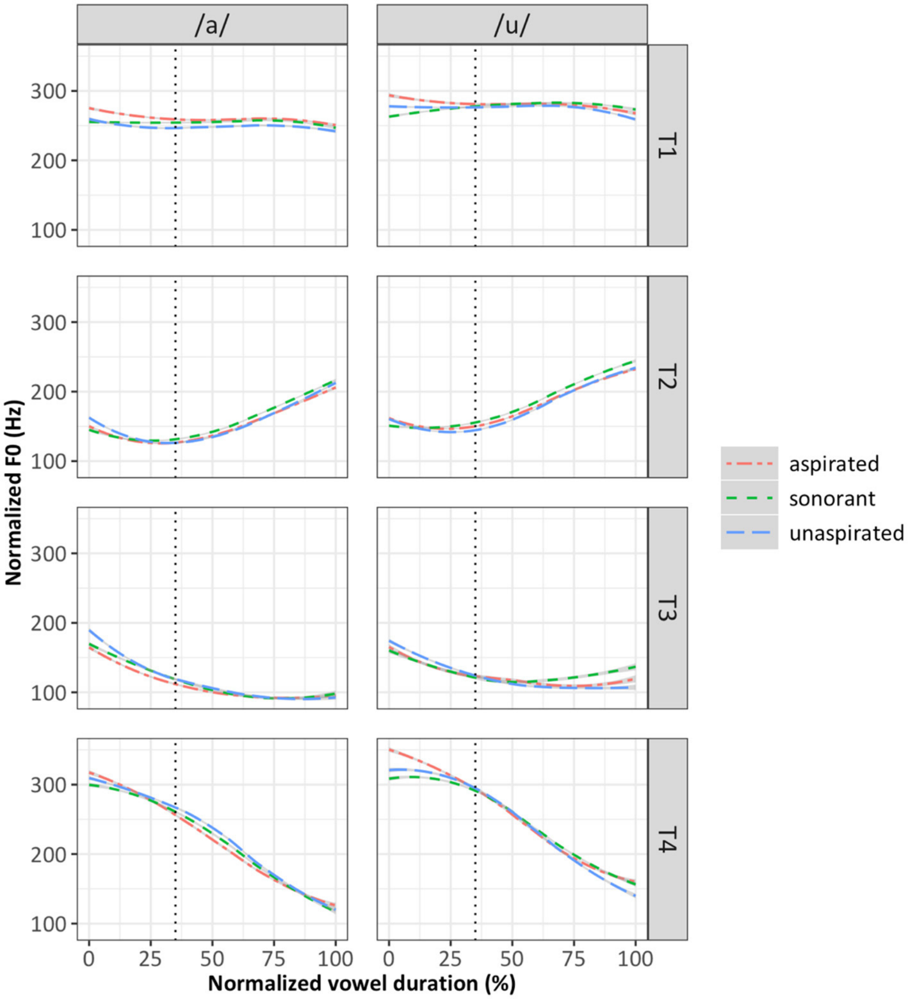

FIGURE 1 | Normalized F0 of Mandarin syllables.

0.0001]. Of interest to the current study, we also found significant effects of TONE on the VOT of aspirated plosives. As shown in **Figure 2**, the VOTs of aspirated plosives were the longest in T3, followed by T2, and T1 and T4 had the shortest VOT.[4] The results of the *post-hoc* Tukey's HSD comparisons are in **Table 2**. The VOTs of the unaspirated plosives did not show such effects of TONE.

Comparing Obstruents and Sonorants

Although the current study mainly aims to examine the f0 difference between aspirated and unaspirated plosives, we also compared f0-SON (f0 following a sonorant onset) with f0-ASP and f0-UNASP. Across different tone and vowel contexts, f0-UNASP was consistently greater than f0-SON, at least at the vowel onset (see **Table 3**). The difference between f0-ASP and f0-SON was less consistent (**Table 4**), varying mostly with the tonal contexts, in the same way as the difference between f0-ASP and f0-UNASP.

The duration of f0 perturbation varied in different tones as well as in different vowel contexts. The difference between f0-ASP and f0-SON mirrored the patterns showed between f0-ASP and f0-UNASP in high-initial tones. The difference between f0-ASP/UNASP and f0-SON also showed some influence of the vowel context. The difference lasted longer in the /u/-contexts than in the /a/-contexts in high-initial tones, but not in low-initial tones.

Interim Summary and Discussion

To summarize, the difference between f0-ASP and f0-UNASP showed opposite directions in high-initial tones and low-initial tones. On the other hand, the most consistent f0 difference across different tonal contexts was observed between f0-UNASP and f0-SON such that f0-UNASP was consistently higher than f0-SON. These outcomes suggest aspiration and voicing (or lack of voicing and aspiration) separately influenced the f0 at the vowel onset.

[4] Note our stimuli for T2 included bilabial /p^ha2/ and /pa2/ instead of /t^ha2/ and /ta2/. As coronal plosives usually have longer VOTs than labial plosives, this is expected to influence the reported VOT values for T2. We suspect that the VOT difference between T2 and T3 would have been exaggerated due to this difference in places of articulation.

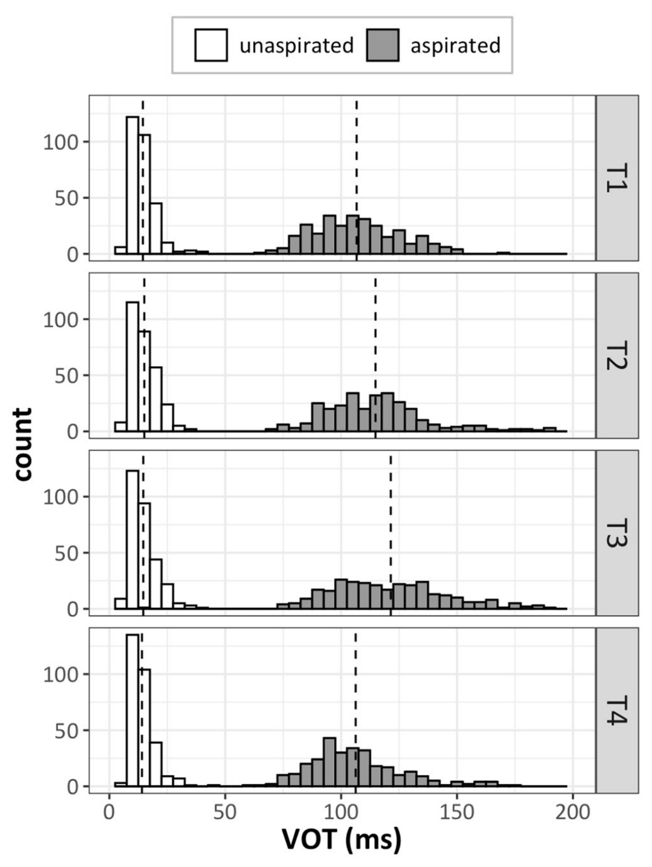

FIGURE 2 | Distribution of VOT across different tones. Dashed lines represent the mean VOT values.

TABLE 2 | Aspirated plosives' VOT in different tonal contexts (Tukey HSD *post-hoc* comparisons).

Tonal contrast	Estimate (β)	df	t ratio	*p*-value
T1–T2	−8.114	2347	−7.271	<0.0001***
T1–T3	−14.730	2347	−13.244	<0.0001***
T1–T4	0.494	2347	0.444	0.9708
T2–T3	−6.616	2347	−5.928	<0.0001***
T2–T4	8.608	2347	7.701	<0.0001***
T3–T4	15.224	2347	13.666	<0.0001***

*Significance codes: *** for p < 0.001, ** for p < 0.01, * for p < 0.05, and (*) for p < 0.1.*

First, aspirated plosives, compared to unaspirated plosives, influenced the f0 in different directions in high- vs. low-initial tones. Among the voiceless plosives, aspiration cooccurred with high f0 in the high-initial tones but with low f0 in the low-initial tones. The duration of this aspiration effect also depended on the tonal context. The difference between f0-ASP and f0-UNASP in the current study lasted the longest in T1 and T3, followed by T4. T2 showed little, if any, perturbation due to aspiration.

Second, although our main goal was to examine the perturbation due to consonant aspiration, we could also observe the voicing effect. F0-SON was consistently lower than f0-UNASP, suggesting that voicelessness raised (or voicing lowered) post-onset f0, consistent with the commonly observed cross-linguistic pattern. This effect was consistent throughout all tones.

The difference between f0-ASP and f0-SON seemed to reflect the interaction of these two effects. That is, if the f0-SON could be considered as the baseline, voicelessness (both unaspirated and unaspirated) raised f0, and in low-initial tones, aspiration lowered f0, resulting in little difference between f0-ASP and f0-SON. On the other hand, in high-initial tones, both aspiration and voicelessness raised f0, leading to a greater difference between f0-ASP and f0-SON.

The effect of vowel height interacted with the tonal contexts such that in high-initial tones, syllables with a high vowel showed greater f0 perturbation than those with a low vowel; in low-initial tones, syllables with a low vowel showed greater perturbation effects.

EXPERIMENT 2: PERCEPTION

Although complicated, the observed f0 perturbation patterns in Experiment 1 can be predicted as a function of the consonant's laryngeal category and the lexical tone. In this regard, the findings from Experiment 1 suggest that Mandarin has a systematic f0 perturbation at least for the tested speakers. Experiment 2 examines the perception of plosive aspiration contrast by the same Mandarin speakers. The purpose is to investigate whether Mandarin speakers, who produce systematically different f0 contours after aspirated and unaspirated plosives, use the f0 information to perceive the plosives' laryngeal categories.

Methods

Participants

The same individuals from Experiment 1 also participated in Experiment 2. Related to the task of Experiment 2, all participants reported to be right-handed.

Stimuli

Perception stimuli were created by recording natural productions of the syllables /t^hu/ in isolation, and manipulating them in Praat (Boersma and Weenink, 2020) to create a series of stops covarying in VOT and f0. A female native Mandarin speaker recorded the base syllables in four tones (i.e., /t^hu1/, /t^hu2/, /t^hu3/, /t^hu4/) in isolation. Aspirated stops were selected as the base tokens and unaspirated tokens were created by removing the aspirated portions from the base tokens. Consistent with previous studies using similar methods (e.g., Francis et al., 2006), removing the aspiration noise and shortening the VOT resulted in more natural sounding tokens than adding in aspiration noise and lengthening the VOT in our pilot works. The high back vowel /u/ was selected because /u/ provides a full set (all four tones) of real Mandarin words for both aspirated and unaspirated alveolar stops. We wanted to avoid the situation in which one of the choices is a word and the other is not. In addition, the vowel contexts did not influence the results in our pilot works using both /a/ and /u/ vowels.

TABLE 3 | F0 difference (z-score): unaspirated–sonorant (Tukey HSD *post-hoc* pairwise comparisons).

Time points		0 (0%)	1 (5%)	2 (10%)	3 (15%)	4 (20%)	5 (25%)	6 (30%)	7 (35%)
Tone	Vowel								
T1	Low	0.13***	0.05	−0.02	−0.05	−0.06$^{(*)}$	−0.06$^{(*)}$	−0.08*	−0.07*
	High	0.17***	0.11***	0.09**	0.05	0.02	0.00	−0.01	−0.02
T2	Low	0.24***	0.15***	0.06$^{(*)}$	0.02	−0.01	−0.04	−0.06$^{(*)}$	−0.07*
	High	0.14***	0.07*	0.02	−0.03	−0.07*	−0.12***	−0.13***	−0.16***
T3	Low	0.27***	0.20***	0.12***	0.09**	0.08*	0.07$^{(*)}$	0.05	0.04
	High	0.22***	0.17***	0.13***	0.10**	0.07*	0.05	0.03	0.00
T4	Low	0.16***	0.09**	0.03	0.02	0.02	0.03	0.03	0.03
	High	0.24***	0.19***	0.17***	0.15***	0.13***	0.13***	0.13***	0.12***

Significance codes: *** *for* $p < 0.001$, ** *for* $p < 0.01$, * *for* $p < 0.05$, *and* $^{(*)}$ *for* $p < 0.1$. *Shaded cells indicate significant f0 differences that are unidirectional starting from time point 0 and continuing without a break.*

TABLE 4 | F0 difference (z-score): aspirated–sonorant (Tukey HSD *post-hoc* pairwise comparisons).

Time points		0 (0%)	1 (5%)	2 (10%)	3 (15%)	4 (20%)	5 (25%)	6 (30%)	7 (35%)
Tone	Vowel								
T1	Low	0.23***	0.21***	0.17***	0.14***	0.12***	0.08**	0.07*	0.06$^{(*)}$
	High	0.44***	0.35***	0.29***	0.22***	0.18***	0.15***	0.12***	0.10**
T2	Low	0.04	0.02	0.00	0.00	−0.02	−0.04	−0.04	−0.04
	High	0.16***	0.08*	0.04	0.00	−0.03	−0.05	−0.07$^{(*)}$	−0.08*
T3	Low	−0.05	−0.06	−0.05	−0.05	−0.03	−0.05	−0.04	−0.04
	High	0.08*	0.01	0.00	−0.03	−0.04	−0.05	−0.06	−0.07$^{(*)}$
T4	Low	0.25***	0.19***	0.12***	0.08*	0.04	−0.01	−0.03	−0.06
	High	0.53***	0.41***	0.31***	0.24***	0.18***	0.13***	0.11***	0.06$^{(*)}$

Significance codes: *** *for* $p < 0.001$, ** *for* $p < 0.01$, * *for* $p < 0.05$, *and* $^{(*)}$ *for* $p < 0.1$. *Shaded cells indicate significant f0 differences that are unidirectional starting from time point 0 and continuing without a break.*

To obtain a fine-grained picture of the respective roles of VOT and f0 in the perception of Mandarin stop aspiration, 49 distinct syllables were initially created from each of the four base tokens (i.e., /t^{h}u1/, /t^{h}u2/, /t^{h}u3/, /t^{h}u4/). The 49 syllables covaried in stop VOT and post-stop f0, by fully crossing 7 steps of VOT and 7 steps of post-stop f0.

The mean VOT duration of the 4 base tokens was 99 ms, and the VOT step size was approximately 14 ms. Starting at the nearest zero crossing point from the end of the stop burst, about 14 ms of aspiration was manually removed incrementally in Praat until the VOT of the base token was around 14 ms. As a result, mean VOT values for each step were as follows: step 1 = 14 ms, step 2 = 28 ms, step 3 = 42 ms, step 4 = 56 ms, step 5 = 72 ms, step 6 = 86 ms, and step 7 = 99 ms.

Post-plosive f0 was manipulated using the TD-PSOLA (Moulines and Charpentier, 1990) implemented in Praat. First, the first 35% of the vowel was selected, and then the pitch curve of the selected vowel portion was simplified with the stylize function in Praat (frequency resolution 2 Hz). The onset f0 for each of the base tokens before manipulation were T1 = 323 Hz, T2 = 241 Hz, T3 = 210 Hz, and T4 = 371 Hz. Then, to create the 7 steps of post-plosive f0, the initial pitch point was either raised or lowered by 20 Hz, 40 Hz, and 60 Hz. F0 during the rest of the 35% of the vowel was proportionately increased or decreased. All the tokens were resynthesized with TD-PSOLA after the manipulation.

The tokens after manipulation were checked by four Mandarin native listeners for their naturalness, and all were judged to be good tokens of the original syllables. We conducted a pilot study with additional four Mandarin listeners, and VOT step 6 (84 ms) and step 7 (99 ms) never elicited different perceptual responses and, thus, VOT step 6 stimuli were removed from the experiment to keep the experiment short. The final set of perception stimuli included 168 (4 tones * 7 steps of f0 * 6 steps of VOT) unique tokens.

Procedure

Experiment 2, the perception experiment, was conducted after the production experiment, out of the concern that listening to the stimuli would influence the subsequent productions of the related sounds. After completing the production experiment, participants took a 5-min break before beginning the perception experiment.

Using PsychoPy (Peirce, 2007), the participants were presented with a forced-choice identification task. While listening to the stimuli, two Chinese characters constituting the aspirated and unaspirated pairs (e.g., 突/t^{h}u1/ vs. 督/tu1/) were

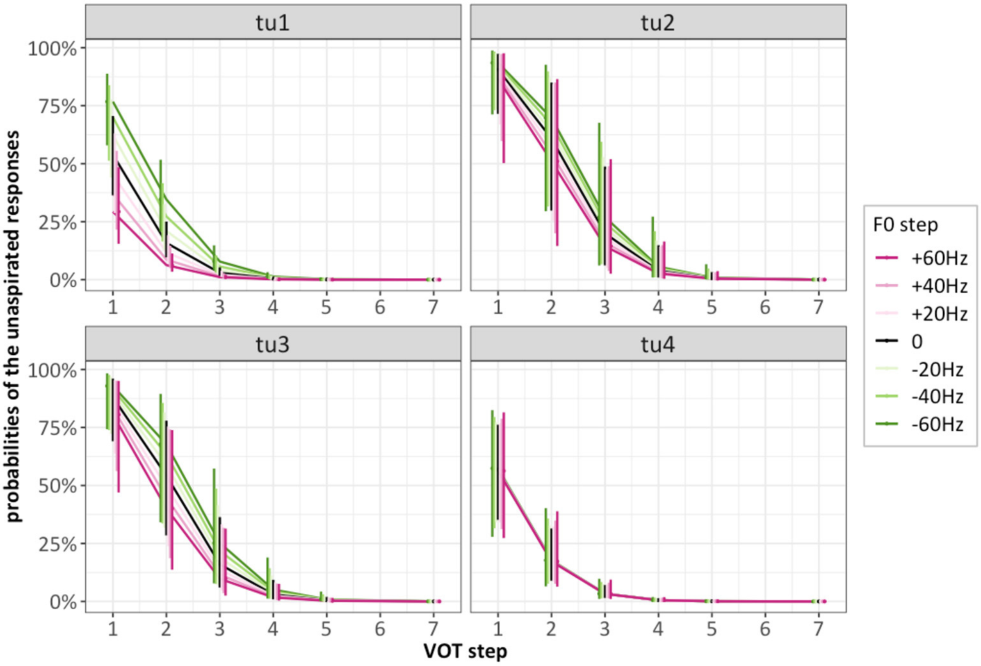

FIGURE 3 | Predicted perceptual responses based on the logistic regression model. Vertical lines represent the 95% confidence interval.

displayed on the laptop screen. Thirteen participants saw the screen with /t^h/- syllables on the left and /t/-syllables on the right, and 12 participants saw the opposite. The auditory stimuli were presented through Sennheiser HD 280 pro headphones. The participants were instructed to choose the word they heard by selecting one of the two characters using a Cedrus button box (model RB-740).

The experiment was blocked by the lexical tones and the order among the blocks was counter-balanced across participants. Within each block, each of the 42 tokens (7 f0 steps * 6 VOT steps) was repeated three times in different random orders. There were self-paced breaks between blocks. The entire task took about 20 minutes.

Results

A total of 12,600 responses (25 participants * 4 blocks * 42 tokens * 3 repetitions) were collected. Prior to the statistical analyses, the responses with the reaction time (measured from the onset of the audio stimuli to the button hit) that are more than 3 standard deviations away from the participant's mean (232 responses, 1.8%) were discarded. Then, to determine the influence of each acoustic property (VOT, post-plosive f0) on the identification of the onset laryngeal category, the responses (aspirated vs. unaspirated) were statistically analyzed using the binary logistic regression models built with the *lme4* packages in R (Bates et al., 2014). The reference category for the responses was aspirated and, thus, the coefficients β represent the log odds of unaspirated responses. The full model initially included VOT STEP, F0 STEP, TONE (T1, T2, T3, T4), and their interactions, as fixed effects. VOT STEP (1–7 without step 6) and F0 STEP (1-7) were included as continuous variables. TONE was orthogonally contrast coded (T1, T4 vs. T2, T3; T1 vs. T4; T2 vs. T3) to examine whether there are significant response differences between the high-initial tones (T1, T4) and the low-initial tones (T2, T3), as well as within the two tonal groups. The random effects structure of the model was determined using a forward best path algorithm (Barr et al., 2013), and the final model included by-SUBJECT and by-WORD intercepts, as well as by-SUBJECT slopes for VOT STEP, F0 STEP, and TONE. Interaction terms between fixed effects were included if they were directly related to our research question or if their inclusion improved the model fit based on a likelihood ratio test ($p < 0.05$). As a result, the final model included F0 STEP * TONE which was central to our research question. The full outcome of this final model is in **Appendix B (Table B3)**. A graph of predicted responses is in **Figure 3**. Raw response data for individual listeners are in the **Supplementary Materials**.

The likelihood ratio tests comparing the best model and the model without the predictor under consideration indicated that all fixed effects significantly influenced the listeners' responses. First, VOT STEP significantly contributed to model fit [$\chi^2 = 45.56, p < 0.0001$]. As shown in **Figure 3**, VOT step 1 elicited the highest rate of unaspirated responses across the four tones and, as the VOT increased, the possibility of unaspirated responses decreased. Second, the F0 STEP was also significant [$\chi^2 = 14.37$, $p = 0.0062$]: the higher the F0 STEP is, the less likely it is to elicit the unaspirated responses. Finally, as for TONE, the first tonal contrast (T1, T4 vs. T2, T3) contributed significantly to model fit [$\chi^2 = 30.47, p < 0.0001$], and high-initial tones (T1 and T4) elicited significantly less unaspirated responses than low-initial tones (T2 and T3) [$\beta = -3.82, p < 0.0001$]. The differences

TABLE 5 | Estimated trend of F0 step on tonal contrast (Tukey HSD *post-hoc* comparisons).

Tonal contrast	Estimate (β)	Standard Error	z. ratio	*p*-value
T1–T2	−0.206	0.134	−1.537	0.4152
T1–T3	−0.150	0.134	−1.122	0.6758
T1–T4	−0.323	0.141	−2.286	0.1013
T2–T3	0.055	0.125	0.442	0.9712
T2–T4	−0.118	0.134	−0.881	0.8148
T3–T4	−0.173	0.134	−1.293	0.5674

between the high-initial tones and the low-initial tones were the most conspicuous when VOT was short, as shown in **Figure 3**. For example, while the unaspirated responses were less than 50% at VOT step 2 in tones 1 and 4, in tones 2 and 3, a similar decrease was at step 3. This indicates that the stimuli belonging to the second step of VOT (28 ms), for instance, more likely elicited aspirated responses in the high-initial tones, but unaspirated responses in the low-initial tones. The second [$p = 0.74$] and third [$p = 0.35$] tonal contrasts were not significant, suggesting that listeners' responses in T1 vs. T4 and T2 vs. T3 were not significantly different.

The interaction F0 STEP: TONE was not significant [$\chi^2 = 5.37$, $p = 0.15$], but was included in the model as it was central to our research question. To verify whether the effects of F0 STEP across different tones, displayed in **Figure 3**, differed significantly, *post-hoc* Tukey tests were performed using the emtrends() function in the *emmeans* package (Lenth, 2020). It has been suggested that *post-hoc* analyses on non-significant interactions can be informative when the main effects of the predictors participating in an interaction are significant (e.g., Wei et al., 2012). The results of these *post-hoc* analyses suggest that the effects of F0 STEP did not differ as a function of TONE. None of the pairwise comparisons were significant, as shown in **Table 5**. Therefore, the current data do not provide evidence that the F0 STEP effects were influenced by tones. Rather, the current outcome appears to suggest that Mandarin listeners associated high post-plosive f0 with aspirated plosives across different tones.

Interim Summary and Discussion

The current findings demonstrate, as expected, that VOT is the primary cue of aspiration contrast in Mandarin. The unaspirated responses decreased as VOT became longer, across all f0 steps and lexical tones. At VOT step 1 (14 ms), which falls in the typical VOT range of the Mandarin unaspirated plosives (e.g., Rochet and Fei, 1991), the listeners provided the highest number of unaspirated responses, and starting from VOT step 4 (56 ms), the listeners tended to give mainly aspirated responses. The VOT categorical boundary for the aspirated-unaspirated plosives seemed to be different between high-initial tones vs. low-initial tones. Specifically, the VOT categorical boundaries occurred one step earlier in the high-initial tone stimuli than in the low-initial tone stimuli. At step 2, the low-initial tone stimuli yielded mostly unaspirated responses whereas the high-initial tone stimuli were more likely to yield aspirated responses (see **Figure 3**).

Although VOT was clearly the most influential cue for the aspiration, the listeners still used post-plosive f0 in deciding whether the plosive was aspirated or not. The current outcomes related to the f0 steps and lexical tones commonly suggest that the listeners associated high post-plosive f0 with the aspirated stops and low post-plosive f0 with unaspirated stops. The stimuli with raised f0 elicited more aspirated responses than those with lowered f0. In addition, stimuli with low-initial tones (T2, T3) elicited significantly more unaspirated responses than stimuli with high-initial tones (T1, T4). This is consistent with the pattern observed in the production experiment in which the aspirated plosives in T2 and T3 had longer VOT than those in T1 and T4 (**Figure 2**). This suggests that lower post-plosive f0, whether it be a part of the lexical tone or not, made the stops with an ambiguous VOT more likely to be judged as unaspirated than as aspirated.

Taken together, the current results suggest that Mandarin listeners extracted both consonantal and tonal information from f0 at the vowel onset. This perceptual pattern, however, did not precisely reflect the f0 perturbation observed in the same speakers' production patterns. In production, the difference between f0-ASP and f0-UNASP was not consistent across different tones, showing the opposite directions in high- vs. low-initial tones. Despite this divergent pattern in production, when VOT was ambiguous, the same speakers gave more aspirated responses in higher f0 steps both in high-initial and low-initial tones.

DISCUSSION

Post-onset F0 in Production

The main findings of Experiment 1, which compares f0-ASP, f0-UNASP, and f0-SON in four tonal contexts, can be summarized as the following. First, the difference between f0-ASP and f0-UNASP shows the opposite directions in high-initial tones and low-initial tones. In high-initial tones, f0-ASP is higher than f0-UNASP whereas f0-ASP is lower than f0-UNASP in low-initial tones. Second, f0-UNASP is consistently higher than f0-SON throughout the tonal contexts. Third, the difference between f0-ASP and f0-SON reflects the combination of these two effects. These outcomes suggest that the f0 at the vowel onset in Mandarin shows two separate perturbation effects, one due to aspiration and the other due to voicing. Between aspirated and unaspirated voiceless plosives, f0-ASP is higher in high-initial tones and lower in low-initial tones than f0-UNASP. Between voiceless plosives and voiced sonorants, voicelessness raises (or voicing lowers) f0 across the tonal contexts. Consequently, the difference between f0-ASP and f0-SON is greater in high-initial tones than in low-initial tones.

The current findings on f0 perturbation due to aspiration are partially consistent with the conflicting previous findings on Mandarin. Our findings in the low-initial tones are in line with Xu and Xu (2003), showing that aspiration lowers the post-plosive f0, compared to f0-UNASP, in low-initial tones. At the same time, we also find that in high-initial tones, aspiration raises the post-plosive f0, again compared to f0-UNASP, and this outcome is consistent with Luo's (2018) findings. The raising effects of the consonantal aspiration in Luo (2018) are greater in high-initial tones, the lowering effects in Xu and Xu (2003) are greater in low-initial tones, and our data show both of these patterns. These findings, taken together, reaffirm the dichotomy between the high-initial and low-initial tones.

The exact source of this dichotomy is puzzling, but we suggest that the tonal dichotomy is consistent with the interpretation that the f0 perturbation due to aspiration in Mandarin is biomechanically motivated. The observed tonal dichotomy can be explained by the differences in the laryngeal settings utilized in different tones. According to Moisik et al. (2014), the larynx height in general is positively correlated with f0 in Mandarin tone productions. As the laryngeal setting influences the vocal fold tension (e.g., Honda et al., 1999; Moisik et al., 2014), in high tones, the larynx is usually raised and the vocal folds are stretched and stiffened whereas the larynx is lowered and the vocal folds are slackened in low tones. When vocal folds are stiffened, they are resistant to vibration (i.e., require a greater volume of air flowing more rapidly than slack folds), but once they are set to vibrate, they vibrate at a high frequency. Also, stiffer vocal folds are often accompanied by a narrower glottal opening during the voiceless portion of a plosive (e.g., McCrea and Morris, 2005; Narayan and Bowden, 2013). On the other hand, slackened vocal folds are more prone to vibration and a wide glottal opening during a plosive.

The difference in the status of the vocal folds and the glottis has two notable consequences in the current study. The first consequence is the VOT difference in high-initial vs. low-initial tones. In the current study, aspirated plosives in high-initial tones have shorter VOT than those in low-initial tones (see **Figure 2**). According to McCrea and Morris (2005) and Narayan and Bowden (2013), stiff vocal folds and a narrow glottal opening result in shorter VOT of aspirated plosives, presumably accompanied by a faster airflow, in high f0 environments than in low f0 environments. The second consequence is the influence of aspiration on the post-plosive f0. Depending on the laryngeal settings for different tones, the influence of plosive aspiration on the post-plosive f0 can take different forms. According to the aerodynamic predictions, as claimed in Xu and Xu (2003), aspirated plosives, with a greater volume of air escaping through glottis between the oral release and the voicing onset, have a lower subglottal air pressure than unaspirated plosives at the voicing onset. This results in the f0-ASP being lower than f0-UNASP. This pattern (f0-ASP < f0-UNASP) appears when the vocal folds are slack and the glottal opening is wider, as in the low-initial tones in Mandarin. We claim that, in the high-initial tones, the aerodynamic effect is manifested in a different form because of the high tension of the vocal folds. As stiff vocal folds are more resistant to vibration and require a faster airflow to vibrate, the subglottal air pressure would not go down as much even in aspirated plosives. That is, the laryngeal setting and the resulting vocal fold tension in the high-initial tones require a higher trans-glottal pressure threshold than those in the low-initial tones, to initiate phonation at the onset of voicing after the plosive release. If the subglottal air pressure were to go down to the same extent regardless of the vocal fold tension, the trans-glottal pressure difference would not have been enough for the stiff folds to vibrate in the high-initial tones. Consequently, in the high-initial tones, the faster airflow in aspirated plosives (than in unaspirated plosives, see also Klatt et al., 1968), when combined with the high tissue tension and the narrow glottal opening, would increase the f0-ASP more than f0-UNASP. Chen (2011) proposes a similar dichotomy (tense vocal folds in a high-f0 context and slackened vocal folds in a low-f0 context, giving rise to distinct f0 perturbation patterns) for cross-linguistic variation. Our findings suggest that the tonal dichotomy can be observed even within a language.

Finally, although we suggest that the f0 perturbation in Mandarin is attributable to the biomechanics of the larynx, the current findings are also consistent, in several different aspects, with the claim that speakers of tonal languages would control the f0 perturbation to enhance (or not to impede) the tonal contrast (e.g., Hombert et al., 1979; Francis et al., 2006). First, the magnitude of the perturbation is greater in high-initial tones than in low-initial tones. Assuming that the high tones are salient in Mandarin (Luo, 2018) and the tones that are already salient do not need to be further enhanced, Mandarin speakers have more room for f0 variation in high-initial tones than in low-initial tones. This, according to Luo (2018), is the reason why the f0 raising due to aspiration is greater in high-initial tones in her study. Our findings differ from Luo's (2018) that we observe not only the f0 raising in high-initial tones but also the lowering in low-initial tones. Still the size of the difference between f0-ASP and f0-UNASP is greater in high-initial tones than in low-initial tones (**Table 1**), consistent with the claim that speakers would restrict the biomechanically-motivated f0 fluctuations when the tonal contrast is less salient and, thus, more vulnerable to misperception. Second, the perturbation lasts longer in the tones with a static f0 contour during the first half of the vowel than in those with a dynamic f0 contour. In Mandarin, the f0 contours for T1 and T3 are relatively steady during the first half of the vowel whereas those for T2 and T4 are more dynamic (see the section: Lexical tones). And the current findings indicate that the difference between f0-ASP and f0-UNASP lasts the longest in T1 and T3, followed by T4, and then T2 (**Table 1**). This seems to provide evidence for the speakers' control over f0 perturbation in a (subconscious) effort to preserve the tonal contrast. When the tones require dynamic f0 changes earlier in the vowel, speakers suppress the f0 variation automatically induced by the onset consonant. Tones with relatively steady f0 contours, on the other hand, would allow for more variability in f0 due to non-tonal factors, such as the aspiration of onset consonants.

Post-onset F0 in Perception

Our production data show that the f0 perturbation in Mandarin varies according to the lexical tones. As discussed in Post-onset f0 in production, this variation appears to be systematic, reflecting different laryngeal maneuvers for different tonal targets. Still, the same speakers, when they are presented with auditory stimuli varying in plosive VOT and post-plosive f0, are more likely to select the aspirated category when the post-plosive f0 is high and when VOT is ambiguous. The associations (high f0-aspirated and low f0-unaspirated) are valid even in the low-initial tones which show an opposite perturbation pattern in the production. In other words, there seems to be an intriguing mismatch between the production and the perception with regard to Mandarin speakers' use of f0 as a cue for consonant aspiration.

We propose several different factors contributing to this apparent mismatch. First, listeners are more attentive to the phonetic patterns present in salient contexts. Since Mandarin high-initial tones are more salient than low-initial tones both phonologically and perceptually, as suggested by Luo (2018), listeners may use the pattern presented in the salient tones that associates high f0 with aspirated plosives even when they perceive the low-initial tone stimuli. The production patterns in less salient low-initial tones are likely to be unattended. Second, the distribution of Mandarin lexical tones also suggests that the perturbation patterns in high-initial tones are more prevalent in the language. Liu and Ma (1986), based on their survey of two different corpora, the National Standard Corpus of Mandarin Words and the Chinese Vocabulary Corpus, show that T4 is the most frequent (32%) and T3 is the least frequent (17%) in Mandarin. T1 and T2 account for 24∼25% of Mandarin words. This means that the two high-initial tones (T1 and T4) compose more than half (56∼57%) of the Mandarin lexicon while the two low-initial tones, when combined, comprise about 40% of the lexicon. In addition, T3 is subject to tone sandhi (Duanmu, 2007), and when followed by another T3, becomes T2, which has the minimal, if any, perturbation due to aspiration (see **Table 1**, and also the same pattern is reported in Luo, 2018). Taking all these together, Mandarin listeners are presumably exposed to the f0 perturbation pattern that f0-ASP is higher than f0-UNASP more frequently than to the opposite pattern. Also, even in the infrequent cases when the listeners actually hear the pattern of f0-ASP < f0-UNASP, they are less likely to attend to this covariation occurring in less salient tonal contexts. Therefore, we claim that the Mandarin listeners' perception reflects the predominant pattern in their production. The perturbation pattern from the low-initial tones (f0-ASP < f0-UNASP) is not robustly represented, as T3 is the least frequent in the language and vulnerable to sandhi, and the perturbation in T2 is weak at best. Consequently, Mandarin listeners are likely to learn, from their native language experience, that high f0 is associated with aspirated plosives and low f0 with the unaspirated plosives, and use the high post-plosive f0 as a secondary cue to consonant aspiration.

Francis et al. (2006) also report a discrepancy between production and perception, in their investigation of the f0 perturbation in Cantonese. Cantonese listeners use post-plosive f0 as a cue for consonant aspiration but Cantonese speakers' production does not provide evidence for the association between high f0 and plosive aspiration. As the listeners' perceptual responses cannot be explained by their native language experience, Francis et al. (2006) claim that the listeners' perception is guided by a language-independent, general auditory enhancing effects among different phonetic properties (e.g., Kingston and Diehl, 1994), which could have been facilitated by the listeners' experience with English. Unlike Francis et al. (2006), we do see evidence for the association between high f0 and aspiration in Mandarin speakers' production. This suggests that the perceptual pattern observed in the current study may not be entirely due to the general auditory effects but, rather, due to the listeners' native language experience. However, we acknowledge that we cannot rule out the potential influence of the English experience. The participants in this study speak English as their second language, residing in Virginia, USA at the time of testing. We still expect the English influence, if any, to be minimal since bilingual listeners' categorization, which requires language-specific phonological judgments, shows the language mode effects (e.g., Antoniou et al., 2012). In this study, the experiments were carried out in Mandarin by a native Mandarin-speaking experimenter, and the perception task asked the listeners to select the Mandarin character matching the stimuli they heard.

Concluding Remarks: Mandarin Aspiration Contrast

The current outcomes confirm that VOT is the phonetic property primarily responsible for Mandarin aspiration contrast. In production, Mandarin aspirated plosives and unaspirated plosives are well-separated by the VOT alone (**Figure 2**), and Mandarin listeners primarily rely on VOT to distinguish the aspirated plosives from the unaspirated ones in perception (**Figure 3**). The VOT boundary, however, seems to vary according to the tonal contexts. The VOT of aspirated plosives is greater in low-initial tones than in high-initial tones in production (**Figure 2**). We suggest that this variation arises from the biomechanics of the larynx as, in high f0 ranges, VOT of aspirated stops decreases due to vocal fold tension (McCrea and Morris, 2005; Narayan and Bowden, 2013). In perception, Mandarin listeners are sensitive to this contextual VOT variation, providing more unaspirated responses in low-initial tones than in high-initial tones (**Figure 3**). Taken together, these findings suggest that the VOT boundary for Mandarin aspiration contrast is flexible and influenced by the tonal contexts. This is comparable to the well-documented covariation between VOT and place of articulation. In production, labial plosives have the shortest VOT, with the plosives of backer places of articulation having longer VOT (e.g., Peterson and Lehiste, 1960; Cho and Ladefoged, 1999). And listeners attend to this systematic variation. For example, the VOT boundary between voiced and voiceless categories is at a lower VOT range in labial plosives than in velar plosives (e.g., Miller, 1977; Benkí, 2001). When the variation in the speech signal is systematic, although it may not be uniform across contexts, the contextual variation does not impede but facilitates listeners' perception.

The current findings also provide evidence for a systematic variation in post-plosive f0 influenced both by the consonant aspiration and by the lexical tone. Depending on whether the tone begins at a high vs. low f0 range, the consonantal influence on f0 takes a different form. This can be attributed to the different laryngeal settings for different tonal targets. Despite the variation in production, Mandarin listeners use the post-plosive f0 as a secondary cue for plosive aspiration, associating high f0 with the aspirated category even in the low-initial tones which show an opposite perturbation pattern in the production. When the stimuli VOT is within a typical range of aspirated or unaspirated plosives, the listeners' responses are predominantly determined by the stop VOT. However, when the VOT is ambiguous (step 2 in high-initial tones and step 3 in low-initial tones, **Figure 3**), high post-plosive f0 stimuli, in general, yielded more aspirated responses despite a fairly large inter-listener variation (see the individual data in the **Supplementary Material**). That being said, the overall perceptual pattern pooled across the listeners may arguably originate from the f0 perturbation patterns in Mandarin production. As the high-initial tones are more salient and more prevalent in the Mandarin lexicon, the listeners attend more to the perturbation patterns present in high-initial tones (f0-ASP > f0-UNASP) than those in low-initial tones (f0-UNASP > f0-ASP). Although post-plosive f0 varies according to the tonal contexts in production, its role as the secondary cue to consonant aspiration in perception does not seem to be modulated by the tonal contexts.

Finally, the current study only reports the pooled results, but we should note that the data exhibit a considerable individual variation in both experiments (see the **Supplementary Material**). In production, some speakers show a quite clear f0 perturbation conforming to the group pattern while others show the conforming pattern only in a few tones but not in the others. In perception, post-plosive f0 does not seem to be an informative cue to consonant aspiration for all listeners, and some listeners seem to use f0 differently than others. The reason for these variations is unclear, and they do not seem to be structured in an immediately noticeable way. Still, this individual variation is intriguing and calls for a focused investigation, which we leave for a future study.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by George Mason University IRB. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YG: study conception and design, data collection and analysis, interpretation of results, and writing the initial draft. HK: supervising, data analysis, data visualization, interpretation of results, and writing and revising the manuscript. Both authors reviewed the results and approved the final version of the manuscript.

FUNDING

This work was supported by the George Mason University Dissertation Completion Grant awarded to YG and by the Linguistics Program at George Mason University.

ACKNOWLEDGMENTS

This work is a revised version of part of YG's dissertation (*Production and Perception of Laryngeal Contrasts in Mandarin and English by Mandarin Speakers*. George Mason University, 2020). The authors would like to thank Georgia Zellou and two reviewers for their constructive comments, the audience at the 178th meeting of the Acoustical Society of America and LabPhon 17 where portions of this work have been presented, Steven Weinberger, Vincent Chanethom, and Doug Wulf for helpful discussions related to this work, and participants for making this study possible.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.896013/full#supplementary-material

REFERENCES

Antoniou, M., Tyler, M. D., and Best, C. T. (2012). Two ways to listen: do L2-dominant bilinguals perceive stop voicing according to language mode? *J. Phone.* 40, 582–594. doi: 10.1016/j.wocn.2012.05.005

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear Mixed-Effects Models Using Eigen and S4. R Package Version 1.1-7*. Available online at: http://CRAN.Rproject.org/package1/4lme4

Benkí, J. R. (2001). Place of articulation and first formant transition pattern both affect perception of voicing in English. *J. Phone.* 29, 1–22. doi: 10.1006/jpho.2000.0128

Boersma, P., and Weenink, D. (2020). *Praat: Doing Phonetics by Computer*, *Version 6.1.12*. Available online at: http://www.praat.org/

Brunelle, M., Tấn, T. T., Kirby, J., and Giang, Ð. L. (2020). Transphonologization of voicing in chru: studies in production and perception. *Lab. Phonol.* 11, 15. doi: 10.5334/labphon.278

Chen, Y. (2011). How does phonology guide phonetics in segment–f0 interaction? *J. Phone.* 39, 612–625. doi: 10.1016/j.wocn.2011.04.001

Chi, Y., Honda, K., and Wei, J. (2019). "Glottographic and aerodynamic analysis on consonant aspiration and onset f0 in Mandarin Chinese," in; *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brighton), 6480–6484.

Cho, T., and Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *J. Phone.* 27, 207–229. doi: 10.1006/jpho.1999.0094

Coetzee, A. W., Beddor, P. S., Shedden, K., Styler, W., and Wissing, D. (2018). Plosive voicing in afrikaans: differential cue weighting and tonogenesis. *J. Phone.* 66, 185–216. doi: 10.1016/j.wocn.2017.09.009

Deng, D. 邓丹, Feng, S. 石锋, and Lu, S. 吕士楠(2006). 普通话与台湾国语声调的对比分析[The contrast on tone between Putonghua and Taiwan Mandarin]. 声学学报*[Sheng Xue Xue Bao – Acta Acoustica]* 31, 536–541.

Dmitrieva, O., Llanos, F., Shultz, A. A., and Francis, A. L. (2015). Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English. *J. Phone.* 49, 77–95. doi: 10.1016/j.wocn.2014.12.005

Duanmu, S. (2007). *The Phonology of Standard Chinese*. New York, NY: Oxford University Press.

Francis, A. L., Ciocca, V., Wong, V. K. M., and Chan, J. K. L. (2006). Is fundamental frequency a cue to aspiration in initial stops? *J. Acoust. Soc. Am.* 120, 2884–2895. doi: 10.1121/1.2346131

Gandour, J. T. (1974). Consonant types and tone in Siamese. *J. Phone.* 2, 337–350. doi: 10.1016/S0095-4470(19)31303-8

Gao, J., and Arai, T. (2019). Plosive (de-)voicing and f0 perturbations in Tokyo Japanese: positional variation, cue enhancement, and contrast recovery. *J. Phone.* 77, 100932. doi: 10.1016/j.wocn.2019.100932

Halle, M., and Stevens, K. N. (1971). A Note on Laryngeal Features. Quarterly Progress Report, Research Laboratory of Electronics, MIT. 101, 198–213.

Hallé, P. (1994). Evidence for tone-specific activity of the sternohyoid muscle in modern standard Chinese. *Lang. Speech* 37, 103–123. doi: 10.1177/002383099403700201

Hanson, H. M. (2009). Effects of obstruent consonants on fundamental frequency at vowel onset in English. *J. Acoust. Soc. Am.* 125, 425–441. doi: 10.1121/1.3021306

Hombert, J. M., Ohala, J. J., and Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language* 55, 37–58. doi: 10.2307/412518

Honda, K. (2004). Physiological factors causing tonal characteristics of speech: From global to local prosody. *Proc Speech Prosody* 2004, 739–744.

Honda, K., Hirai, H., Masaki, S., and Shimada, Y. (1999). Role of vertical larynx movement and cervical lordosis in F0 control. *Lang. Speech* 42, 401–411. doi: 10.1177/00238309990420040301

Hoole, P., and Honda, K. (2011). "Automaticity vs. feature-enhancement in the control of segmental f0," in *Where do phonological features come from?: Cognitive, physical and developmental bases of distinctive speech categories Language Faculty and Beyond (LFAB): Internal and external variation in linguistics*, eds N. Clements and R. Ridouane (Amsterdam: John Benjamins), 133–171.

House, A. S., and Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *J. Acoust. Soc. Am.* 25, 105–113. doi: 10.1121/1.1906982

Jessen, M., and Roux, J. C. (2002). Voice quality differences associated with stops and clicks in Xhosa. *J. Phone.* 30, 1–52. doi: 10.1006/jpho.2001.0150

Kingston, J. (2007). "Segmental influences on f0: automatic or controlled?" in *Tones and Tunes, Volume 2: Experimental Studies in Word and Sentence Prosody*, eds Gussenhoven and T. Riad (Berlin: Mouton de Gruyter), 171–201.

Kingston, J., and Diehl, R. L. (1994). Phonetic knowledge. *Language* 70, 419–494. doi: 10.1353/lan.1994.0023

Kirby, J. (2018). Onset pitch perturbations and the cross-linguistic implementation of voicing: Evidence from tonal and non-tonal languages. *J. Phone.* 71, 326–354. doi: 10.1016/j.wocn.2018.09.009

Kirby, J., and Ladd, D., R. (2016). Effects of obstruent voicing on vowel F0: evidence from "true voicing" languages. *J. Acoust. Soc. Am.* 40, 2400–2411. doi: 10.1121/1.4962445

Klatt, D. H., Stevens, K. N., and Meade, J. (1968). "Studies of articulatory activity and airflow during speech in sound production in man," in *Annals of the New York Academy of Science*, eds A. Bouhuys (New York, NY), 42–55.

Kohler, K. J. (1982). F0 in the production of lenis and fortis plosives. *Phonetica* 39, 199–218. doi: 10.1159/000261663

Kuang, J. (2017). Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *J. Acoust. Soc. Am.* 142, 1693–1706. doi: 10.1121/1.5003649

Lehiste, I., and Peterson, G. E. (1961). Some basic considerations in the analysis of intonation. *J. Acoust. Soc. Am.* 33, 419–425. doi: 10.1121/1.1908681

Lenth, R. (2020). *emmeans: Estimated Marginal Means, aka Least-Squares Means. R Package Version 1.4.5.* Available online at: https://CRAN.R-project.org/package=emmeans

Liu, L. Y. 刘连元, and Ma, Y. F. 马亦凡(1986). 普通话声调分布和声调结构频度[The distribution of Mandarin tones and the frequency of tonal phrases]. 语文建设 *[Language Planning]* 3, 21–23.

Löfqvist, A., Baer, T., McGarr, N. S., and Story, R. S. (1989). The cricothyroid muscle in voicing control. *J. Acoust. Soc. Am.* 85, 1314–1321. doi: 10.1121/1.397462

Luo, Q. (2018). *Consonantal Effects on F0 in Tonal Languages (Doctoral dissertation)*. Michigan State University, East Lansing.

McCrea, C. R., and Morris, R. J. (2005). The effects of fundamental frequency levels on voice onset time in normal adult male speakers. *J. Speech Lang. Hear. Res.* 48, 1013–1024. doi: 10.1044/1092-4388(2005/069)

Miller, J. L. (1977). Nonindependence of feature processing in initial consonants. *J. Speech Lang. Hear. Res.* 20, 519–528. doi: 10.1044/jshr.2003.519

Mohr, B. (1971). Intrinsic variations in the speech signal. *Phonetica* 23, 65–93. doi: 10.1159/000259332

Moisik, S. R., Lin, H., and Esling, J. H. (2014). A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS). *J. Int. Phon. Assoc.* 44, 21–58. doi: 10.1017/S0025100313000327

Moulines, E., and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun.* 9, 453–467. doi: 10.1016/0167-6393(90)90021-Z

Narayan, C., and Bowden, M. (2013). Pitch affects voice onset time (VOT): a cross-linguistic study. *Proc. Meet. Acoust.* 19, 060095. doi: 10.1121/1.4800681

Ohde, R. N. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *J. Acoust. Soc. Am.* 75, 224–230. doi: 10.1121/1.390399

Peirce, J. W. (2007). PsychoPy—psychophysics software in python. *J. Neurosci. Methods* 162, 8–13. doi: 10.1016/j.jneumeth.2006.11.017

Peterson, G. E., and Lehiste, I. (1960). Duration of syllable nuclei in English. *J. Acoust. Soc. Am.* 32, 693–703. doi: 10.1121/1.1908183

R Core Team. (2021). *A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Available online at: https://www.R-project.org

Rochet, B. L., and Fei, Y. (1991). Effect of consonant and vowel context on Mandarin Chinese VOT: production and perception. *Can. Acoust.* 19, 105.

Wei, J., Carroll, R. J., Harden, K. K., and Wu, G. (2012). Comparisons of treatment means when factors do not interact in two-factorial studies. *Amino Acids* 42, 2031–2035. doi: 10.1007/s00726-011-0924-0

Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (1990). Gradient effects of fundamental frequency on stop consonant voicing judgments. *Phonetica* 47, 36–49. doi: 10.1159/000261851

Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times. *J. Acoust. Soc. Am.* 93, 2152–2159. doi: 10.1121/1.406678

Whalen, D. H., and Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *J. Phone.* 23, 349–366. doi: 10.1016/S0095-4470(95)80165-0

Xiao, H. 肖航(2010). 现代汉语通用平衡语料库建设与应用[The construction and application of the general modern Chinese balanced corpus]. 华文世界 *[Chinese World]*. 106, 24–29.

Xu, C. X., and Xu, Y. (2003). Effects of consonant aspiration on Mandarin tones. *J. Int. Phon. Assoc.* 33, 165–181. doi: 10.1017/S0025100303001270

Xu, Y. (1997). Contextual tonal variations in Mandarin. *J. Phone.* 25, 61–83. doi: 10.1006/jpho.1996.0034

Xu, Y., and Xu, A. (2021). Consonantal f0 perturbation in American English involves multiple mechanisms. *J. Int. Phon. Assoc.* 149, 2877–2895. doi: 10.1121/10.0004239

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Communication

ORIGINAL RESEARCH
published: 21 June 2022
doi: 10.3389/fcomm.2022.844862

Within-Speaker Perception and Production of Two Marginal Contrasts in Illinois English

Jennifer Zhang[1], *Lindsey Graham*[1], *Marissa Barlaz*[1] *and José Ignacio Hualde*[1,2]*

[1] *Department of Linguistics, University of Illinois at Urbana-Champaign, Urbana, IL, United States,* [2] *Department of Spanish and Portuguese, University of Illinois at Urbana-Champaign, Urbana, IL, United States*

The notion of marginal contrasts and other gradient relations challenges the classification of phones as either contrastive phonemes or allophones of the same phoneme. The existence of "fuzzy" or "intermediate" contrasts has implications for language acquisition and sound change. In this research, we examine production and perception of two marginal contrasts [ɑ-ɔ] ("*cot-caught*"), where two original phonemes are undergoing a merger, and [ʌi-aɪ] ("*writer-rider*"), where a single original phoneme has arguably split into two contrastive sounds, albeit in a limited manner. Participants born and raised in Illinois were asked to provide recordings of *cot-caught* and *writer-rider* pairs embedded in sentences, followed by the target word in isolation. They then completed ABX and two-alternative forced choice two-alternative forced choice (2FC) perception tasks with stimuli produced by two native speakers from the Chicagoland area. Results showed that the [ʌi-aɪ] contrast, which has been defined as marginal in other work, is actually currently more phonetically and phonologically stable than [ɑ-ɔ] for the group of speakers that we have tested, with a more robust link between production and perception. The *cot-caught* merger appears to have progressed further, compared to what had previously been documented in the region. Our results and analysis suggest different sound change trajectories for phonological mergers, regarding the coupling of production and perception, as compared with phonemic splits.

Keywords: marginal contrast, merger, split, Canadian raising, production, perception

OPEN ACCESS

Edited by:
Joseph V. Casillas,
Rutgers, The State University of New Jersey, United States

Reviewed by:
Margaret Renwick,
University of Georgia, United States
Marta Ramírez Martínez,
Wartburg College, United States
Ryan Bessett,
The University of Texas Rio Grande Valley, United States

****Correspondence:***
José Ignacio Hualde
jihualde@illinois.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *28 December 2022*
Accepted: *29 April 2022*
Published: *21 June 2022*

Citation:
Zhang J, Graham L, Barlaz M and Hualde JI (2022) Within-Speaker Perception and Production of Two Marginal Contrasts in Illinois English. Front. Commun. 7:844862. doi: 10.3389/fcomm.2022.844862

INTRODUCTION

The words in a language are commonly analyzed in terms of unique phonological units, which by themselves are meaningless but combine according to the constraints of the language to bring about meaning (Hockett, 1958, 1960). This system of categorizing sounds assumes a specific set of phones for each language, with phones falling into one of two categories: *phoneme* (contrastive) or *allophone* (non-contrastive). Traditionally, two sounds are considered to be contrastive if, in at least one phonological environment, the choice of phone may result in lexical minimal pairs; the choice of phone cannot be predicted from the environment alone. Conversely, if the choice between two sounds *can* be predicted from their phonological environment, then the two sounds are allophones. Many phonological processes appear to ignore non-contrastive features, and contrast-based theories hold that the only features that can be phonologically active are those that serve to distinguish and contrast members of the underlying phonemic inventory (see Kiparsky, 1985; Hall, 2007; Dresher, 2009).

However, numerous researchers have pointed out the existence of distinctions between phones which cannot be easily categorized as either phonemic or allophonic (e.g., Goldsmith, 1995; Ladd, 2006, 2014; Nadeu and Renwick, 2016). Hall (2013) offers a comprehensive overview of these intermediate phonological relationships and provides a typology illustrating the many different ways in which contrasts can be marginal. In the literature, such relationships have previously been referred to as *semi-phonemic* (e.g., Bloomfield, 1939; Crowley, 1998), *quasi-phonemic* (e.g., Scobbie et al., 1999; Hualde, 2005), *weak contrast* (e.g., Hume and Johnson, 2001; Walker, 2005; Martin and Peperkamp, 2011), *partial contrast* (e.g., Hume and Johnson, 2003; Chitoran and Hualde, 2007; Kager, 2008), *gradient phonemicity* (e.g., Boulenger et al., 2011; Ferragne et al., 2011), and *marginal contrast* or *marginal phoneme* (e.g., Vennemann, 1971; Kiparsky, 2003; Edwards and Beckman, 2008; see also Hall, 2013; Renwick et al., 2016). Even in cases of phonological neutralization, where a contrast that is neutralized in a specific environment is still considered to be present elsewhere, some researchers have interpreted neutralization as an example of a "partial contrast," intermediate between full contrast and full allophony (Hume and Johnson, 2003; Kager, 2008). There are also cases where the distribution of a contrast in the lexicon may not be as reliably or consistently employed as expected, although the sounds themselves may be clearly distinct phonetically (Renwick and Ladd, 2016).

The notion of marginal contrasts and other gradient relationships challenges the division of phones into strict phonemic categories. The existence of marginal contrasts has implications for models of speech perception and language acquisition (both first and additional language) that rely on learner identification of contrastive phonological units and also has implications for sound change, in that speakers can acquire a distinction that is not necessarily utilized to identify words in speech.

Additionally, a speaker's ability to perceive marginal contrasts may not be directly correlated to their ability to produce that contrast and vice versa. Studies of sound changes in progress have shown that perception and production often do not proceed symmetrically, with changes occurring earlier in perception than in production (Di Paolo and Faber, 1990; Herold, 1990; Harrington et al., 2012; Kleber et al., 2012; Kuang and Cui, 2018), although some evidence has been found for a production lead when the relevant cues for production and perception are misaligned (e.g., Coetzee et al., 2018). Listeners may also still be able to perceive a contrast that they no longer produce (Labov, 1994; Hay et al., 2013; Coetzee et al., 2018; Pinget et al., 2020). Differences in perception and production in the actuation of a sound change may misalign, as perception and production may in fact be based on different targets or exemplars (Garrett and Johnson, 2013).

Speaker intuitions can also be a valuable resource for examining metalinguistic awareness of marginal contrasts. Previous research with Catalan (Nadeu and Renwick, 2016; Renwick and Nadeu, 2018) and Italian speakers (Renwick and Ladd, 2016), both populations with marginal mid vowel contrasts and the commonly used metalinguistic language to describe said contrast ("closed" and "open" mid vowels), has found speakers to be relatively accurate judges of their own productions. However, the prevalence of mismatches between production and speaker intuition involving members of a mid vowel contrast pair, relative to mismatches between pairs of mid vowels and corner vowels [i, a, u], separate the marginal mid vowel contrast on some dimension of phonological closeness.

In this article, we are concerned with the interaction between perception and production in two cases of marginal contrast in Illinois American English: [ɑ-ɔ], as in *cot* vs. *caught* (Experiment 1), and [ʌi-aɪ] as in *writer* vs. *rider* (Experiment 2). These two cases of marginal contrast differ in their diachronic provenance. The former represents an ongoing merger of two phonemes. The latter, instead, is a case of phonemic split, as it has arguably resulted from the phonological recategorization of allophones as (quasi-)contrastive units. This phenomenon is often known as *Canadian raising*. In both cases, we are interested in determining to what extent the degree to which the two categories are separated in individual speakers' productions determines their behavior in perception, as well as the relation of speakers' intuitions about contrastiveness to their own production and perception. Although there is a substantial literature on each of the two vowel phenomena we examine here, we are not aware of previous research that has compared the production-perception link in both a merger in progress and a split in progress for the same group of speakers.

EXPERIMENT 1: A MERGER IN PROGRESS: PRODUCTION AND PERCEPTION OF [ɑ-ɔ] ("*COT-CAUGHT*")

Background and Research Question

One example of a contrast that could be considered marginal in some varieties of present-day American English is the [ɑ-ɔ] ("*cot-caught*") low back vowel pair. The merger of these two phonemes was first attested in the US in the 1930s (Kurath, 1939) in parts of western Pennsylvania and eastern New England. Labov et al. (2006) later documented the distribution of the *cot-caught* merger, showing that the merger was highly advanced or completed in western Pennsylvania, and progressing in eastern New England and the western half of the United States. In contrast, the Inland North, the Mid-Atlantic, and the South were identified as regions that showed resistance to the low back merger. When the data for the *Atlas of North American English* (Labov et al., 2006) were collected, results of minimal pair perception tests for speakers in the Inland North (including Chicagoland, part of the area under study) showed no trace of a merger, with participants universally responding that the presented minimal pairs were different from one another. The maintenance of the /ɑ/ vs. /ɔ/ contrast was attributed to the fronting of /ɑ/, part of the Northern Cities Chain Shift (Labov, 1994; Clopper et al., 2005), making it rather distinct from /ɔ/. The low back merger was also found to be most advanced in syllables closed by nasal consonants and most conservative before velar /k/.

TABLE 1 | Contrastive contexts in US varieties without merger of /ɑ/ and /ɔ/.

Following context	Vowel	
	/ɑ/	**/ɔ/**
t	bot, cot, knot, not, rot, clot, dot, jot, lot, shot	bought, caught, naught, wrought
d	cod, nod, odd, sod, pod, god, mod, rod	cawed, gnawed, awed, sawed, pawed, broad, clawed, flawed
k	chock, hock, stock, wok, dock, lock, rock	chalk, hawk, stalk, walk, talk
n	don, con, Jon, Ron, swan	dawn, brawn, lawn, pawn
l	collar, doll, dollar	caller, ball, call, crawl, haul, mall, shall, stall
θ	goth	cloth, moth

TABLE 2 | Contexts of phonological neutralization in varieties without complete merger.

Following context	Vowel	
	/ɑ/	**/ɔ/**
p	bop, cop, drop, pop, top	–
b	blob, cob, job, knob, lob, mob, sob, swab, rob	–
m	com, mom, prom	–
g	–	blog, dog, fog, flog, hog, log
ŋ	–	long, song, wrong
f	–	loft, off, scoff, soft
s	–	boss, loss, moss, sauce, toss
0	–	law, saw

Even in varieties where /ɑ/ and /ɔ/ are clearly contrastive phonemes, the distribution of the two phones is not entirely free, and the presence or lack of a contrast is sometimes predictable from context. As Labov et al. (2006: 57) explain, historically, /ɑ/ descends from Middle English short /o/, with the addition of some /o/ words directly borrowed from French and some words where /a/ was rounded after /w/ (*watch, want, wander*, etc.). The resulting phone occurs before all but two consonants, /v/ and /ʒ/, in American English. In contrast, /ɔ/ has a more limited distribution. This vowel is, for the most part, a direct continuation of the Middle English diphthong /aw/ (which had a number of Old English and Old French sources). In addition, a number of words that had Middle English /o/ have been transferred to the /ɔ/ class, e.g., *dog, long, loss* (before /g/, /ŋ/, and voiceless fricatives, but without affecting all lexical items with these following contexts). Presently, for some speakers in Illinois, a contrast is attested in the phonological contexts shown in **Table 1**, where both phones occur (see also Labov et al., 2006: 57).

In the contexts in **Table 2**, on the other hand, the contrast appears to have been neutralized for all speakers in the Midwest dialect that we explore here. Labov et al. (2006: 57), describe a slightly different distribution, including a possible contrast before /g/, as in *log* vs. *dog*, that does not seem to exist in the geographical area under study. These authors do in fact report variation before /g/.

Even for contexts where a robust contrast has been reported, e.g., before /l/, the realization of the contrast may be less certain given the lack of representative words or given gaps in the lexicon. The merger between the two vowels appears to proceed in gradient fashion, occurring first before nasal consonants (Labov et al., 2006).

Based on informal observation, we suspect that the merger is currently more advanced in our target population (young speakers from northern and central Illinois) than some decades ago. We expect to find three or even four types of speakers: (a) speakers with a clear contrast in production and perception, (b) speakers who have merged the two phonemes in production and do not perceive them as different vowels, (c) as in other cases of mergers in progress, we also expect to find some speakers who have a marginal contrast in production, but cannot reliably identify or discriminate the two historical phonemes; that is, a merger in perception may precede a merger in production (Labov, 1994, 2011). Finally, some recent research indicates that in some mergers in progress there are listeners who can still perceive a contrast that they no longer produce (Hay et al., 2013; Coetzee et al., 2018; Pinget et al., 2020); thus, we may also expect to find a group (d) of speakers who do not produce the contrast but can reliably perceive it. Depending on the types of speakers found, these groups may help elucidate potentially differing patterns regarding the progression of the *cot-caught* merger.

Methods

Participants

Thirty-six participants were recruited among the undergraduate student population at an Illinois university to participate in this study. Of these participants, 11 did not complete all tasks and were excluded from analysis. Since the focus of this study is variation within northern and central Illinois, an additional 5 participants were excluded as they reported being born in a different country or state. The remaining 20 participants (14 females, 6 males) all reported being born and raised in Illinois. Of these, 14 are from Chicago and its suburbs, 4 from Central Illinois, and 2 from Illinois near the St. Louis area. We decided not to exclude the two St. Louis-area speakers[1] because this area has been shown to participate in some of the vowel changes that are found in Chicago, such as the ones under study (Labov, 2007). For the place where each participant was raised, see **Figure 1** (in Section Discussion). Their ages ranged from 18 to 25. Participants were volunteers or received extra credit for their participation in an undergraduate linguistics course. These subjects participated in one production task and two perception tasks.

[1] All speakers in this study, due to their status as students at this university, were immersed in an environment with substantial input from Chicago area speakers. Illinois residents have comprised over 70% of the incoming student population in recent years, with over 70% of the Illinois population (21,936 out of 30,347 students in 2021) representing the Chicago metropolitan area. However, we acknowledge the heterogeneity of our limited sample size.

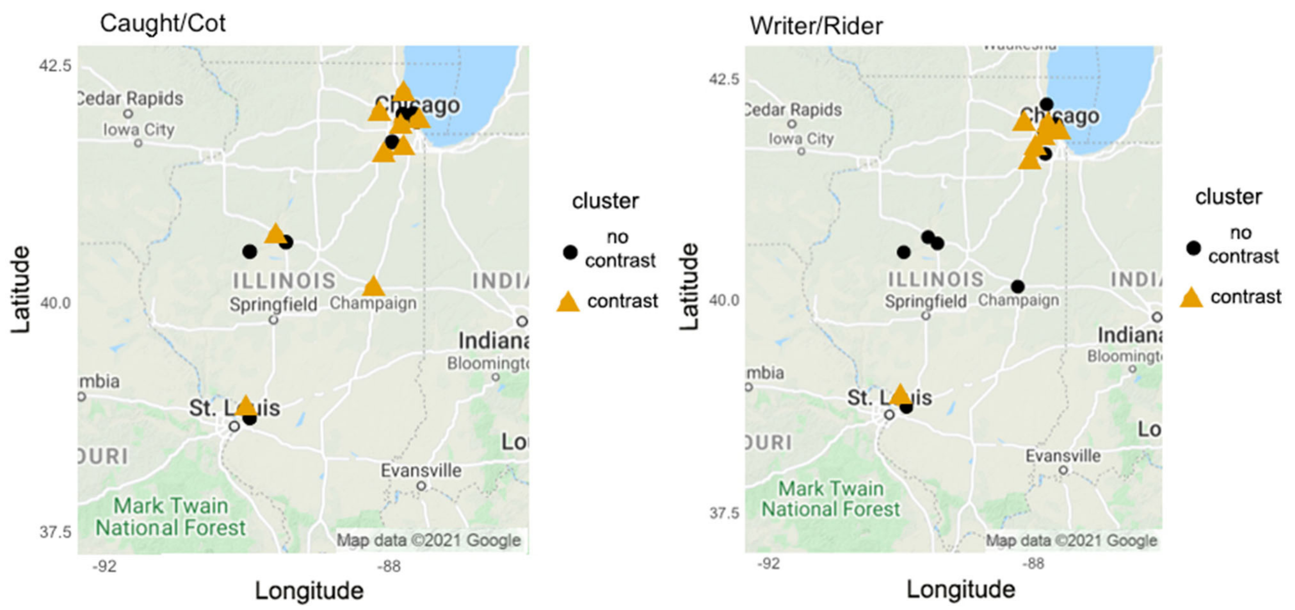

FIGURE 1 | Participant locations by production cluster for cot-caught (left) and writer-rider (right). Participant locations by production cluster (black circle: no contrast [low Pillai score]; orange triangle: contrast [high Pillai score]; Google, n.d.).

Stimuli for Production Study

For the production task, the goal was to create balanced lists of 20 pairs for each contrast under study. For the *cot-caught* contrast, the stimuli consisted of 13 monosyllabic minimal pairs with an alveolar coda (e.g., *caught* vs. *cot*), 3 monosyllabic near-minimal pairs (e.g., *laud* vs. *lot*), and 4 monosyllabic non-minimal pairs to complete the set of 20. Stimuli for our Experiment 2 (20 pairs) were also presented together. Filler items consisted of 10 pairs distinguished by their codas (e.g., *bet* vs. *bed*) and 10 homophone pairs (e.g., *flower* vs. *flour*). This resulted in 60 total pairs for a total of 120 productions. The stimuli for production are shown in **Table 3**.

Stimuli for Perception Study

For our perception study, our goal was to create balanced lists of 10 minimal pairs. The 10 minimal pairs (20 words) for the *cot-caught* contrast were all monosyllabic with coda consonants /t, d, n, k/. The stimuli used for the perception tasks are also shown in **Table 3**. Fillers in the perception tasks included 10 minimal pairs (20 words) distinguished by their codas (e.g., *mat* vs. *mad*, *mate* vs. *made*) and 4 homophone pairs (8 words) (e.g., *metal* vs. *medal*) as distractors. Tokens created for our Experiment 2 (9 minimal pairs, 18 words) on Canadian raising (see Experiment 2) also were presented together. Each word was presented two times, resulting in a total of 132 tokens.

The stimuli were produced by two native speakers, one female (Speaker F) and one male (Speaker M) from the Chicagoland area. These two model speakers were recruited to produce the stimuli because they reported producing a contrast between both *cot-caught* words and *writer-rider* words, and they both grew up in Illinois, like the participants in our experiments. The stimuli were recorded in a soundproof booth, using a Marantz PDM 750 solid state recorder and an AKG C5C20 head-mounted microphone at a sampling rate of 44.1 kHz. Based on formant values, all target stimuli included in the perception task showed a difference in vowel quality between *cot*-words and *caught*-words, although this difference was not always of the same magnitude or produced in the same manner. **Figure 2** shows average time-normalized formant trajectories for the stimuli produced by each of the two model speakers. Note that *cot*-words have higher values for both formants than *caught*-words, indicating a lower and less retracted articulation for [ɑ] than for [ɔ].

Procedure

Participants first completed a background questionnaire to provide demographic information and confirm that they had been born and raised in Illinois.

Production

For the first experimental task, participants were asked to provide recordings of 120 target words (20 *caught-cot* pairs, 20 *writer-rider* pairs, and 20 filler pairs). Because of the COVID-19 pandemic situation, conducting the experiment in a phonetics laboratory was not feasible at that time. Instead, participants were asked to record themselves in a quiet room, using their phones[2] or laptops. The recording material was presented *via* PowerPoint slides, and a copy of the PowerPoint slides was shared with each

[2] In a recent study, Freeman and De Decker (2021) report that recording with Apple devices may result in possibly deviant formant frequencies, which may affect the phonetic analysis of back and lower vowels in particular. We would like to add this caveat, since many of our participants submitted recordings from Apple devices.

TABLE 3 | Stimuli used in production and perception tasks.

Production				
Experiment 1: *cot-caught* stimuli		**Experiment 2: *writer-rider* stimuli**		
ɑ	**ɔ**	**ʌi**	**aɪ**	**Filler**
Tot	Taught	Writer	Rider	Apple
Not	Naught	Writing	Riding	Bade
Cot	Caught	Biter	Bider	Bait
Bot	Bought	Biting	Biding	Banana
Rot	Wrought	Cited	Sided	Bat
Sot	Sought	Sighting	Siding	Bed
Pod	Pawed	Whiter	Wider	Bet
Cod	Cawed	Light	Lied	Dear
Nod	Gnawed	Bite	Bide	Deer
Sod	Sawed	Bright	Bride	Died
Mod	Mawed	Ice	Eyes	Flour
Don	Dawn	Rice	Rise	Flower
Odd	Awed	Cite	Side	Kiwi
Lot	Laud	White	Wide	Knew
Blot	Brought	Kite	Buys	Lab
Rod	Broad	Nice	Hide	Lap
Pot	Pawn	Night	Ride	Led
Con	Lawn	Tight	Slide	Lessen
Dot	Flawed	Twice	Tide	Lesson
Clot	Brawn	Vice	Wise	Let
				Mad
				Made
				Mango
				Mat
				Mate
				Med
				Medal
				Met
				Metal
				Missed
				Mist
				New
				Pat
				Pedal
				Petal
				Ring
				Road
				Rowed
				Tied
				Wade
				Weak
				Wed
				Weighed
				Wet
				Wring

Perception					
Experiment 1: *cot-caught* stimuli		**Experiment 2: *writer-rider* stimuli**			
ɑ	**ɔ**	**ʌi**	**aɪ**	**Filler**	
Tot	Taught	Writer	Rider	Bade	Bait
Cot	Caught	Writing	Riding	Bat	Pat
Bot	Bought	Biter	Bider	Bed	Bet
Pod	Pawed	Biting	Biding	Died	Tied
Cod	Cawed	Cited	Sided	Lab	Lap
Nod	Gnawed	Sighting	Siding	Led	Let
Don	Dawn	Whiter	Wider	Mad	Mat
Tok	Talk	Insighter	Insider	Made	Mate
Wok	Walk	Citer	Sider	Met	Mat
Stock	Stalk			Wed	Wet
				Lessen	Lesson
				Medal	Metal
				Missed	Mist
				Pedal	Petal

participant. Each target word was embedded in a sentence (e.g., "The word ____ in English [means/refers to/is…]") which was presented on one slide, followed by the same target word in isolation on the next slide. The tokens were presented in pseudo-random order such that each presented token was not followed by a member of a potential minimal pair. The recording session was divided into four blocks, and participants were asked to submit their recordings at the completion of each block. The first production of each block consisted of a filler sentence and word.

ABX

Following production, participants completed the first perception task: an ABX task administered online *via* Qualtrics. The stimuli consisted of target words in isolation, with an interstimulus interval of 500 ms. Participants heard a presented ABX series and were asked to select whether X was the same word as the first word they heard (A) or the second (B) by clicking on the number "1" or "2" on the screen. Upon making a selection, the next ABX series was automatically presented. Tokens produced by Speaker F were used for A and B of the ABX task, and tokens produced by Speaker M were used for X. The use of two speaker voices requires some level of abstraction by the participant, particularly as our speakers differ in sex. The stimuli were presented in pseudo-random order such that no two types of the same category (*cot-caught*, *writer-rider*, coda-distinguished filler, homophone filler) were presented sequentially; the presentation of one category of stimuli was always followed by a different category, e.g., *rider-writer* (*rider-writer* pair) followed by *mist-missed* (filler-pair) followed by *cawed-cod* (*cot-caught* pair), followed by *mate-made* (coda-distinguished pair). There was a total of 132 items presented, of which 40 tokens (20 pairs) were representative of the *cot-caught* contrast. Tokens created for our Experiment 2 on Canadian raising (see Experiment 2) also were presented together.

Two-Alternative Forced Choice Identification (2FC)

Participants then completed the second perception task: a two-alternative forced choice (2FC) word identification task administered online *via* Qualtrics. The stimuli presented were the same as in the ABX task, but participants were instead asked to identify an auditorily presented word by clicking on one of

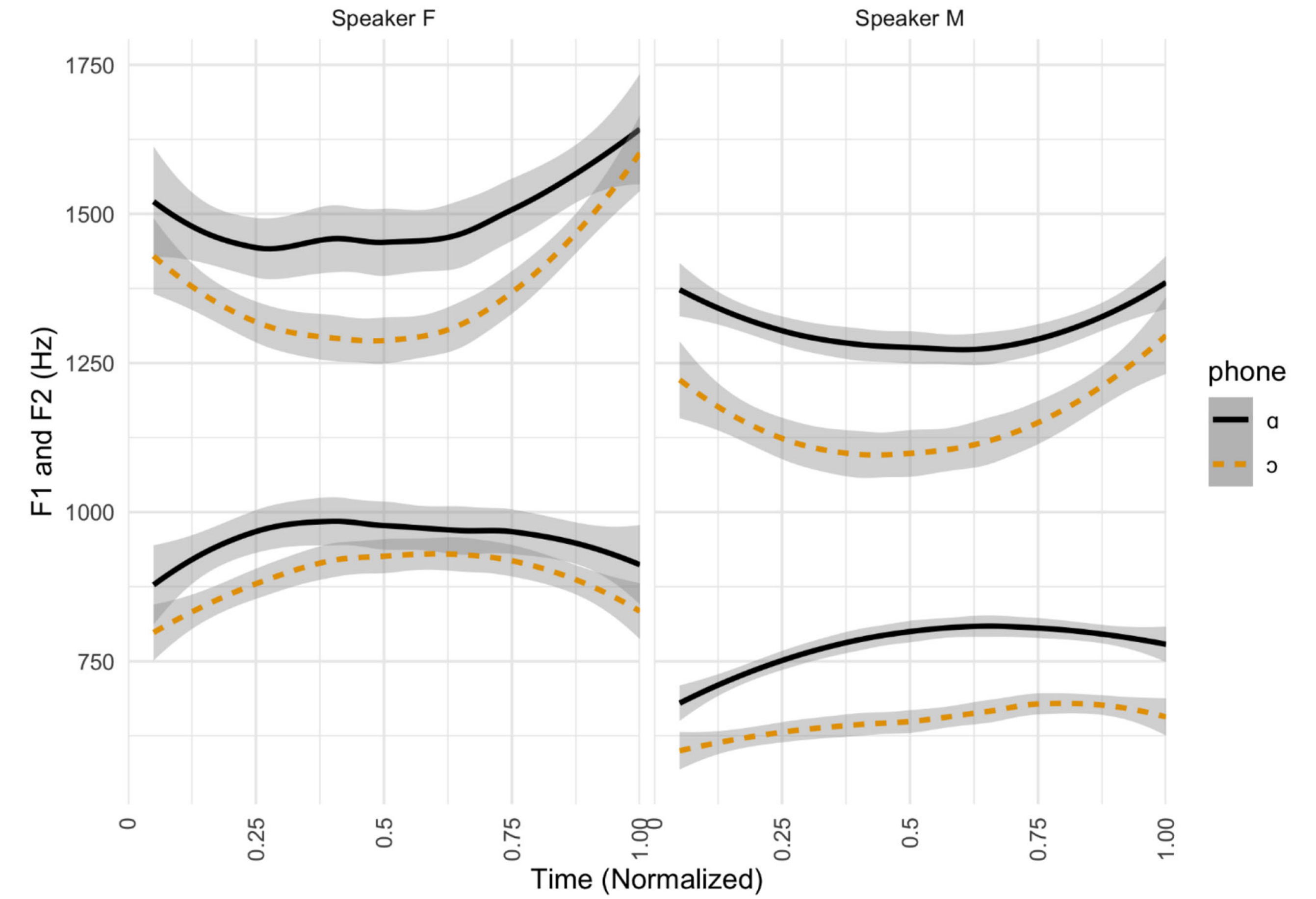

FIGURE 2 | Formant contours for cot-caught stimuli. Time-normalized formant contours (F1 and F2) in Hz for the two speakers providing perception stimuli, separated by target phone. Formants for *cot*-words are shown in solid black lines and formants for *caught*-words in dashed orange lines.

two words presented on the screen. For example, they would hear the word *cot* and either click on the presented text <cot> or <caught>. Upon making a selection, the task automatically moved to the next 2FC item.

Exit Survey

Finally, participants completed an exit survey which probed their phonological intuitions of the contrast as spoken by themselves (e.g., "Do you think you pronounce *cot* and *caught* in the same way?"), their parents or guardians, and by their social circles. They were also asked to describe any differences in their pronunciation [A similar methodology was used in Renwick and Nadeu (2018)].

Acoustic Analysis of Production Data

The target words were first force-aligned with the Montreal Forced Aligner (McAuliffe et al., 2017), then corrected by hand in Praat (Boersma and Weenink, 2021). Waveform and spectrogram information was used for manual corrections. When the preceding consonant was an obstruent, the left boundary of the vowel was placed at the first zero uprising after the onset of glottal vibration, when formant structure was visible. When the preceding consonant was a fricative, the left vowel boundary was similarly placed at the first zero uprising when formant structure was visible, after the offset of high energy frication noise. When the preceding consonant was a sonorant, changes in formant structure and intensity were used to place segmental boundaries. The right boundary of the vowel was similarly determined by decreases in intensity and changes in formant structure. The vowels were manually assigned labels that would correspond to the presence, rather than merger, of a phonological contrast. Following segmentation, F1 and F2 values were automatically extracted at the 50% duration of the vowel.

Statistical Treatment

As a measure of distance between vowel distributions, we calculated Pillai scores for each vowel pair at the 50% duration of the vowel for each participant (Nycz and Hall-Lew, 2013; Jibson, 2021). A higher Pillai score, closer to 1, results from greater distance and less overlap between vowel pairs, indicating a stronger contrast. A lower Pillai score, closer to 0, results from greater overlap, indicating a weaker contrast or no contrast at all. The Pillai scores were then submitted to a k-means cluster analysis using the function *kmeans* from the *stats* package in R (R Core Team, 2019). The analysis was run for 2-4 groups.

We decided to use cluster analysis as opposed to determining a threshold for the classification of participants as having one phoneme or two (as in, e.g., Labov et al., 2006), precisely because we want to allow for the possibility of having intermediate situations between merger or not merger and split or not split.

Since we are interested in determining the relation between production and perception, we ran correlations (*cor.test* from *stats* in R) between Pillai scores and perception accuracy results. Linear mixed effects regressions were run on accuracy rates and formant values with the function *lmer* in the package *lme4* (Bates et al., 2015) and *p*-values were obtained with the *emmeans* package (Lenth, 2022). In addition, we considered the extent of participants' phonological intuitions concerning the existence of a contrast in their speech or lack thereof, and how this corresponded to their performance in our two perception tasks.

Results

Production Results for Experiment 1

Figure 3 displays average F1 and F2 values over normalized time for words belonging to the traditional /ɑ/ class (*cot*) and words belonging to the /ɔ/ class (*caught*). Each participant is shown on their own plot, and the plots are organized from lowest Pillai score (Participant 003) to highest (Participant 026). Speakers ranged from no discernible contrast in Participant 003 (Pillai score = 0.02) to a very clear contrast in Participant 026 (Pillai score = 0.89).

Figure 4 shows the vowel plots for Participants 003 (lowest Pillai score = 0.02) and 026 (highest Pillai score = 0.89), showing an example of the difference in degree of overlap for speakers with the merger and with the contrast. The vowel plot for Participant 003 shows a clear overlap between *cot*-type words and *caught*-type words. In comparison, the two types of vowels are clearly distinct for Participant 026.

Participants were clustered first based on production alone, using cluster analysis specified for 2 to 4 clusters. Based on the total within-cluster sum of squares, the best clustering resulted in 2 groups according to their productions, which we may think of as mergers and non-mergers. The merger group includes 13 participants with Pillai scores ranging from 0.02 to 0.30, and the non-merger group includes 7 participants with Pillai scores ranging from 0.35 to 0.89.

Perception Results for Experiment 1

Participants were also independently clustered based on average perception accuracy between ABX and 2FC. Participants did not fall into the same clusters for production as for perception, so the two clustering analyses were visually combined to show inconsistencies between mergers in perception and production. The resulting groups are shown in **Figure 5**, along with the correlation between production of [ɑ-ɔ] and each perception task.

As can be seen in **Figure 5**, participants fall into one of four possible groups according to Pillai-score-based clustering: (1) NO contrast in perception and NO contrast in production (black in **Figure 5**), (2) NO contrast in perception, YES contrast in production (orange), (3) YES contrast in perception, NO contrast in production (blue), and (4) YES contrast in both perception and production (green). For speakers with the contrast in both production and perception, formant values for [ɑ] and [ɔ] were significantly different for F1 ($p < 0.05$) and F2 ($p < 0.001$). Those without the contrast showed no significant differences among formant values. Perception accuracy rates between the NO/NO and YES/YES groups were significantly different for the 2FC task ($p < 0.05$) but not for ABX ($p = 0.23$).

Participants 002 and 006 are examples of speakers in group 2, who showed a contrast in their production (Pillai scores = 0.35 and 0.47), but their perception accuracy was around chance (40-65%). In the opposite direction, Participants 011 and 031 are examples of speakers in group 3, those who could perceive the contrast with accuracy rates ranging from 70 to 80%, but whose productions had low Pillai scores (0.04 and 0.29) and were therefore considered to be merged in production. Group 3 in particular was not specifically hypothesized to exist, based on the assumption that vowel mergers in perception tend to precede merger in production, yet it comprises 35% of our participants. These findings do align, however, with recent research regarding perception and production inconsistencies (e.g., Hay et al., 2013; Coetzee et al., 2018; Pinget et al., 2020).

Overall, the correlations between perception and production were very weak ($R^2 = 0.06$ between production and ABX, $R^2 = 0.09$ between production and 2FC). Perception accuracy for individuals clustered as non-mergers was 73% (averaged between ABX and 2FC), whereas average accuracy for those clustered as mergers was 69%; perception accuracy was thus similar regardless of their clustered status as mergers or non-mergers.

Exit Survey Results for Experiment 1

Based on their responses to the exit survey, participants who were clustered as non-mergers in both perception and production appeared to be metalinguistically aware of the contrast. For those who reported having a contrast ($n = 12$), 9 of them also showed the contrast in production. The other 3 participants who self-reported a distinction were instead clustered as showing a merger in their production. Fewer participants reported having merged productions ($n = 8$), and their productions were split between the merger group and the non-merger group. This data is visualized in the violin plots as seen in **Figure 6**, which show participants' self-reported distinctions compared with their Pillai scores at 50% vowel duration.

As for the correlation between participants' phonological intuitions and their performance in the perception tasks, participants who reported making a contrast between *cot* and *caught* outperformed participants who reported not making a distinction, as can be seen in **Figure 7**. This is an expected result. Accuracy was somewhat higher for both groups of participants in the forced-choice identification task.

Furthermore, participants who reported making a contrast in production also described differences in vowel pronunciation on the exit survey, with a number of participants transcribing [ɑ] in *cot* as <ah> and [ɔ] in *caught* as <aw> (e.g. "*cot* is more 'KAHt' *caught* is more 'CAWt'" and "In the word *cot*, the 'o' makes more of an 'ah' sound rather than the 'aw' sound found in *caught*"). Our results suggest that speakers who believe they have a contrast

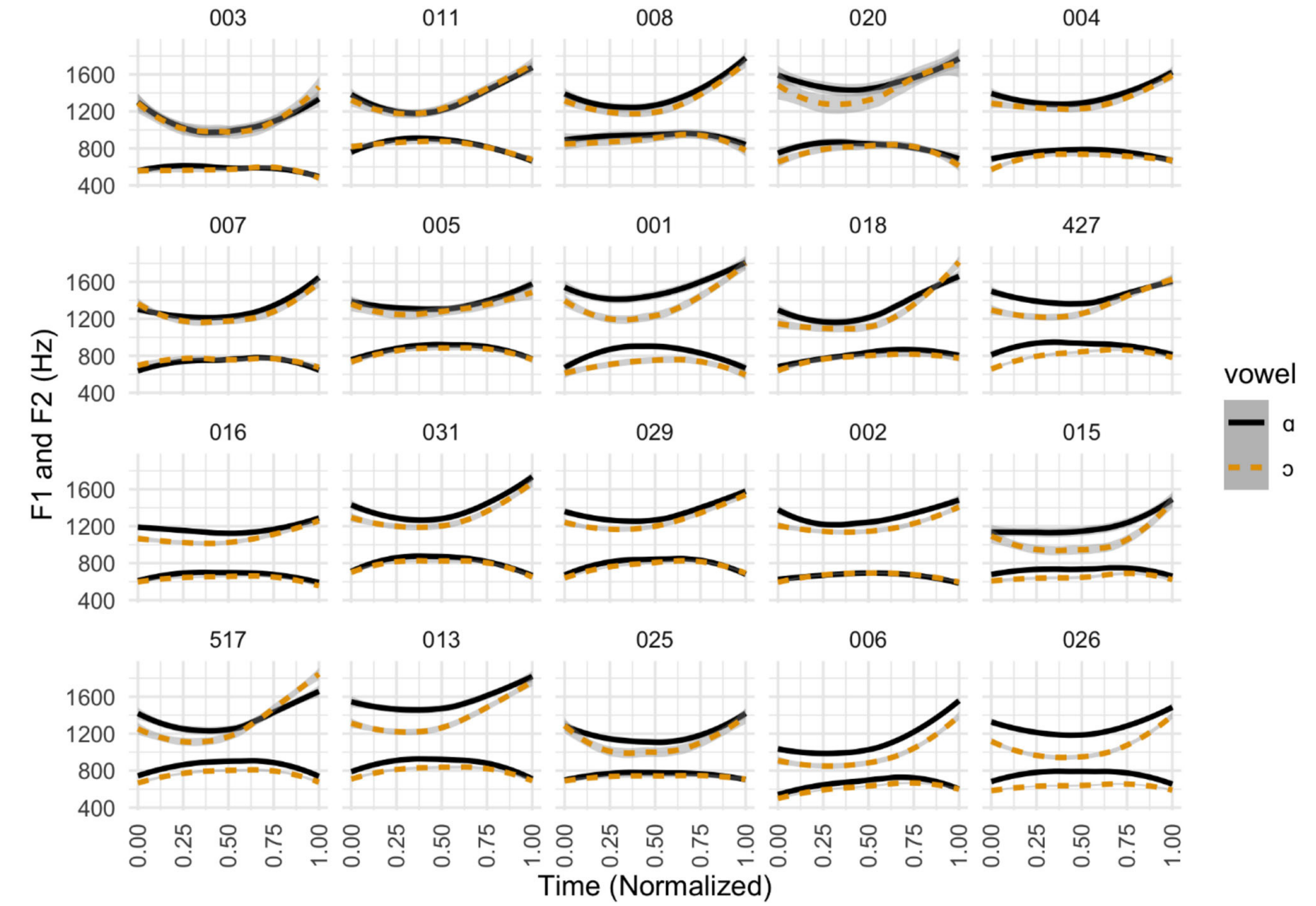

FIGURE 3 | Formant contours for cot-caught by participant. Individual time-normalized formant contours (F1 and F2) in Hz, separated by target phone. Participants are organized from lowest to highest Pillai score at 50% vowel duration.

do in fact produce it. However, their perceptual discrimination abilities, based on clustering, do not always mirror the contrast as found in production, again based on clustering. The resulting four groups from our clustering analyses also show different patterns of perception and production as they relate to a merger in progress. Out of our 20 speakers only 12 reported having a phonological contrast in their own speech, and less than half of them (3) fell within the non-merger clusters in both production (high-Pillai-score cluster) and perception (average ABX and 2FC accuracy). The historical phonemic contrast between /ɑ/ and /ɔ/ now thus has the status of a marginal or fuzzy contrast for this group of speakers, with most speakers showing variable behavior in perception and production that is inconsistent with the existence of a robust contrast between two phonemes.

EXPERIMENT 2: A MARGINAL SPLIT: CANADIAN RAISING

Background and Research Question

The other marginal contrast that we are concerned with is the [ʌi-aɪ] ("*writer-rider*") split, a phenomenon traditionally known as Canadian raising (Joos, 1942; Chambers, 1973, 1989, 2006). Historically, Canadian raising is an instance of an (incomplete) phonemic split, where a phonemic unit develops distinct allophones, a process opposite of the merger between [ɑ] and [ɔ]. The basic distribution of the Canadian raising diphthongs is that [ʌi] occurs before voiceless consonants (e.g., *write, sight*) and [aɪ] appears before voiced consonants and word-finally (e.g., *ride, buy*). However, this complementary distribution is rendered opaque by the neutralization of /t/ and /d/ as a flap before an unstressed vowel (Herd et al., 2010), leading to minimal pairs differentiated only by the quality of the diphthong, as in *writer* [ɹʌiɾɚ] vs. *rider* [ɹaɪɾɚ]. Before a flapped alveolar stop, phonological raising is purported to be triggered not by the phonetic realization of the stop, but rather by its underlying phonological specification (see also Mielke et al., 2003; Bermúdez-Otero, 2004; Idsardi, 2006; Pater, 2014). Additional instances of unconditional raising in monomorphic words have also been reported, e.g., *tiger, spider* (Vance, 1987; Dailey-O'Cain, 1997; Graham, 2019)[3]. Raising of the nucleus has

[3] It has been noticed [e.g., Hualde et al. (2021)] that in some monomorphemic words like *cider* and *spider*, the higher diphthong is sometimes found before the

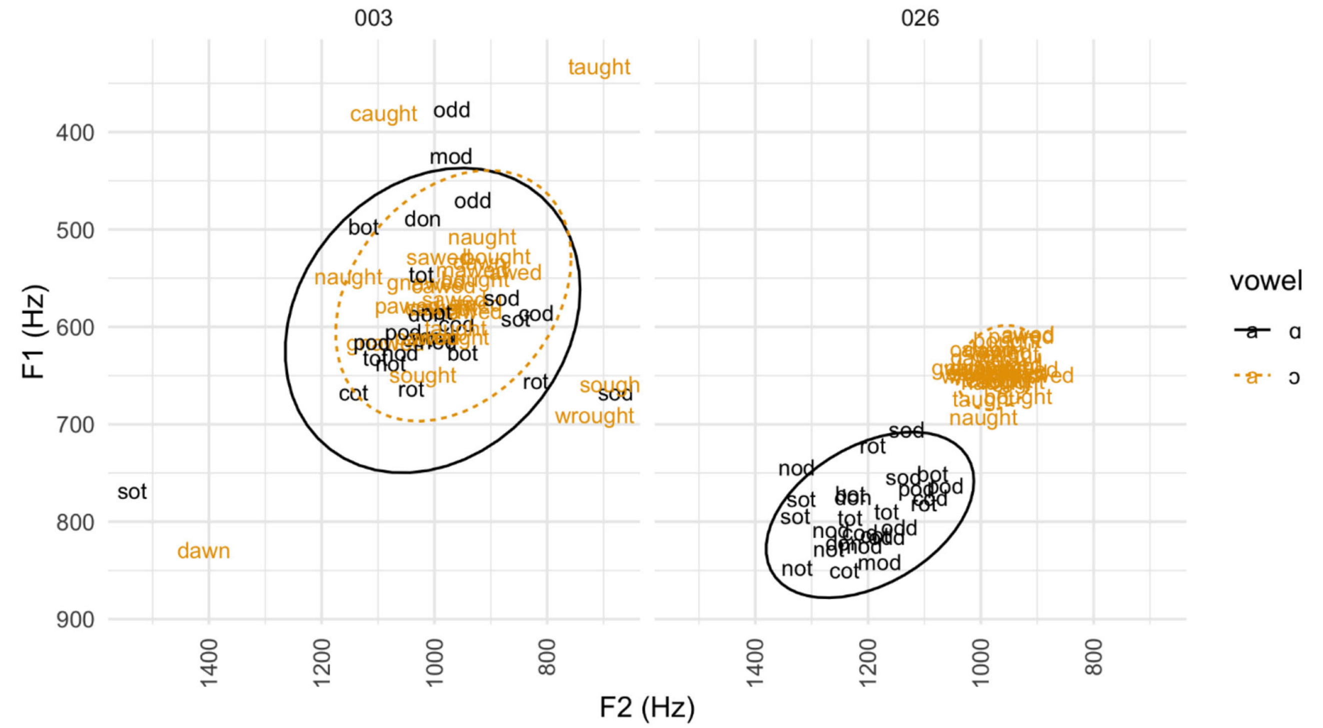

FIGURE 4 | Vowel plot for cot-caught productions with low vs. high Pillai scores. Vowel plot of individual tokens representative of the *cot-caught* contrast for Participant 003 (lowest Pillai score) and Participant 026 (highest Pillai score). *Cot*-words are graphed in black and *caught*-words in orange. Ovals indicate the 95% confidence interval (2 standard deviations) for each vowel.

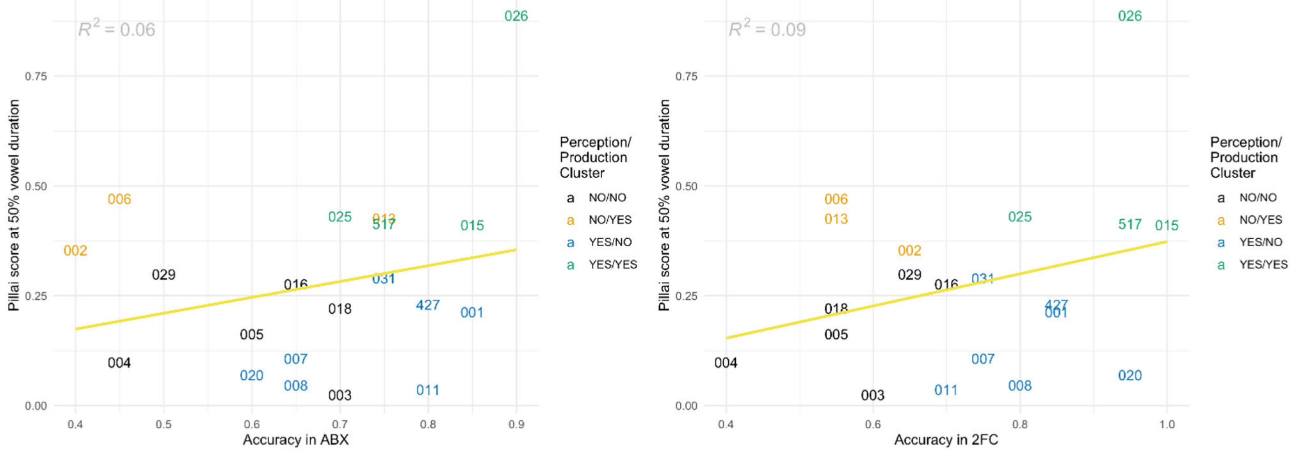

FIGURE 5 | Correlations between Pillai scores at 50% vowel duration and ABX perception accuracy (left) or 2FC perception accuracy (right) for cot-caught. Participants are clustered based on perception and production independently, with the two clustering analyses visually combined to show 4 groups, where the first NO or YES of each label signifies the contrast in perception and the second NO or YES signifies a contrast in production. NO/NO (participant clustered in the group with no contrast based on Pillai scores for both production and perception) is shown in black. NO/YES (participant in no-contrast group in perception, but in contrast group in production) is in orange. YES/NO (participant in contrast group in perception, but in no-contrast group in production) is in blue and YES/YES (contrast in both production and perception) is in green.

been also reported before syllabic /r/, as in *fire* [fʌiɚ]; but in this phonological context as well, it is possible to have minimal pairs, as in *hire* [hʌiɚ] vs. *higher* [haɪɚ], where the second member of the pair retains the quality of the word-final diphthong in the underived form *high*.

tap, contrary to what would be expected from the spelling. This shows a tendency for this diphthong to spread outside the context where it emerged (cf. also *tiger, fire*, without a following coronal stop, etc.). In our stimuli the diphthong is always followed by a morpheme-final coronal stop.

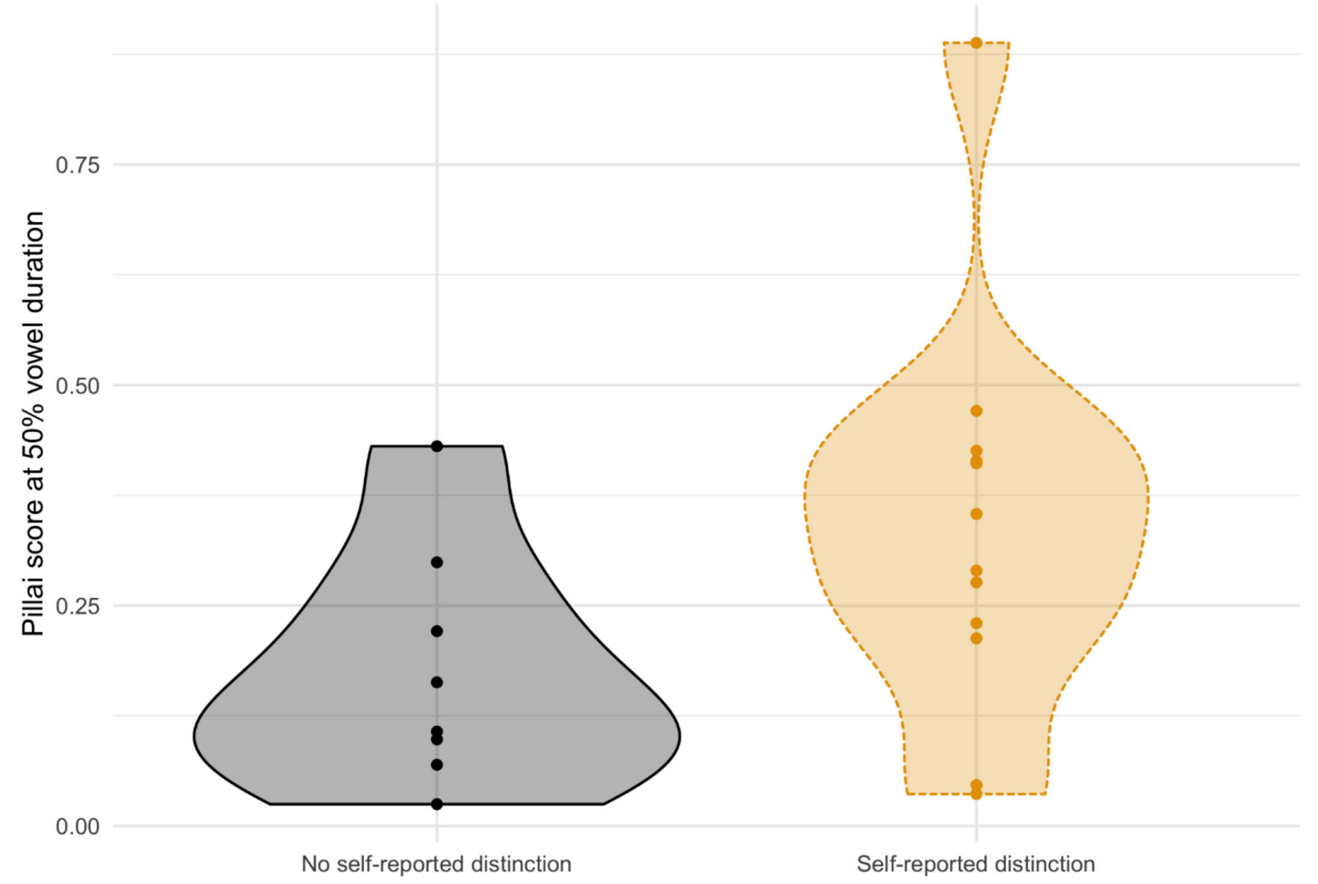

FIGURE 6 | Pillai scores by self-reported contrast for cot-caught. Pillai scores for *caught-cot* by participant, separated by self-reported distinction of the contrast.

Despite its name, the [ʌi-aɪ] split has been documented in various regions of the US. In fact, for East Virginia, the phenomenon was described as early as the first decades of the twentieth century (Shewmake, 1925, 1943). The split has since been documented in areas such as Minnesota (Vance, 1987), Rochester, New York (Vance, 1987), Michigan (Milroy, 1996; Dailey-O'Cain, 1997), Philadelphia, Pennsylvania (Fruehwald, 2008, 2016), and Fort Wayne, Indiana (Davis et al., 2020). Progression of the split is at perhaps a less advanced stage in the Fort Wayne area, with Davis and colleagues finding evidence for an incipient phase of phonetic conditioning. Whereas in Canadian English both diphthongs /ai/ and /au/ undergo raising of the nucleus before a voiceless consonant (e.g., *about* vs *loud*), in many US varieties only /ai/ appears to be affected. Davis et al. (2020) have proposed the term "American raising" to refer to the phenomenon in varieties where the relevant effects are found for /ai/ but not for /au/.

Within northern and central Illinois, the area under study, the [ʌi-aɪ] contrast has been reported for the Chicago area. Kilbury (1983) provides minimal and near-minimal pairs from his own Chicago variety, noting that speakers in his area differed in their intuitions regarding the pronunciation of *writer-rider* pairs. This idea found recent support from Hualde et al. (2021), whose formant trajectory analysis showed notable interspeaker variation in degree of production of this contrast. Research on perception of the [ʌi-aɪ] contrast has also found that speakers of this variety also vary in their ability to differentiate minimal pairs in perception. The Chicago-area speakers tested in Hualde et al. (2017) were able to discriminate the sounds [ʌi-aɪ], but not at ceiling accuracy, unlike other contrasts tested in the same experiment, indicating that the [ʌi-aɪ] contrast is somewhat less robust. Though in Hualde et al. (2017) both production and perception were tested, the correlation between production and perception for individual speakers was not examined. Strickler (2019) analyzed the production and perception of speakers from Fort Wayne, Indiana, an area where Canadian raising appears to be spreading throughout the community. She found that speakers seemed unable to perceive a more advanced form of the split than the forms they produced.

Here we focus on the speech of young native English speakers from northern and central Illinois, a geographical area where both the [ɑ-ɔ] and the [ʌi-aɪ] contrasts appear to have different degrees of "robustness" for different speakers. The question that we wish to ask is how a speaker's ability to perceive marginal contrasts relates to the robustness of that contrast in their own production. We are also interested in learning

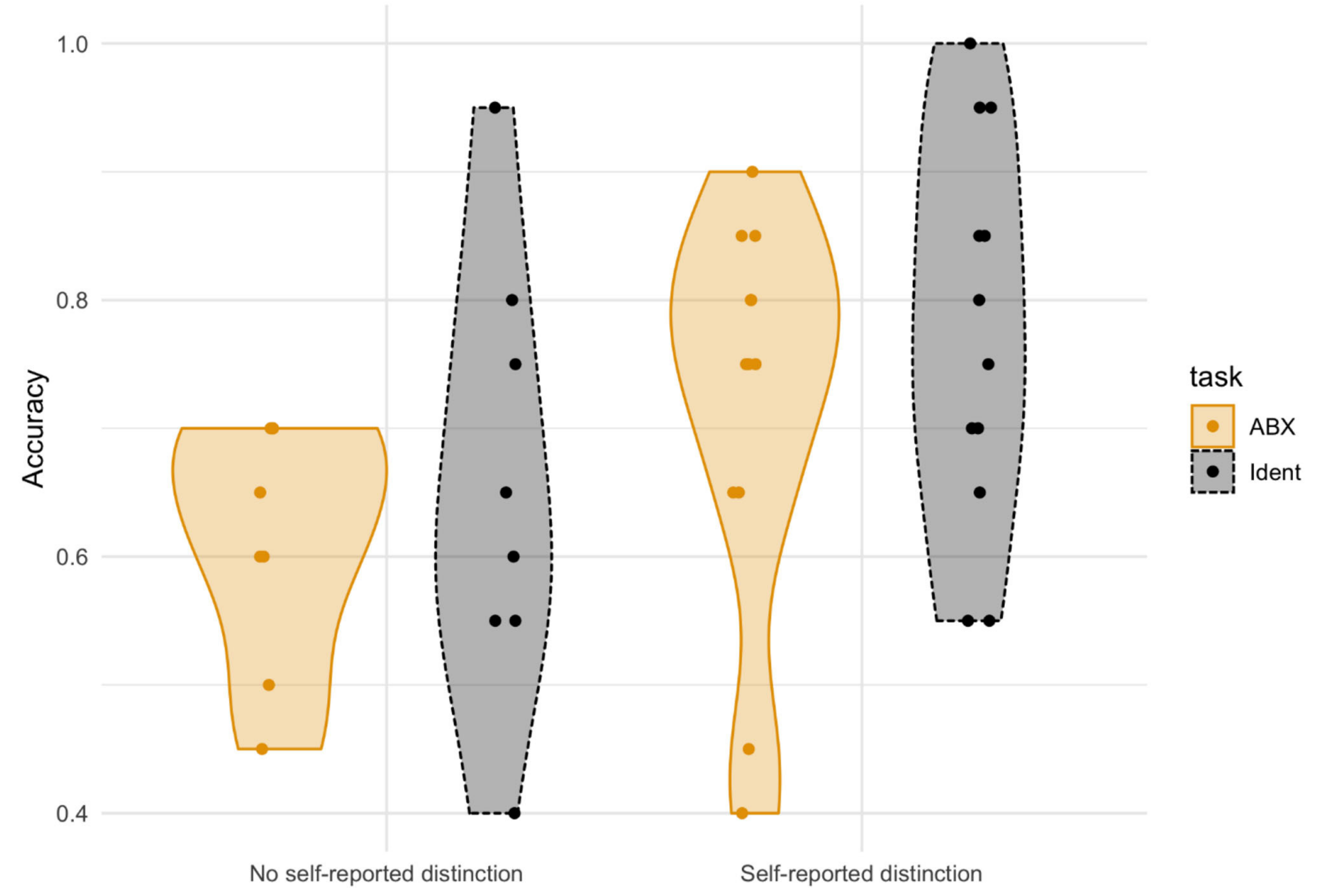

FIGURE 7 | Accuracy in perception by self-reported contrast for cot-caught. Accuracy in the two perception tasks (ABX and 2FC Identification), separated by self-reported distinction of *caught-cot*.

to what extent speakers' intuitions regarding phonological contrasts correlate with their own performance in perception and production.

Methods

The same speakers participated in this experiment and in Experiment 1, as both experiments were run together. The stimuli for the production task, presented together with the other stimuli described in Experiment 1, consisted of 7 bisyllabic minimal pairs, 7 near-minimal pairs (e.g., *bite* vs. *bide*), and 6 non-minimal pairs (12 monosyllabic words).

The stimuli for the two perception tasks were produced by the same two model speakers who produced stimuli for the *cot-caught* distinction. Due to the environment conditioning the *writer-rider* contrast, minimal pairs are infrequent, and a maximum possibility of 8 bisyllabic pairs and 1 trisyllabic pair were compiled for the perception task. Of the 132 total items presented, 36 tokens (18 pairs) were representative of the *writer-rider* contrast. The procedure was described in Experiment 1, and the full list of stimuli for production and perception are shown in **Table 3**.

Figure 8 shows time-normalized average formant values for Speaker F and Speaker M separately. These contours show that both speakers have clearly distinct productions of the contrast, with Speaker F differentiating the sounds in frontness (F2) early on and height (F1) toward the middle and end of the diphthong. Speaker M also has distinct vowels, but to a less drastic degree, and differentiates using F1 early on in production and F2 later.

The acoustic analysis was also performed as in Experiment 1, but rather than taking formant measurements at 50% of the duration of the vowel, these measurements were taken at 30 and 80% of the duration. The use of multiple timepoints for diphthong analysis captures differences in the nucleus and the offglide (Hillenbrand, 2013; Hualde et al., 2017). While some have proposed 20 and 80% as optimal timepoints, our data suggest that 30% is a preferred first measure. As noted by others, this timepoint is well within the nucleus but avoids coarticulatory effects (Berkson et al., 2017). Two Pillai scores were calculated for each speaker using these measurements, which were then used for clustering. As with Experiment 1, this process resulted in two groups which can be interpreted as speakers with the contrast in production and speakers without. A similar clustering analysis was also performed for the perception data.

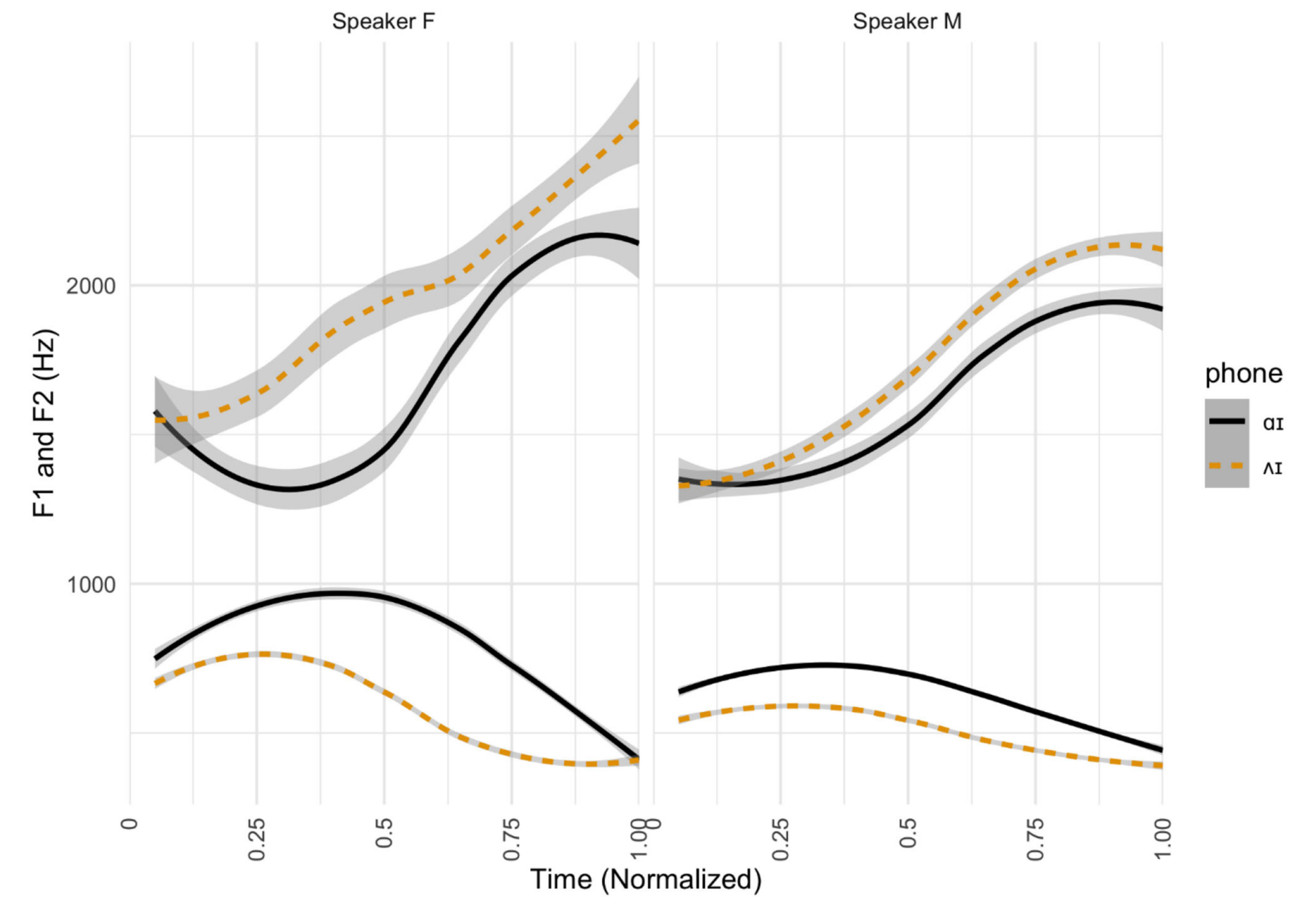

FIGURE 8 | Formant contours for writer-rider stimuli. Time-normalized formant contours (F1 and F2) in Hz for the two speakers providing perception stimuli, separated by target phone. Formants for *rider*-words are shown in black and formants for *writer*-words in orange.

Results

Production Results for Experiment 2

As in Experiment 1, we present the results of our production task before considering the perception results and the correlations between perception and production.

In **Figure 9**, we show average time-normalized F1 and F2 tracings for each participant. From the upper left to the bottom right, participants are organized by Pillai score (low to high). Pillai scores at 30 and 80% of the duration of the diphthong turned out to be highly correlated (R^2 = 0.90), so this figure orders plots based on the 30% Pillai scores alone, and the 80% scores are not shown.

As can be observed, some participants such as 025 (Pillai score = 0.93 at 30% and 0.83 at 80%) and 031 (Pillai score = 0.83 at 30% and 0.76 at 80%) have hardly any overlap in formant trajectories between *writer*-type words (in dashed orange) and *rider*-type words (solid black). Other speakers show near-total overlap, including 002 (Pillai score = 0.02 at 30% and 0.03 at 80%) and 004 (Pillai score = 0.13 at 30% and 0.07 at 80%).

The use of Pillai scores takes into account both distance between vowel clusters and overlap between them. **Figure 10** shows the vowel plots for Participants 013 and 025, showing an example of the drastic difference in degree of overlap for speakers without a contrast (013) and for those with the split (025).

The vowel plot for Participant 013 shows a clear overlap between *writer*-type words and *rider*-type words. The two types of vowels are clearly distinct for Participant 025, as there is notable distance between the production clouds and no overlap.

Perception Results for Experiment 2

Results showing the correlation between Pillai scores (at 30% and 80% vowel duration) and perception accuracy (for ABX and 2FC) are shown in **Figure 11**.

Participants were first clustered based on production alone, resulting in a no-contrast group of 9 participants with Pillai scores 0.01 to 0.42, and a split group of 11 participants with Pillai scores ranging from 0.43 to 0.93. Participants were also independently clustered based on average perception accuracy between ABX and 2FC. As in Experiment 1, we cross-classified participants considering their clustering in production and their clustering in perception, which would potentially yield up to four groups. However, in contrast with *cot-caught*, the combination

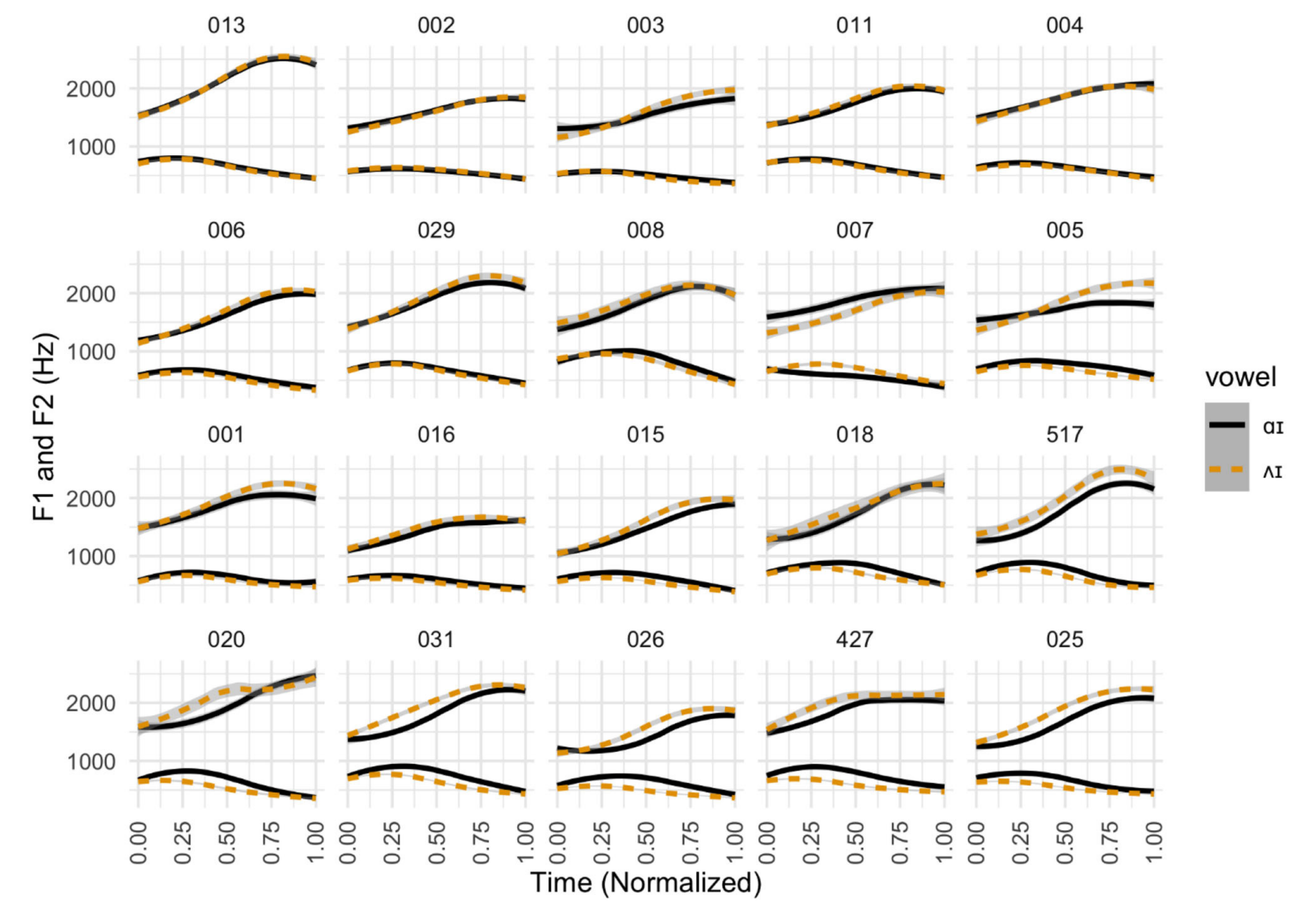

FIGURE 9 | Formant contours for writer-rider by participant. Individual time-normalized formant contours (F1 and F2) in Hz, separated by target phone. Speakers are organized from lowest to highest Pillai score at 30% duration of the vowel.

of both independent clustering procedures resulted in almost the same clusters in production and perception, with only one single participant for which production and perception-based clustering do not match. We therefore have a group of 9 participants with low accuracy in perception and low Pillai scores in production, consistent with lack of a phonological contrast (in black in **Figure 11**), a second group of 10 participants with high accuracy in perception and high Pillai scores in production (in blue) and a single participant with low accuracy in perception but a high Pillai score in production (in orange).

For speakers with the contrast in both production and perception, formant values for [ʌi] and [aɪ] were significantly different for F1 and F2 at 30% and at 80% ($p < 0.01$). Those without the contrast showed no significant differences among formant values for the two diphthongs. Perception accuracy rates between the NO/NO and YES/YES groups were significantly different for both the ABX ($p < 0.01$) and 2FC tasks ($p < 0.001$).

Our resulting clusters for *writer-rider* show that perception and production pattern very closely together, and that speakers with inconsistencies between perception and production are rather rare. Unlike our results for *cot-caught*, clustering analyses showed inconsistencies between perception and production for only one participant (Participant 016). We believe this to be an interesting result, perhaps pointing to a generalizable difference between mergers and splits.

Correlations between perception and production were moderate to strong, with the best correlation obtained when Pillai scores at 80% of the duration of the vowel were compared with results of the 2FC task ($R^2 = 0.71$). Overall, our results suggest that speakers with the *writer-rider* contrast in production will perform better in perception of that contrast. Mean perception accuracy rates by production support this: those classified as speakers who produced the contrast had an average accuracy of 85% while those who were classified as not producing a contrast were at 62%. This is much larger than the difference found for *cot-caught* in average perception accuracy between the two tasks.

Exit Survey Results for Experiment 2

Metalinguistic awareness of the *writer-rider* contrast in a participant's own speech appears to be mixed at first blush, but further analysis of the responses shows that this is due, for the most part, to the possibility of pronouncing /t/ and /d/ differently in the context of flapping or orthographical differences. Of those

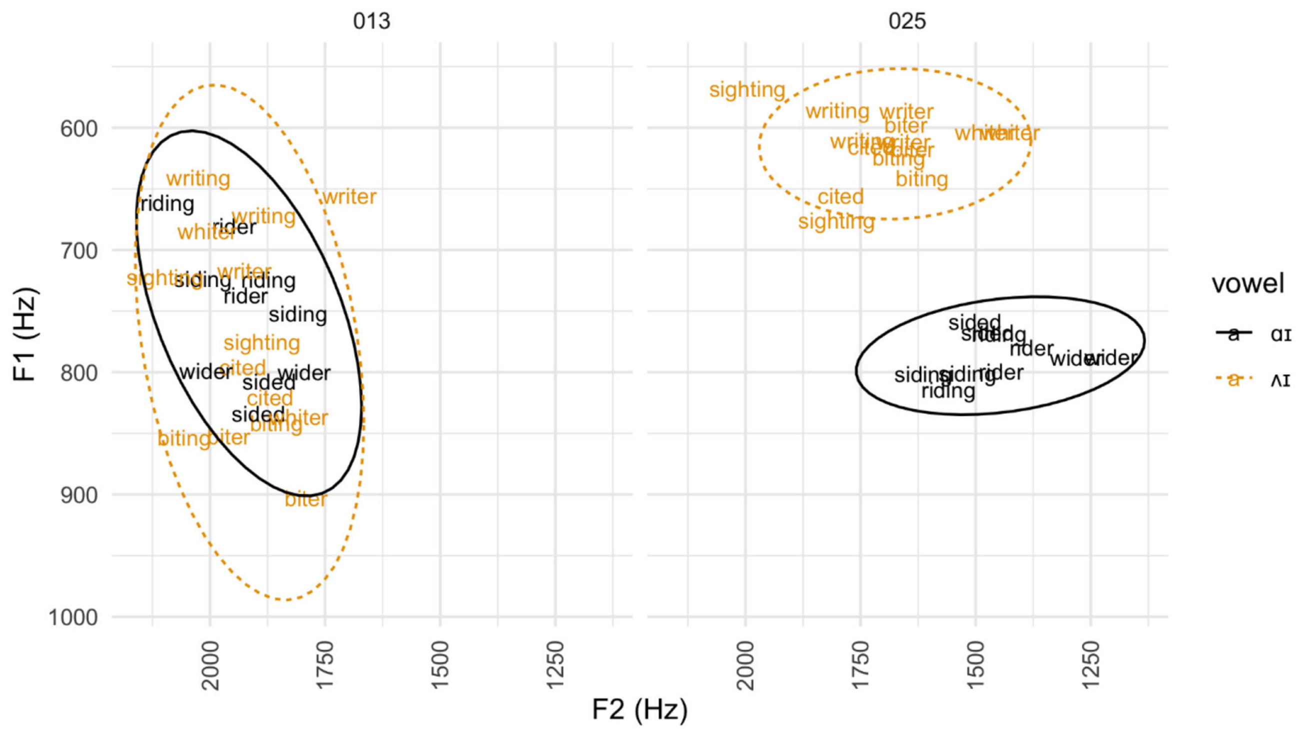

FIGURE 10 | Vowel plot for writer-rider productions with low vs. high Pillai scores. Individual tokens of words representative of the *writer-rider* contrast for Participant 013 (lowest Pillai score) and Participant 025 (highest Pillai score). Ovals indicate 95% confidence intervals for each vowel.

who reported having a contrast ($n = 16$), only half of them were classified within the cluster with high Pillai scores in production. However, the 8 participants who did show a contrast in their production were participants who described differences in the vowel (e.g., "The vowel in *rider* lasts longer, and the mouth opens more for that vowel" and "The 'i' in *writer* is a shorter sound and sounds more like 'uh-ee.' The 'i' in *rider* is longer and sounds more like 'ah-ee.'"), with the majority of them ($n = 6$) describing differences in vowel duration specifically. Instead, participants who did not show a contrast in production primarily described differences in the consonant (e.g., "emphasis on the 'd' consonant," "In *rider,* I pronounce the D more," and "There is a pronunciation of the 't' in *writer* not in *rider.*"). It therefore appears that, for those who participate in the Canadian raising split, they do have some level of metalinguistic awareness about changes or differences in the vowel. Other participants appeared to be more heavily influenced by orthographic <t> and <d> in their intuitions about their own productions. Participant intuitions according to the distinction they reported in their speech, relative to their Pillai scores at 30% vowel duration and their perception accuracy rates, are visualized in **Figure 12**. Almost all speakers who reported a vowel contrast in their own speech obtained very high Pillai scores in production and none of the speakers who reported lack of contrast had a Pillai score above 0.5.

Regarding the relationship between reported phonological contrast and perception, speakers who reported that they produced a phonological contrast in the diphthong clearly outperformed other speakers, most having over 80% accuracy in the 2FC identification task. As can also be seen in **Figure 12**, speakers who reported a difference in the pronunciation of the consonant show a very wide range of accuracy in that same task.

Our results for *writer-rider* show that perception and production of the contrast pattern very closely together. Those who were clustered as having a split in production were also clustered as being able to discriminate [ʌi] and [aɪ]. Those who were clustered instead as not producing a contrast were also clustered as being unable to discriminate the two phones. Intra-speaker inconsistencies between perception and production were rare. Speakers who participated in the split also tended to be aware of differences in vowel quality between the target vowels in *writer* and *rider*. These results are clearly different from those of Experiment 1, on the [ɑ-ɔ] contrast. Interestingly whereas Canadian raising has been described as resulting in a marginal contrast, this contrast appears to be more robust when production and perception are considered than the merger in progress that we are also analyzing here.

DISCUSSION

Research on perception and production has largely been conducted under the assumption of a binary distinction based on phonemic category. However, the contrasts under study here, [ɑ-ɔ] ("*cot-caught*") and [ʌi-aɪ] ("*writer-rider*"), cannot be easily categorized as either phonemic or

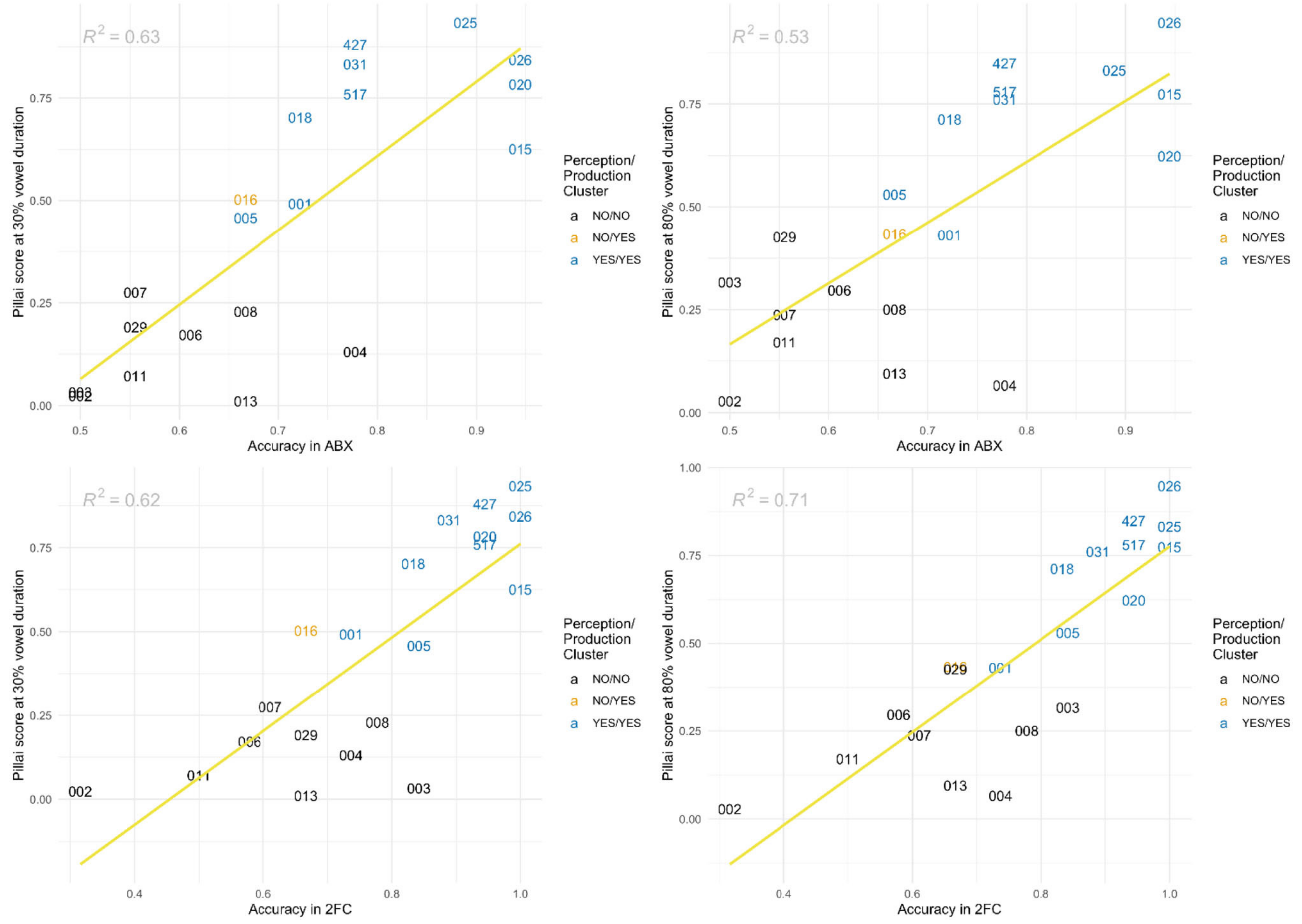

FIGURE 11 | Correlations between ABX perception accuracy and Pillai scores at 30% vowel duration (top left) and 80% vowel duration (top right) and between 2FC perception accuracy and Pillai scores at 30% vowel duration (bottom left) and 80% vowel duration (bottom right) for writer-rider. Participants are clustered based on perception and production independently, with the two clustering analyses visually combined to show 3 groups, where the first NO or YES of each label signifies the contrast in perception and the second signifies a contrast in production. NO/NO (participant clustered in the group with no contrast based on Pillai scores for both perception and production) is shown in black. NO/YES (participant in no-contrast group in perception, but in contrast group in production) is in orange. YES/YES (contrast in both perception and production) is in blue.

allophonic, due to ongoing sound changes involving each pair of phones.

Regarding production first, an examination of Pillai scores for the two contrasts shows larger Pillai scores (greater distance between vowels) for those who produced a contrast between *writer-rider*, compared to smaller Pillai scores for those who appeared to maintain a contrast between *cot-caught*. In other words, although some speakers did produce a discernible contrast between [ɑ-ɔ], the relative distance between the two phones is smaller than that between [ʌi-aɪ] of *writer-rider*. The number of participants in our experiment with relatively large Pillai scores is also greater in our *writer-rider* experiment than in the *cot-caught* experiment. The *writer-rider* split appears to have progressed to a point where it is more stable than the current state of the *cot-caught* contrast, both in terms of phonetic stability, in that the contrast is realized consistently with a greater distance between vowels, and in terms of phonological stability, in that speakers with the contrast patterned very closely in production and perception and were able to identify the contrast in terms of their own vowel productions.

As for the production-perception link, participants with a production contrast (high Pillai scores) for *writer-rider* showed higher accuracy in both perception tasks (ABX: 80%; 2FC: 89%) relative to those who had a contrast in production for *cot-caught* (ABX: 69%; 2FC: 78%). Results for perception accuracy for raisers and non-raisers of *writer-rider* are comparable to previously found results by Strickler (2019) for speakers in the Fort Wayne, Indiana area.

Additionally, the correlations between perception and production for *writer-rider* are much stronger than those for *cot-caught*, further supporting the view that the *writer-rider* split has progressed to a more stable point than the *cot-caught*

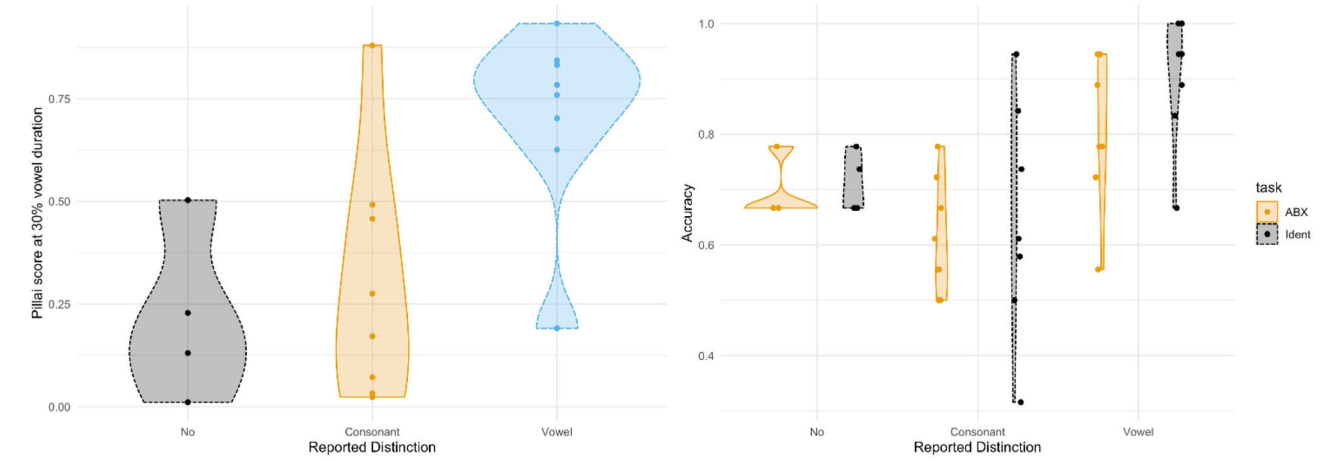

FIGURE 12 | Pillai scores at 30% vowel duration (left) and perception accuracy rates (right) by self-reported distinction for writer-rider. Pillai scores and individual mean accuracy for perception tasks at 30% for *writer-rider,* separated by reported distinction. No, no self-reported contrast; Consonant, reported contrast and described difference as consonantal; and Vowel, reported contrast and described difference in terms of vowel quality.

contrast has at present. Those with the *writer-rider* contrast in production were highly accurate in perception, whereas those without the contrast had poorer performance (85% vs. 65% mean perception accuracy between ABX and 2FC). In comparison, for *cot-caught*, participants with the contrast in production averaged 73% in overall perception accuracy while those without the contrast averaged 69%; compared to the 20% difference between groups for *writer-rider,* the 4% difference for *cot-caught* seems to further indicate weaker perception-production correlation for the latter pair. These findings suggest that speakers in northern and central Illinois may receive variable input regarding the *cot-caught* contrast, although the current status of merger progression cannot be established solely from our experimental results due to a lack of real- or apparent-time data. However, compared to what had previously been documented for northern and central Illinois (including the Chicagoland area, from which we have 14 participants) (e.g., Labov et al., 2006), a number of speakers in our study have fully merged the two sounds. The place of where each participant was raised, along with their status as mergers or non-mergers, can be seen in **Figure 1**.

For *cot-caught*, these maps show a mix of speakers with and without the merger in both Chicagoland and central Illinois, supporting the idea that the merger has advanced in recent years beyond a mere transitional state for some speakers. Of the two speakers from southwest Illinois (the St. Louis area), one appears to maintain the *cot-caught* contrast while the other has merged the two phonemes. Location data for the *writer-rider* contrast showed less change from past data. Speakers in Chicagoland largely have the contrast, although not all of them do; speakers in central Illinois do not. One of the speakers from the St. Louis area shows evidence for the *writer-rider* split, which is a feature of Chicago English. This finding is consistent with previous work by Labov et al. (2006) and the general pattern showing that the St. Louis corridor, along Interstate Highway 55, has served to transmit sound changes from Chicago to the St. Louis area (Labov, 2007). These comparisons, albeit made with a relatively small sample size, tentatively suggest that listeners continue to receive variable input as to the status of the *cot-caught* merger in Illinois, while the geographical distribution of the *writer-rider* split has stayed largely the same.

For each contrast, our clustering analyses separated speakers into two groups for production (high vs. low scores, interpretable as "produce the contrast" and "do not produce the contrast") and two groups for perception ("can perceive the contrast" and "cannot perceive"). Based on the distribution and overlap of speakers in each of these groups, our results appear to support interesting differences regarding the link between perception and production for splits as compared with mergers, and for the trajectories of both sound changes.

For the *writer-rider* split (Experiment 2), two main groups emerged: those who could produce and perceive a contrast between *writer-rider*, and those who could not produce or perceive the contrast. Only one participant was clustered separately as having a contrast in production but not perception. Perception and production showed moderate to strong correlations, suggesting that speakers with the *writer-rider* contrast in production will perform better in perception of that contrast; this corresponds with results found by Strickler (2019). There was only one participant who, despite being clustered as part of the group that produces a contrast, had relatively low accuracy rates in perception. These results show that production and perception largely go together in this case.

Sound Change Trajectories

At the beginning of certain sound changes, such as phonemic splits, listeners may attend to secondary cues which are initially non-contrastive. Responses from speakers who participated in the split seemed to indicate that they were aware at some level of a difference in vowel duration, which has been found to vary depending on whether the flap is an underlying /t/ (138.44 ms) or /d/ (157.72 ms) (Hualde et al., 2017). Strickler (2019) also reports a vowel duration difference of ∼15–17ms. Attention to secondary cues as a path to sound change has been previously suggested for other changes such as vowel fronting (Harrington et al., 2012; Kleber et al., 2012), and perceptual weighting of F0 vs. other primary cues (Kuang and Cui, 2018).

For phonological mergers, such as *cot-caught* (Experiment 1), our results show a more complicated relationship between perception and production, reflective of *cot-caught* as a more marginal contrast for our group of speakers. Although correlations between perception and production were very weak [in line with Baranowski (2013) and Hay et al. (2013)], our clustering analyses illuminated patterns among inconsistencies between perception and production. Our clustered participants were representative of four different groups, including those who were merged in their perception but maintained a contrast, as well as those whose productions were possibly merged but were still able to discriminate in perception.

A substantial body of work has shown that in mergers in progress, merger in perception may come before a merger in production (Labov, 2011: 334), with previous studies finding support for a perception lead (e.g., Di Paolo and Faber, 1990; Herold, 1990). Recent support for a perception lead comes from Pinget et al. (2020), who examined the devoicing of initial labiodental fricatives and initial bilabial stops in Dutch, a process that appears to be resulting in a merger or near merger. For those individuals who participated in this sound change, Pinget and colleagues found that most individuals tended to change their perceptual patterns before changing their production patterns in speech.

However, there have been mixed results regarding the directionality of mergers. In the same study, Pinget et al. (2020) found that even when productions of the voiced and voiceless categories were merged, results showed that perception lagged behind production, in that some participants were still able to discriminate a contrast that they no longer produced. Baranowski (2013), examining the *pin-pen* [ɪ-ɛ] merger in Charleston, South Carolina, found that speakers were more likely to be merged in production than perception, although there was no significant difference between production and perception for *cot-caught*. Austen (2020), tracking the distribution of the *pin-pen* [ɪ-ɛ] merger across the United States, found that almost all speakers who merged the two phones were still able to discriminate them above chance (but below 100% accuracy) in a two-alternative forced choice identification task. For the *Ellen-Allen* merger in New Zealand (Thomas and Hay, 2005; Hay et al., 2013), and for the *cot-caught* merger in Hawai'i and the western United States (Hay et al., 2013), speakers who were merged in production could still discriminate between both pairs in perception.

This group of speakers, classified as those who are merged in production but are still able to discriminate the target phones, comprised 35% of our participants; despite participating in the *cot-caught* merger, a large number of speakers still appeared to maintain the ability to perceive a contrast. This lends support to the hypothesis that the merger is still in progress in northern and central Illinois and that productions of the *cot-caught* contrast may vary at the level of individual speakers. However, it cannot be ruled out that there is instead stable variation in the regions under study, given the lack of real- or apparent-time data in our study.

While sound changes may begin with changes in perception for individual speakers, not all speakers participate in a sound change at the same rate. In a speech community where a merger is still in progress, different speakers may continue to produce both merged and unmerged variants, requiring listeners to maintain a contrast in their perceptual systems. This suggests that, for individual speakers, completion of the merger in perception may lag behind completion of said merger in production (Janson and Schulman, 1983; Pinget et al., 2020). In contrast, for phonemic splits and other sound changes that result in phones being added to a speaker's phonological inventory, no such lag in perception may be expected, as evidenced by the high correlations between perception and production of *writer-rider* by speakers in our study.

To summarize, two marginal contrasts as perceived and produced by the same group of speakers were examined in this study, with one marginal contrast representative of an ongoing phonological merger and the other representative of a phonemic split. For speakers in this study, overall lower accuracy levels in perception for *cot-caught*, and greater overlap in Pillai scores for production, suggest that this contrast is more marginal. The *writer-rider* contrast appears to exhibit more robustness in this community, although perception accuracy was not at ceiling (relative to control items, where perception accuracy was near or at ceiling), suggesting that *writer-rider* may still be considered an example of a marginal contrast (e.g., Hualde et al., 2017, 2021), albeit one that has become more stable in the speech community under study. Strikingly different results were obtained regarding clustering of participants based on their perception and production of the two contrasts, perhaps pointing to different sound change trajectories for mergers as compared with phonemic splits.

DATA AVAILABILITY STATEMENT

Participants of this study did not consent for their data to be shared publicly, so supporting data is not available.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Office for the Protection of Research Subjects, University of Illinois at Urbana-Champaign. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JH, JZ, and LG contributed to conception and design of the study. JZ and LG recruited participants and organized the database. MB performed the statistical analysis. JZ wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.844862/full#supplementary-material

REFERENCES

Austen, M. (2020). Production and perception of the Pin-Pen merger. *J. Linguist. Geogr.* 8, 115–126. doi: 10.1017/jlg.2020.9

Baranowski, M. (2013). On the role of social factors in the loss of phonemic distinctions1ss1. *Engl. Lang. Linguist.* 17, 271–295. doi: 10.1017/S1360674313000038

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4ee4. *J. Stat. Softw.* 67, 1–48 doi: 10.18637/jss.v067.i01

Berkson, K., Davis, S., and Strickler, A. (2017). What does incipient/ay/-raising look like?: A response to Josef Fruehwald. *Lang.* 93, e1e81–e1e91. doi: 10.1353/lan.2017.0050

Bermúdez-Otero, R. (2004). Raising and flapping in Canadian English: grammar and acquisition. In: *Paper Presented at CASTL Colloquium, University of Tromsø*. Available online at: http://www.bermudez-otero.com/tromsoe.pdf

Bloomfield, L. (1939). Menomini morphophonemics. *Travaux du cercle linguistique de Prague.* 8, 105–115.

Boersma, P., and Weenink, D. (2021). *Praat: Doing Phon. by Computer [Computer program]. Version 6.2.04*. Available online at: http://www.praat.org/ (retrieved on December 18, 2021).

Boulenger, V., Hoen, M., Jacquier, C., and Meunier, F. (2011). Interplay between acoustic/phonetic and semantic processes during spoken sentence comprehension: An ERP study. *Brain lang.* 116, 51–63. doi: 10.1016/j.bandl.2010.09.011

Chambers, J. K. (1973). Canadian raising. *Can. J. Linguist.* 18, 113–135. doi: 10.1017/S0008413100007350

Chambers, J. K. (1989). Canadian raising: blocking, fronting, etc. *Am. Speech.* 64, 75–88. doi: 10.2307/455114

Chambers, J. K. (2006). Canadian raising in retrospect and prospect. *Can. J. Linguist. 51(2/3)* 105–118. doi: 10.1017/S000841310000400X

Chitoran, I., and Hualde, J. I. (2007). From hiatus to diphthong: the evolution of vowel sequences in Romance. *Phonology.* 24. 37–75. doi: 10.1017/S095267570700111X

Clopper, C. G., Pisoni, D. B., and De Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *J. Acous. Soc. Am.* 118, 16611661–1676. doi: 10.1121/1.2000774

Coetzee, A. W., Beddor, P. S., Shedden, K., Styler, W., and Wissing, D. (2018). Plosive voicing in Afrikaans: Differential cue weighting and tonogenesis. *J. Phon.*. 66, 185–216. doi: 10.1016/j.wocn.2017.09.009

Crowley, T. (1998). The voiceless fricatives [s] and [h] in Erromangan: O One phoneme, two, or one and a bit? *Australian J. Linguist.* 18, 149–168. doi: 10.1080/07268609808599565

Dailey-O'Cain, J. (1997). Canadian raising in a midwestern US city. *Lang. Variation Change.* 9, 107–120. doi: 10.1017/S0954394500001812

Davis, S., Berkson, K., and Strickler, A. (2020). Unlocking the mystery of dialect B: a note on incipient /aɪ/-raising in Fort Wayne, Indiana. *Am. Speech.* 95, 149–172. doi: 10.1215/00031283-7603207

Di Paolo, M., and Faber, A. (1990). Phonation differences and the phonetic content of the tense-lax contrast in Utah English. *Lang. variation Change.* 2, 155–204. doi: 10.1017/S0954394500000326

Dresher, B. E. (2009). *The Contrastive Hierarchy in Phonology*. Cambridge University Press. p. 121. doi: 10.1017/CBO9780511642005

Edwards, J.an, and Mary, E., Beckman. (2008). Some cross-linguistic evidence for modulation of implicational universals by language-specific frequency effects in phonological development. *Lang. Learn. Develop.* 4, 122–156. doi: 10.1080/15475440801922115

Ferragne, E., Bedoin, N., Boulenger, V., and Pellegrino, F. (2011). The perception of a derived contrast in Scottish English. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)* 667–670.

Freeman, V., and De Decker, P. (2021). Remote sociophonetic data collection: Vowels and nasalization over video conferencing apps. *J. Acoust. Soc. Am.* 149:11211–1223. doi: 10.1121/10.0003529

Fruehwald, J. (2008). The spread of raising: opacity, lexicalization, and diffusion. *Univ Pa. Work. Papers Linguist* 14, 83–92.

Fruehwald, J. (2016). The early influence of phonology on a phonetic change. *Language.* 92, 376–410. doi: 10.1353/lan.2016.0041

Garrett, A., and Johnson, K. (2013). "Phonetic bias in sound change," in *Origins of Sound Change: Approaches to Phonologization, Vol. 1*, 51–97.

Goldsmith, J. A. (1995). Phonological theory. *Handbook Phonol. Theor.* 1–23. doi: 10.1111/b.9780631201267.1996.00003.x

Google (n.d.). Google Maps view of the state of Illinois. Available online at: https://goo.gl/maps/2Uu6nU1GakXJ3z6v7

Graham, L. (2019). Production and contrastiveness of Canadian Raising in Metro-Detroit English. In Sasha, C., Paola, E., Marija, T., and Paul, W. (eds.) *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia. p. 127–131. Canberra: Australasian Speech Science and Technology Association.

Hall, D. C. (2007). *The Role Representation of Contrast in Phonological Theory*. Toronto: University of Toronto.

Hall, K. C. (2013). A typology of intermediate phonological relationships. *Linguist. Rev.* 30, 215–275. doi: 10.1515/tlr-2013-0008

Harrington, J., Kleber, F., and Reubold, U. (2012). The production and perception of coarticulation in two types of sound change in progress. *Speech Plan. Dyn.* 39–62.

Hay, J., Drager, K., and Thomas, B. (2013). Using nonsense words to investigate vowel merger1rr1. *Engl. Lang. Linguist.* 17, 241–269. doi: 10.1017/S1360674313000026

Herd, W., Jongman, A., and Sereno, J. (2010). An acoustic and perceptual analysis of/t/and/d/flaps in American English. *J. Phon.* 38, 504–516. doi: 10.1016/j.wocn.2010.06.003

Herold, R. (1990). *Mechanisms of merger: The implementation distribution of the low back merger in Eastern Pennsylvania.* University of Pennsylvania dissertation.

Hillenbrand, J. M. (2013). Static and dynamic approaches to vowel perception. In *Vowel Inherent Spectral Change*. p. 9–30. Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-14209-3_2

Hockett, C. F. (1958). *A Course in Modern Linguistics*. doi: 10.1111/j.1467-1770.1958.tb00870.x

Hockett, C. F. (1960). The origin of speech. *Sci. Am.* 203, 88–97. doi: 10.1038/scientificamerican0960-88

Hualde, J. I. (2005). Quasi-phonemic contrasts in Spanish. In V. Chand, A. Kelleher, A. J. Rodriguez and B. Schmeiser (eds.), *Proceedings of the 23rd West Coast Conference on Formal Linguist*. p. 374–398. Somerville, MA: Cascadilla Press.

Hualde, J. I., Barlaz, M., and Luchkina, T. (2021). Acoustic differentiation of allophones of /aɪ/ in Chicagoland English: Statistical comparison of formant trajectories. *J. Int. Phon. Assoc.* 1–31. doi: 10.1017/S0025100320000158

Hualde, J. I., Luchkina, T., and Eager, C. D. (2017). Canadian Raising in Chicagoland: The production and perception of a marginal contrast. *J. Phon.* 65, 15–44. doi: 10.1016/j.wocn.2017.06.001

Hume, E., and Johnson, K. (2001). A model of the interplay of speech perception and phonology. In Elizabeth Hume and Keith Johnson (eds.) *The Role of Speech Perception in Phonology*, 3–26. San Diego: Academic Press. doi: 10.1163/9789004454095

Hume, E., and Johnson, K. (2003). The impact of partial phonological contrast on speech perception. In: *Proceedings of the Fifteenth International Congress of Phonetic Sciences.*

Idsardi, W. J. (2006). Canadian raising, opacity, and rephonemicization. *Can. J. Linguist.* 51, 119–126. doi: 10.1017/S0008413100004011

Janson, T., and Schulman, R. (1983). Non-distinctive features and their use. *J. Linguist.* 19, 321–336. doi: 10.1017/S0022226700007763

Jibson, J. (2021). Assessing merged status with Pillai scores based on dynamic formant contours. *Proc. Linguist. Soc. Amer.* 6, 203–212. doi: 10.3765/plsa.v6i1.4961

Joos, M. (1942). A phonological dilemma in Canadian English. *Language*. 141–144. doi: 10.2307/408979

Kager, R. (2008). Lexical irregularity and the typology of contrast. In K. Hanson and S. Inkelas (eds.), *The Nature of the Word: Essays in Honor of Paul Kiparsky*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262083799.003.0017

Kilbury, J. (1983). Talking about phonemics: Centralized diphthongs in a Chicago-area idiolect. In F. Agard, G. Kelley, A. Makkai, and V. B. Makkai (Eds.), *Essays in Honor of Charles F. Hockett*. p. 336–341. Leiden: Brill.

Kiparsky, P. (1985). Some consequences of lexical phonology. *Phonology.* 2, 85–138. doi: 10.1017/S0952675700000397

Kiparsky, P. (2003). Analogy as optimization: 'Exceptions' to Sievers' Law in Gothic. In Aditi Lahiri (ed.) *Analogy, Levelling, Markedness: Principles of Change, Phonology Morphology*. p. 15–46. Berlin: Walter de Gruyter. doi: 10.1515/9783110899917.15

Kleber, F., Harrington, J., and Reubold, U. (2012). The relationship between the perception and production of coarticulation during a sound change in progress. *Lang. Speech.* 55, 383–405. doi: 10.1177/0023830911422194

Kuang, J., and Cui, A. (2018). Relative cue weighting in production and perception of an ongoing sound change in Southern Yi. *J. Phon.* 71, 194–214. doi: 10.1016/j.wocn.2018.09.002

Kurath, H. (1939). *The Linguistic Atlas of New England*. Providence: Brown University Press.

Labov, W. (1994). *Principles of Linguist. change. Vol. 1: Internal factors*. Cambridge, Mass., and Oxford: Blackwell.

Labov, W. (2007). Transmission and diffusion. *Lang.* 83, 344–387. doi: 10.1353/lan.2007.0082

Labov, W. (2011). *Principles of Linguistics Change, Volume 3: Cognitive Cultural Factors*. John Wiley and Sons. doi: 10.1002/9781444327496

Labov, W., Ash, S., and Boberg, C. (2006). *The Atlas of North America English: Phonetics, Phonology and Sound Change*. Walter de Gruyter. doi: 10.1515/9783110167467

Ladd, D. R. (2006). "Distinctive phones" in surface representation. In Louis M. Goldstein, D. H. Whalen and Catherine T. Best (eds.), *Laboratory Phonology*. Vol. 8. p. 3–26. Berlin: Mouton de Gruyter.

Ladd, D. R. (2014). *Simultaneous Structure in Phonology, 28*. OUP Oxford. doi: 10.1093/acprof:oso/9780199670970.001.0001

Lenth, R. V. (2022). *Emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.7.3. Available online at: https://CRAN.R-project.org/package=emmeans

Martin, A., and Peperkamp, S. (2011). Speech perception and phonology. In M. van Oostendorp, C. J. Ewen, E. Hume and K. Rice (eds.), *The Blackwell Companion to Phonology*. p. 2334–2356. Oxford: Wiley-Blackwell. doi: 10.1002/9781444335262.wbctp0098

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In*:* *Interspeech*. p. 498–502. doi: 10.21437/Interspeech.2017-1386

Mielke, J., Armstrong, M., and Hume, E. (2003). Looking through opacity. *Theor. Linguist.* 29, 123–139. doi: 10.1515/thli.29.1-2.123

Milroy, J. (1996). Variation in /ai/ in Northern British English, with comments on Canadian Raising. *Univ. Pa. Work. Papers Linguist.* 3, 16.

Nadeu, M., and Renwick, M. E. (2016). Variation in the lexical distribution and implementation of phonetically similar phonemes in Catalan. *J. Phon.* 58, 22–47. doi: 10.1016/j.wocn.2016.05.003

Nycz, J., and Hall-Lew, L. (2013). Best practices in measuring vowel merger. *Proc. Mtgs. Acoust.* 20, 060008. doi: 10.1121/1.4894063

Pater, J. (2014). Canadian raising with language-specific weighted constraints. *Language*. 230–240. doi: 10.1353/lan.2014.0009

Pinget, A. F., Kager, R., and Van de Velde, H. (2020). Linking variation in perception and production in sound change: Evidence from Dutch obstruent devoicing. *Lang. Speech*. 63, 660–685. doi: 10.1177/0023830919880206

R Core Team (2019). *R: A Language environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available online at: https://www.R-project.org/

Renwick, M. E., and Nadeu, M. (2018). A survey of phonological mid vowel intuitions in central Catalan. *Lang. Speech*. 62, 164–204. doi: 10.1177/0023830917749275

Renwick, M. E. L., and Ladd, D. (2016). Phonetic Distinctiveness vs. Lexical Contrastiveness in Non-Robust Phonemic Contrasts. *J. Assoc. Lab. Phonol.* 7. 1–29. doi: 10.5334/labphon.17

Renwick, M. E. L., Vasilescu, I., Dutrey, C., Lamel, L., and Vieru, B. (2016). Marginal contrast among romanian vowels: evidence from ASR and functional load. In: *Proceedings of Interspeech*. p. 2433–2437. San Francisco, CA. Available online at: http://www.isca-speech.org/archive/Interspeech_2016/pdfs/0762.PDF

Scobbie, J. M., Hewlett, N., and Turk, A. E. (1999). Standard English in Edinburgh and Glasgow: The Scottish vowel length rule revealed. In P. Foulkes and G. Docherty (eds.), *Urban Voices: Variation Change in British Accents*. p. 230–245. London: Arnold.

Shewmake, E. F. (1925). Laws of pronunciation in eastern Virginia. *Modern Lang. Notes*. 40, 489–492. doi: 10.2307/2914584

Shewmake, E. F. (1943). Distinctive Virginia pronunciation. *Am. Speech*. 18, 33–38. doi: 10.2307/487265

Strickler, A. (2019). Within-speaker perception and production of dialectal /aɪ/-raising. In *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia* (pp. 3205–3209).

Thomas, B., and Hay, J. (2005). A pleasant malady: The Ellen/Allan merger in New Zealand English. *Te Reo*. 48, 69.

Vance, T. J. (1987). "Canadian Raising" in some dialects of the northern United States. *Am. Speech.* 62, 195–210. doi: 10.2307/454805

Vennemann, T. (1971). The phonology of Gothic vowels. *Lang.* 47, 90–132. doi: 10.2307/412190

Walker, R. (2005). Weak triggers in vowel harmony. *Natural Lang. Linguist. Theor.* 23, 917–989. doi: 10.1007/s11049-004-4562-z

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 27 July 2022
doi: 10.3389/fcomm.2022.840041

Recall of Own Speech Following Interaction With L2 Speakers: Is There Evidence for Fuzzier Representations?

Frances Baxter, Ghada Khattab, Andreas Krug and Fengting Du*

Speech and Language Sciences Section, School of Education, Communication and Language Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom

OPEN ACCESS

Edited by:
Joseph V. Casillas,
Rutgers, The State University of New Jersey, United States

Reviewed by:
Laura M. Morett,
University of Alabama, United States
Stefano Coretta,
University of Edinburgh, United Kingdom

****Correspondence:***
Ghada Khattab
ghada.khattab@ncl.ac.uk

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *20 December 2021*
Accepted: *16 June 2022*
Published: *27 July 2022*

Citation:
Baxter F, Khattab G, Krug A and Du F (2022) Recall of Own Speech Following Interaction With L2 Speakers: Is There Evidence for Fuzzier Representations? Front. Commun. 7:840041. doi: 10.3389/fcomm.2022.840041

The aim of this study was to test claims that speakers of a first language (L1) incur cognitive and linguistic processing costs when interacting with second language (L2) speakers. This is thought to be due to the extra cognitive effort required for mapping incoming L2 speech signals onto stored phonological, lexical and semantic representations. Recent work suggests that these processing costs may lead to poorer memory of not only the L2 speech, but of one's own produced speech during an interaction with an L2 speaker. Little is known about whether this is also moderated by working memory (WM) capacity and/or the L2 interlocutor's proficiency. In a partial replication study of Lev-Ari et al., 54 healthy L1 English participants performed a WM test and then read a story and answered inference questions about it from a confederate in one of three conditions: the confederate was either a) a fellow L1 speaker; b) a Chinese L2 speaker of English with advanced proficiency or c) a Chinese L2 speaker of English with intermediate proficiency. Following a distractor task, participants were asked to recall their own answers in a surprise response-recognition questionnaire. Participants recognized their responses more accurately after interacting with the L1 speaker compared with the advanced L2 speaker but not compared with the intermediate L2 speaker. WM capacity correlated with higher accuracy when interacting with the L1 speaker, but with lower accuracy when interacting with the intermediate L2 speaker. These results suggest that effortful processing of input may lead to fuzzier lexical and/or semantic representations of one's own produced speech. However, the lack of significance in recall accuracy between the L1 and the intermediate L2 condition suggests other factors may be at play. Qualitative analyses of the conversations provided insights into strategies that individuals adopt to reduce cognitive load and achieve successful communication.

Keywords: speech processing, communication with L2 speakers, accent perception, L2 proficiency, working memory, recall

INTRODUCTION

Speech perception is a complex process which involves all levels of the grammar in a dynamic and graded manner. Portions of the speech signal need to be mapped on to stored lexical (and in some models pre-lexical) forms, which in turn trigger semantic and syntactic representations (e.g., McClelland and Elman, 1986; Gaskell and Marslen-Wilson, 1997; Luce and Pisoni, 1998; Meyer and Schiller, 2011). When the acoustic signal is degraded, listeners take longer to map what they hear onto stored representations; in terms of word-form processing, increased activation of competitors may take place, leading to more competition between words, weaker activation of target and/or longer time for selection of best match. The influence of the quality of the acoustic signal on perception has mostly been tested under controlled situations using noise or manipulated phonetic detail (e.g., Connine et al., 1993; Andruski et al., 1994; Norris et al., 1995). Listening to second language (L2) speech has not typically been considered in psycholinguistic models of speech processing, but some of the same characteristics of input which varies from the listener's representations described above apply to L2 speech. L2 speakers[1] typically make use of resources from their dominant language to scaffold second language production, resulting in phonetic (and other linguistic) patterns which deviate from those of first language (L1) speakers (Iverson et al., 2003; Wolter, 2006; Lev-Ari and Keysar, 2012). These patterns include phonetically similar material as well as distant targets which may be influenced by the orthography, false friends, and a host of other linguistic factors. When processing L2 speech, L1 listeners may encounter high variability in the acoustic realization of words, potentially leading to lexical competition and "effortful listening" (Van Engen and Peelle, 2014, p. 2), requiring more cognitive resources for successful perception (Munro and Derwing, 1995b; Clarke and Garrett, 2004; Lev-Ari and Keysar, 2012; Van Engen, 2015; Lev-Ari et al., 2018). Under such circumstances, listeners may rely more on top-down processing to comprehend the message, considering the interaction as a whole, and understanding the gist in order to maximize interpretation of linguistic structures (Newman and Connolly, 2009; Goslin et al., 2012; Lev-Ari, 2015).

Input from speakers with lower L2 proficiency is expected to lead to more mismatches between the acoustic signal and L1 listeners' stored lexical representations, potentially leading to lower comprehension and requiring more cognitive resources to encode what is heard into memory (Munro and Derwing, 1995a; Van Engen and Peelle, 2014; Van Engen, 2015). Working memory (WM) is the cognitive system where incoming and stored information are integrated during online speech perception and memory encoding in conversation. In general, the poorer the intelligibility of the speech, the more listeners rely on working memory (WM) for encoding and comprehension (Francis and Nusbaum, 2009). Therefore, WM is more active when the speech signal is degraded or when acoustic mismatch increases in situations such as listening to L2 accented speech. This is thought to lead to encoding of less detailed semantic and conceptual representations into long-term memory (Rönnberg et al., 2008, 2013). Lev-Ari and Keysar (2012) tested this in an L2 speech processing context by investigating whether participants were better able to detect word changes in a story when listening to an L1 than an L2 speaker. They found that listeners remember fewer details of what an L2 speaker says compared with an L1 speaker due to their expectation of lower competence of the L2 speaker. This leads to increased reliance on contextual cues to deduce content, a process modulated by WM capacity: listeners with high WM increased their reliance on context and were subsequently less accurate at detecting word changes when the story was told by the L2 speaker. Lev-Ari (2015) suggested that participants with high WM are better able to adapt their language processing and can rely more on top-down processes to aid understanding of L2 speech.

The demands in terms of language processing as a result of listening to an L2 accent may be attenuated after more exposure to the accent, even after a couple of minutes. Clarke and Garrett (2004) exposed L1 English listeners to sentences spoken by Spanish and Chinese L2 speakers of English, as well as by L1 English speakers. Reaction time was at first longer for L2 speech but L1 listeners adapted quickly and any deficit in comprehension attenuated, even after listening to L2 sentences for just 1 min. This suggested that increased interactions between L1 and L2 speakers may support both parties to better compensate for the "processing costs" when listening to accented speech.

Expectations by the listener regarding the language proficiency of an L2 interlocutor can help listeners predict phonetic/phonological, semantic and syntactic features and adapt to these in order to maximize the success of an interaction and the recall thereof (Hailstone et al., 2012; Hanulíková et al., 2012). However, one possible methodological confound relates to social factors which may cloud one's perception of the difficulty in processing L2 speech, or one's willingness to attend to it. For instance, attitudes toward an L2 accent and stereotyping can negatively impact listeners' comprehension and linguistic processing (Fuertes et al., 2002; Lindemann, 2002; Dunton et al., 2011; Lippi-Green, 2012). Sociolinguistic research suggests that listeners with negative attitudes about L2 speakers may not accept the burden of communication during an interaction (Lindemann, 2002). They may also unconsciously perceive L2 speech as less able to reliably convey information and therefore rely more on the context of the interaction (Lev-Ari and Keysar, 2012; Lev-Ari et al., 2018).

While the above work focuses on processing interlocutor speech, a recent contribution to this research suggests that the memory of one's own spoken responses may also be impacted when interacting with an L2 speaker. In a unique study, which forms the basis of the experimental procedure developed here, Lev-Ari et al. (2018) constructed interactions between L1 and L2 speakers of English and tested if an L1 listener's recall of their own produced utterances was influenced by speaker condition. Participants read a story and were interviewed by an L1 or L2 confederate with inference questions about the story. Afterwards, participants performed a multiple-choice memory recognition

[1] The bulk of the research we review here typically refers to L2 speech as "non-native" or "foreign-accented", but we make a concerted effort to avoid these terms given their negative connotations.

questionnaire of their own responses. Participants had a better memory for their own responses after an interaction with a fellow L1 speaker and were more likely to remember all their responses if they were interacting with an L1 rather than an L2 speaker. Participants were also more likely to choose a distractor response to represent their own answer or a "false alarm" when interacting with L2 rather than L1 speakers. Lev-Ari et al. (2018) argue that this impact is due to the more effortful integration of the incoming speech signal with the listeners' stored lexical and semantic representations and of their sociolinguistic expectations of the L2 speaker. However, WM capacity was not directly tested in this procedure.

In sum, processing an L2 accent has been shown to incur processing costs in terms of the lexical accuracy and semantic detail of recall, with mediating factors such as familiarization, (perceived) linguistic proficiency of the L2 speaker, and attitudes toward L2 speakers. More recent research suggests that this cost may extend to recall of an L1 interlocutor's own speech, but this line of enquiry is still in its infancy. The present study extends this work by examining whether L2 proficiency plays a role in L1 listeners' recall of their own produced speech when interacting with L2 speakers, and the role of WM in such recall. In particular, we seek to answer the following questions:

1) Does interacting with an L2 speaker of English have a negative impact on the recall of L1 speakers' own produced speech during an L1-L2 interaction? This part will be done through a conceptual replication of Lev-Ari et al. (2018) study.
2 Does a lower proficiency in the L2 have a more negative impact on the recall of L1 interlocutors' own produced speech? This is an additional factor which was not part of Lev-Ari et al. (2018) study.
3) Does WM mediate an L1 speaker's ability to recall their own answers after an interaction with L1 or L2 speakers? This factor was explored in an earlier study by Lev-Ari (2015).

Our working hypothesis is that participants who engage in an interaction with an L2 speaker will process fewer lexical and/or semantic details of their own speech, and hence recognize fewer of their own responses compared to when interacting with an L1 speaker. This is due to the cognitive effort involved in processing L2 acoustic input that does not match own stored phonetic details for the intended lexical target or which makes it harder to identify the intended target. This may shift attention away from own answers during the interaction. Any phonetic accommodation to the L2 speaker that is achieved in the fly may also make it harder to subsequently retrieve the message if the acoustic output does not match stored forms. The effect is predicted to be greater in the intermediate proficiency condition, due to the expected greater distance between the phonetic detail of the input and stored representations. We also predict more "false alarms" to be selected in the L2 intermediate condition than in the L2- advanced condition, with the fewest in the L1 condition. Further, participants' Working Memory scores are predicted to show a negative correlation with their recall score when interacting with L2 speakers, and a positive one when interacting with L1 speakers. This is based on previous findings (Lev-Ari, 2015) which show that individuals with high WM increase their reliance on context when interacting with L2 speakers compared with individuals with low WM, thereby remembering less lexical detail of what was said.

MATERIALS AND METHODS

Participants

The participants were 54 L1 English speakers who were studying speech and language therapy and linguistics-related degrees at a university in the north-east of England. They were all females aged between 19 and 30 years with no history of speech, language or communication needs and had no knowledge of this experimental procedure before the debrief.

Confederates

Three confederates were selected through an interview by the first two authors and remunerated for their participation. They were informed of the true aims of the experiment, were offered training on the experimental procedure, and were instrumental in the deception strategy. They were matched for gender (all female), age range (18–30) and education with each other and with participants (the confederates were also students at the same university). One confederate was a speaker of English as a first language (L1) and the other two were Mandarin L2 English speakers, with average scores of 6.5 and 8 out of 9 respectively on the International English Language Testing System (IELTS, 2007). The IELTS overall score represents the aggregate results of speaking, listening, reading and writing skills and is presented in bands from 5 to 9; band 6 demonstrates effective command of language with some inaccuracies and misunderstandings, while band 8 represent fully operations command of the language with only occasional inaccuracies. The two participants were hence regarded as having intermediate proficiency (L2_I) and advanced English proficiency (L2_A), respectively. Recruiting L2 confederates from the same L1 language background ensured that differences between them were in the degree rather than nature of L1 influence on the L2 since the characteristics of L2 accents are relatively consistent across speakers from the same L1 backgrounds (Bradlow and Bent, 2008). The choice of Mandarin as the L2 was that of convenience, due to the large population of Mandarin speakers in and around the university where recruitment took place.

Procedure

Seventeen participants were randomly matched with the L2_I confederate, 19 with the L2_A confederate and 18 with the L1 confederate. Unbeknownst to them, each participant was scheduled to arrive at the authors' research lab at the same time as their matched confederate and was made to believe that both were participants in the study. After initial instructions given by the first author, each participant and their confederate were seated in front of a computer to complete a WM test (Section Working Memory Testing Phase). Once the WM task was completed, the participants had a short break and moved on to the experimental task. These were recorded using an Edirol R-09 recorder with a sampling rate of 44,100 Hz and 16-bit amplitude resolution.

In order to keep testing instructions constant during the experiment and to ensure replicability of procedures, all instructions to participants were standardized as a script that was rehearsed and delivered by the first author. After giving instructions to the participants about the tasks that they would engage in, the researcher left to an adjoining room with an observation window so that the proceedings only focused upon the confederate-participant interaction and not that of researcher-participant. The tasks are described in chronological order below.

Tasks and Scoring

Working Memory Testing Phase

The RSPAN (Automated Reading Span Test) test (Daneman and Carpenter, 1980) was first conducted and administered using Millisecond software in Inquisit and took 15–20 minutes to complete. In order to protect the status of confederates, they were instructed to complete the RSPAN at the same time as their matched participant during each trial. Briefly, the RSPAN consists of a series of sets, each set alternating between the presentation of a sentence to participants which they have to judge on plausibility and a letter after every sentence which they need to memorize and recall in order at the end of each set. The score included in the analyses is the absolute RSPAN, a measure the number of perfectly recalled sets in terms of letters and their order. For example, if a participant correctly recalled 2 letters in a set size of 2, the absolute score would be 2; otherwise, they would get 0 absolute score. This was used as a latent variable for WM capacity since it requires the ability to integrate different sources of information in a set amount of time.

Reading Comprehension and Surprise Memory Phase

The participant and confederate were then informed that they would silently read a story and pick a "random" color out of a box to decide who asks questions about it and who answers these. The experiment was set so that the confederate always asked the questions.

The 200-word text (**Table 1**) which was adapted for the story comprehension activity during the test was sourced from the Discourse Comprehension Tests, set B (Brookshire and Nicholas, 1993). This is a highly readable, clearly structured narrative, from which inference questions could be developed for the questions.

In order to ensure consistency in the linguistic content of the questions and limit differences in delivery to accent, a script with the questions was provided to all confederates ahead of the experiment and they had the chance to practice these. Seven inference questions based on the text were provided for confederates to ask participants after they finished reading the text (**Table 2**). The participants were free to respond to each question at any length.

TABLE 1 | Story used for the reading comprehension task.

George Smith was a quiet French bookkeeper. None of his friends believed him when he told them he was going to walk across Niagara Falls on a tightrope. But here he was, one spring day, looking at the rope which was stretched 50 meters above the falls. George's wife stood beside him trying to convince him not to try such a foolish stunt. She told him to think of his family. George just shook his head stubbornly and told her that this was something he had to do. Then he began to practice for the crossing by walking back and forth on a narrow wooden beam. By the time George was ready to begin the crossing, almost a thousand people had gathered to watch him. George stood uncertainly at the edge of the river. The long rope swayed slightly in the breeze. Slowly he set out across the rope toward the Canadian side of the falls. Twenty minutes later, a television reporter came up to him and asked, "What are you going to do next?" George thought for a moment and then answered, "I guess I'll walk back across the rope. I left my car on the other side."

TABLE 2 | Inference questions.

1. Why didn't any of George's friends believe him when he said he was going to cross Niagara Falls?
2. Why did his wife try to convince him not to do it?
3. Why was George determined to do it?
4. Why did George hesitate uncertainly at the edge of the river?
5. Why was the TV reporter interested in approaching George?
6. Why did people gather to watch him?
7. Why did he hesitate before answering the reporter?

TABLE 3 | Example inference question with potential responses.

Why didn't any of George's friends believe him when he said he was going to cross Niagara Falls?

- He'd never done anything like that before.
- It was something really extreme.
- His friends didn't see him as adventurous.
- A quiet person would not be expected to do that.
- A bookkeeper would be unlikely to do that.
- He was an unlikely character for dangerous stunts.
- His friends thought he was joking/wasn't serious.
- It's the kind of stunt people joke about doing.
- His friends did not think he was trained to do it.

All participants then completed a five-minute distractor picture-puzzle task which served to intercept the instant memory of their responses. The task consisted of 16 sets of pictures (four in each set); the participants were asked to examine each set and write down a three-letter word that best describes the four pictures within it. After this, participants were given 5 min to complete a surprise memory questionnaire (inference questions with possible answers). Here, the same seven questions that were asked of participants during the reading task were shown to them again with potential responses (**Table 3**). They could choose more than one response if necessary. Before experimental procedures began, an informal pilot study was conducted on peers of a similar demographic to the target sample, which informed the range of possible answers in constructing the memory questionnaire. Participants were asked to circle the responses which best represented their spoken answer in the interview. If participants were outside of three standard deviations from the overall mean number of responses or false alarms, their data were excluded from the analysis. Using measures such as the mean and standard deviation on the count data of the current study is not unproblematic but effectively identified two participants who circled, on average, five answers per question, compared to the overall mean of fewer than two responses. These participants

TABLE 4 | Scoring example.

During the reading comprehension task:

- Confederate: Why did his wife try to convince him not to do it?
- Participant: Because she thought it was dangerous.

During the surprise memory task, the participant then selected the following three answers on the response sheet:
('*' means it is not what they said during the interview)

- Because it was dangerous.
- She was trying to protect him. *
- Because he could have died. *

were excluded. The data from another participant was also excluded because they had retained the story text from the task and read from it, rendering the recall measure void.

Scoring of Inference Questions With Potential Responses

The first author calculated participants' recognition of responses they gave during the interview by comparing the responses they selected in the surprise recall test with the responses they provided during the recorded interview. Participants could only get 100% when the answers they selected perfectly matched what they said during the interaction with the confederate. Partial scores were awarded in other scenarios. For instance, if the participant chose one answer when there was scope to select another, then only 50% was awarded to that question; or if the participant gave one answer which agreed with the response along with two others which deviated from it (false alarms) they scored 33% (e.g., **Table 4**). False alarms were of particular interest because they offered a window into whether participants had fuzzier linguistic representations of what they said. If participants had answered a question in the reading task with "I don't know", this question was subsequently not used in the inference questions. Scores obtained were then compared with a second iteration of scoring on 100% of the sample (Intra-Rater) and with 20% of recordings from an independent coder (Inter-Rater). Cohen's (1960) was used for testing Intra- and Inter-rater reliability, with the first yielding a value of 0.88, suggesting high degree of agreement; the smaller dataset for Inter-rater reliability (10 observations) did not render it optimal for Cohen's Kappa testing as the two sets of scores were too similar, with the highest disagreement in scores being 7%. However, Pearson's r for the Inter-rater reliability was at 0.96, suggesting high agreement between the first author and the independent coder.

Language Background and Debrief

After the main tasks, participants completed a language background questionnaire which also included a self-rating measure of how often participants communicated with Mandarin L2 speakers of English.

To keep in line with ethical procedures, participants were given a verbal debrief accompanied with an explanation sheet after they had completed the experiment. All participants reported to have been successfully misled by the role of the confederate and were given the opportunity to ask any questions about the research.

STATISTICAL ANALYSES AND RESULTS

Recall

This subsection addresses the participants' recall of their own speech, as measured by the percentage score from the surprise memory phase of the experiment. Lev-Ari et al. (2018) used mixed effects simple logistic regression in their study to identify significant effects. For the current dataset, the original intention was to conduct the analysis *via* mixed effects ordinal logistic regression to allow for a more nuanced coding of recall because the response variable in ordinal logistic regression can have more than two levels. However, both ordinal and simple logistic regression models resulted in issues of singular fit, particularly for the random effect of participant on recall. This suggests that the dataset in its current shape was insufficient for models that included random effects. Rather than fitting underpowered models, we decided to aggregate the data by participant. As a result, rather than having seven potential data points for each participant (one per question), we used one average recall score per participant. We acknowledge that aggregating the data results, first, in the loss of information on within-participant variation and, second, the necessity to interpret the results with caution due to increased type I and type M/S error rates. However, we considered it a practical solution to interrogate the data without fitting models that are too complex for the dataset.

Since the independent variable, that is the average recall score per participant, was bounded in the interval $0 \leq y \leq 1$, a beta regression rather than a linear regression was fitted to the data. Beta regressions are commonly used for proportion data, such as the one in the current dataset, because the data have natural limits (0 and 1) and often do not follow a normal distribution. For beta regressions, the response variable usually cannot take the extreme values $y = 0$ and $y = 1$. However, some participants in our study scored perfectly for all answered questions, resulting in an average recall score of $y = 1$. Therefore, following Smithson and Verkuilen (2006), all recall scores were transformed with the following formula, which included the sample size $n = 54$:

$$y_{transformed} = \frac{\left(y_{raw} \cdot (n - 1) + 0.5\right)}{n}$$

The model included condition, WM and response length as fixed effects. An interaction term between condition and WM was also included. The coding of each of these factors, the rationale behind including them and the predictions for their effect on the recall of the participants' own speech are provided below:

Condition

Condition is a categorical variable with three levels that correspond to whether the participants interacted with an L1 confederate, an L2_A or an L2_I confederate. We predicted that participants' recall of their own responses would follow this pattern: L1 > L2_A > L2_I.

TABLE 5 | Beta regression model output for recall of participants' own speech.

Coefficient	Estimate	Std. error	*Z*-value adjusted	*P*-value adjusted
Intercept	1.24	0.19	6.64	< 0.001
Condition				
L1 vs. L2_I	−0.24	0.26	−0.90	0.368
L1 vs. L2_A	**–0.59**	**0.25**	**−1.98**	**0.048**
L2_I vs. L2_A	−0.35	0.26	−0.94	0.348
Working memory	0.02	0.01	1.38	0.168
Response length	−0.02	0.02	−0.74	0.460
Condition (L1 vs. L2_I) x working memory	**−0.04**	**0.02**	**−2.26**	**0.024**
Condition (L1 vs. L2_A) x working memory	−0.01	0.01	−0.01	> 0.99
Condition (L2_I vs. L2_A) x working memory	0.03	0.02	−1.62	0.105

The bold values indicate signals with significant result.

WM

WM is a continuous predictor and corresponds to the participants' absolute RSPAN score. This score adds up the number of letters from all sets that were perfectly recalled by the participants. To allow for a more sensible interpretation of the model, the RSPAN score was centered before it was added to the model. The predictions for WM are less straight forward. On the one hand, higher WM usually results in better recall, which is evident from how WM is measured in the RSPAN procedure. However, Lev-Ari and Keysar (2012) found that their participants' recall of L2 speech was worse if their WM was higher. They argued that the adjustment to L2 speech required cognitive resources and, thus, was only possible for participants with higher WM. However, the participants' adjustment was found to lead to less-detailed representations and, as a result, worse recall. It is not yet clear if this finding also extends to the recall of one's own speech. Based on these considerations, it was predicted that, in interactions with L1 confederates, higher WM would be beneficial and correlate with better recall. In interactions with L2 confederates, higher WM would result in fuzzier representations of one's own spoken output. To test the potentially differential effect of WM on the L1 vs. L2 conditions, an interaction between condition and WM was included.

Response Length

Response length is a continuous predictor and corresponds to the number of words that the participants used to respond to a question. It was centered. We predicted that participants would recall their responses better if they gave shorter responses because there would usually be fewer items to recall. Words that were used by the participants to ask for a repetition of the question were not counted. Hesitation markers, such as *um* and *erm*, were also disregarded for the word count.

Table 5 provides the output of the beta regression model. Beta regression outputs the model coefficients in log odds or logits, which can be transformed into probabilities. For example, the estimate for the intercept in **Table 5** ($x_0 = 1.24$) refers to the reference level of condition (i.e., L1 confederate) with WM and response length being kept at constant levels. Since WM and response length were centered, these constant levels refer to the mean WM and the mean response length. Thus, the probability for an overall perfect recall score when participants interacted with an L1 confederate, had average WM and an average response length is:

$$\frac{\exp(1.24)}{1+\exp(1.24)} \approx 0.776 \ = 77.6\%$$

The logits and the corresponding probabilities for the other conditions can be calculated by adding the model estimates to the intercept. For instance, the probability of an overall perfect recall score for participants who interacted with an L2_A confederate and had average WM as well as average response length is:

$$\frac{\exp(1.24 \ - 0.59)}{1+\exp(1.24 - 0.59)} \approx 0.657 = 65.7\%$$

The pairwise comparisons between conditions from **Table 5**[2] show one significant effect. Participants' recall of their own speech is worse when they interact with an L2_A as compared to an L1 confederate. Additionally, there one of the interaction terms was significant. Participants' recall is mediated by WM in that, when comparing interactions with an L1 vs. L2_I confederate, increasing WM has a detrimental effect on recall after interactions with L2_I confederates. These two effects are addressed in more detail below.

The mean percentage scores across conditions are shown in **Figure 1**. On average, participants' recall scores are 76.1, 65.6, and 70.2% for the L1, L2_A and L2_I condition, respectively. This corresponds to the significant difference between the L1 and L2_A conditions.

Figure 2 helps to better understand the interaction effect between condition and WM for L1 vs. L2_I confederates. The participants' RSPAN score is plotted against their average recall score. Condition is coded by color. Each point in the plot represents one participant. Lines of best fit were added to show the relationship between WM and condition. The range of available RSPAN scores varied between conditions. Therefore, the horizontal span of the lines is not equally large across the three conditions. As can be seen, WM is beneficial for recall in the L1 condition. In the L2_I condition, however, higher WM resulted in worse recall of the participants' own speech.

False Alarms

False alarms were defined as answers that the participants recalled as their own speech during the interview phase although they had not given these answers previously. False alarms

[2] The model coefficients for the pairwise comparison between the L2_I and L2_A conditions were taken from a model with the reference level L2_I for condition. P-values (and the corresponding z-values) were adjusted via Bonferroni-Holm corrections to account for the three pairwise comparisons.

are already incorporated into the recall measure presented in Subsection Recall. For example, a recall score of 50% could encode a question for which a participant should have circled one answer only but in reality circled an additional answer (i.e., gave a false alarm). In addition to the above analyses, false alarms are considered here separately because they encode how fuzzy a specific participant's lexical and semantic representations were. If a participant gave one or several false alarms for a question, their memory representations were likely fuzzier.

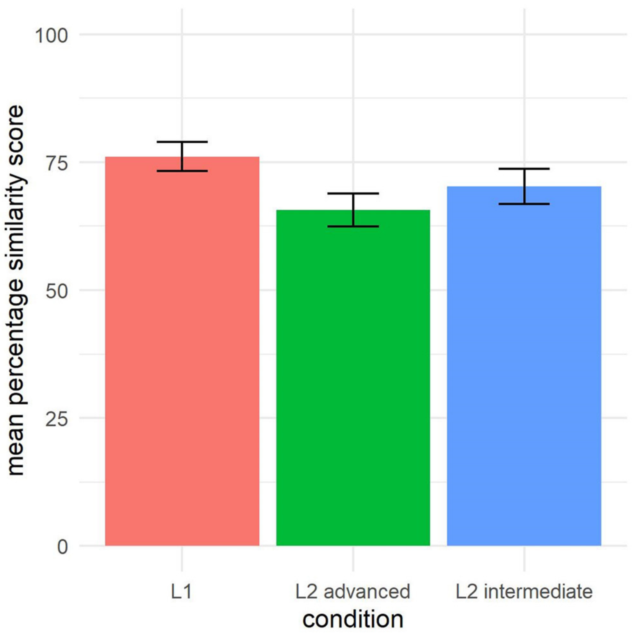

FIGURE 1 | Recall as mediated by condition (error bars: standard error).

Mixed effects simple logistic regression models were used to analyse the false alarms in the recall task. Since there were only relatively few cases with more than one false alarm (23 out of 360 responses), the data were coded in a binary fashion, with responses either containing false alarms or not. The models did not result in any singular fit or convergence issues, which suggested sufficient power and did not warrant for an aggregation of the data. The fixed effects in the models were condition, centered WM and centered response length. An interaction term between condition and WM was also added to the model. In addition to these fixed effects, the models included two random effects:

Participant

Since the occurrence of false alarms might vary beyond the fixed predictors specified above, random intercepts were included in the models. By-participants slopes were not appropriate because of the between-participants design of the study.

Question

The same seven questions were used for all participants across the three conditions. Therefore, by-question random intercepts were fitted as well as, initially, by-question random slopes for the fixed effect of condition. These random slopes were later dropped as the random effects structure proved too complex for the dataset.

The predictions for the occurrence of false alarms were in line with the predictions for the recall score in Subsection Recall. False alarms were predicted to be more prominent when the

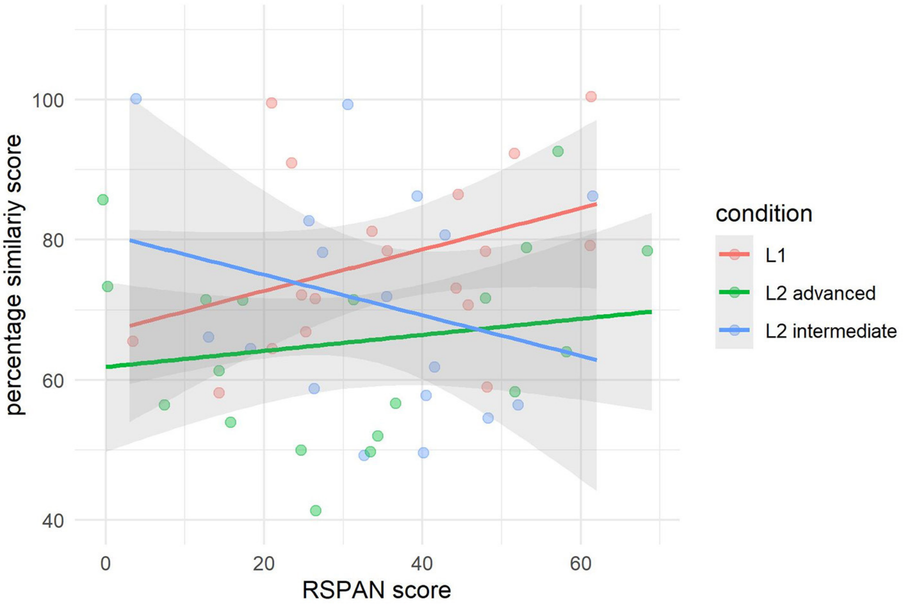

FIGURE 2 | Recall as mediated by condition and WM.

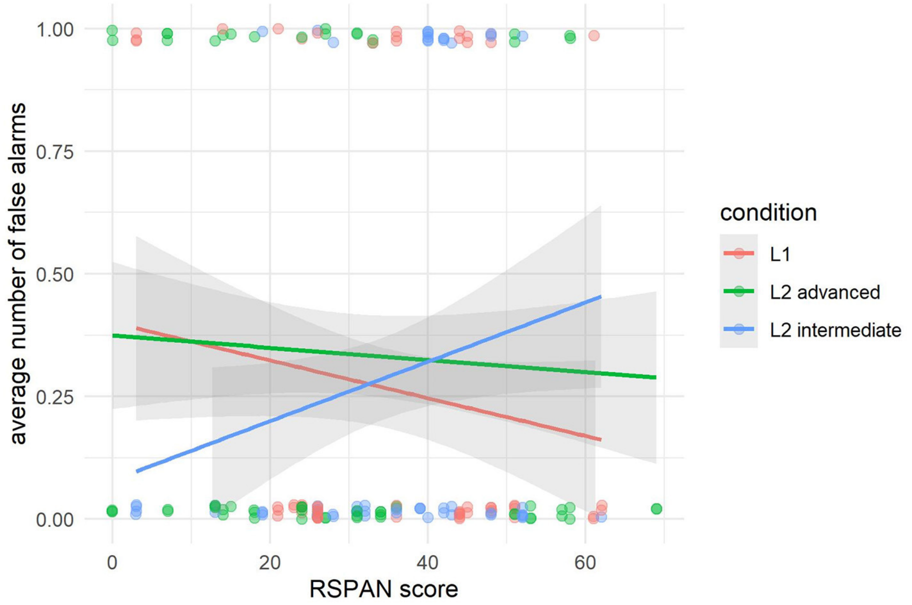

FIGURE 3 | Presence of false alarms as mediated by condition and WM.

participants interacted with L2_I as compared to L2_A and L1 confederates. The effect of condition was predicted to be mediated by WM, with higher WM improving recall in the L1 condition but decreasing it in the L2 condition, especially if the confederate had an intermediate command of English. Longer responses were predicted to result in a higher probability that false alarms would occur.

Likelihood ratio comparisons were used to identify significant effects. The full model was systematically reduced by one fixed effect or interaction term. Model comparisons then showed if the effect or interaction in question had a significant effect on the occurrence of false alarms[3]. The model comparisons showed a significant effect of the interaction between condition and WM ($p = 0.031$). The other effects in the model did not reach significance: condition ($p = 0.507$), WM ($p = 0.733$) and response length ($p = 0.228$).

Figure 3 shows the significant interaction effect in further detail. WM, as measured by the RSPAN score, is shown on the horizontal axis. Each point in the graph represents one of the seven questions. Because of the binary coding of false alarms, a participant's response to each question either did or did not contain a false alarm. Some jitter was added to the points so that they would not overlap for each participant. Lines of best fit are shown in different colors, one for each condition. The interaction effect stems from the different effects of increasing WM for L2 confederates on the one hand, and L1 as well as L2_A confederates on the other hand. Participants with higher WM are more likely to give false alarms in the L2_I than in the other two conditions. This indicates worse recall of one's own speech in interactions with an L2_I confederate, provided that WM is high.

Response Length

Following an informal observation that participants gave comparably shorter answers to the L2 confederate with intermediate proficiency than to the other two confederates, an exploratory analysis was carried out to quantify this observation and report on a potential structural difference in interactions with L2 speakers. Response length was operationalised as the number of words in a participant's response to an inference question (see Subsection Recall).

To see if the participants' response length varied significantly across the three conditions, linear mixed effects models were used. Condition was added as a fixed effect with three levels (L1, L2_A and L2_I). Based on the observations during the experiment, it was predicted that responses would be shorter, that is contain fewer words, in interactions with the L2_I confederate. The models further included the random effects specified for the models in Subsection Recall.

No significant effect of condition on response length was found through the model comparisons ($p = 0.080$). However, the average number of words per response per conditions, as shown in **Figure 4**, displays a trend in the data that is in line with the qualitative comment above. The average response length was 8.8 words ($sd = 4.3$ words) for participants who interacted with the L2_I confederate, 11.7 words ($sd = 7.4$ words) for participants who interacted with the L1 confederate and 12.5 words ($sd = 7.6$ words) for participants who interacted with the L2_A confederate. Although the difference between the conditions is not significant, the shorter responses to the L2_I confederate are informative

[3] P-values for fixed effects that were included in the interaction (i.e. condition and WM) were identified by comparing a model without the interaction and a model without the interaction and without the fixed effect in question.

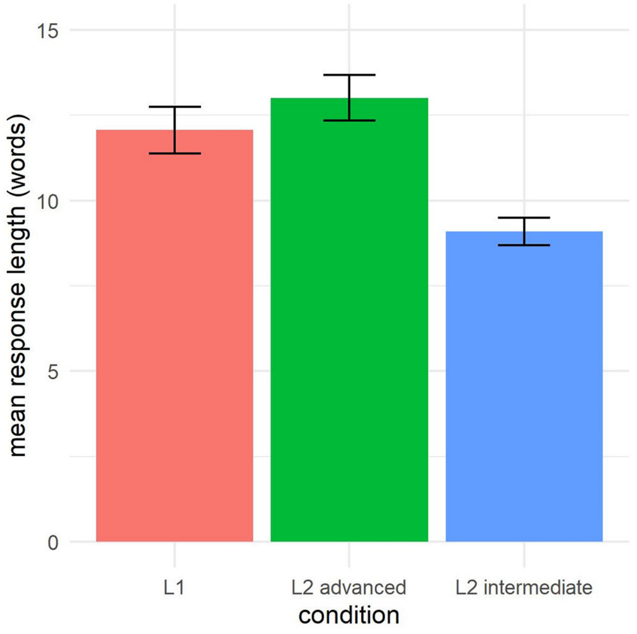

FIGURE 4 | Response length as mediated by condition (error bars: standard error).

and will be discussed in the following, along with the preceding results.

DISCUSSION

This study investigated the potential effect of interacting with L2 speakers on the recall of one's own speech. The aim was to explore whether there are additional processing costs when listening to L2 accented speech as described in previous research (e.g., Van Engen and Peelle, 2014; Van Engen, 2015; Lev-Ari et al., 2018), and whether these are modulated by each of L2 proficiency of the speaker and/or the WM of the L1 speaker. The study found a cognitive disadvantage for participants who interacted with L2 compared with L1 confederates, but only in the L2_A condition. In other words, participants remembered their own responses more accurately, with their recall score higher in the L1 interlocutor condition than in the L2_A condition, but not when compared with the L2_I condition. There was therefore no general across-the-board effect of L2 interaction on recall. These results only partially replicate Lev-Ari et al. (2018) results and are somewhat surprising, since one would have expected lower English proficiency of the confederate to lead to more effortful processing for L1 participants and therefore fuzzier lexical and/or semantic representations, leading to worse recall. They suggest that other factors may have been at play during the task, which affected the communication and degree of orientation to the L2_I speaker. Given that the participants were answering questions about a passage they had just read ahead of their interaction with the confederates, their processing did not only solely consist of bottom-up processing of the linguistic signal; rather, they will have been able to use contextual cues from the passage. Individual differences may also have played a role, since speech processing does not only involve processing of the linguistic signal; the use of extralinguistic and contextual information (e.g. knowledge about the world, expectations from particular situations) is commonly incorporated into the listening process, influencing how individuals come to understand the same discourse (Garman, 2012).

The WM results show a significant interaction between WM capacity and each of recall and false alarms in the participants recall of own produced speech, but with opposing effects depending on the language background of the confederate: higher WM led to better recall and fewer false alarms following communication with the L1 confederate, but worse recall and increased false alarms following communication with the L2_I confederate. These results suggest that speakers with high WM can benefit from integrating social-indexical information in their processing of an accent in the familiar/more compatible L1 condition (Drager, 2011), but this integration is more effortful in the L2 condition and leads to fuzzier lexical and semantic representations of one's own responses. This is the first study to extend previous WM findings to the less detailed recall of one's own produced speech. While the ability to use WM resources in challenging listening conditions enables listeners to orient their attention to their interlocutor and recall more of what they hear (Van Engen et al., 2012), this might have adverse effects on one's own memory of their speech. Another reason for the fuzzier recall of one's own spoken utterances may be due to speakers also adapting their own speech to that of their interlocutor, leading them to remember their own responses less accurately due to the greater mismatch between the acoustics of the response they produced and their own lexical representations (Akeroyd, 2008; Rönnberg et al., 2008, 2013; Lev-Ari and Keysar, 2012). While we did not analyse the speech of our participants in their interactions with confederates, a large proportion of the participants were speech and language therapy trainees who are expected to be particularly skilled at orienting their speech to the listener. While this will have improved their recognition of what the interlocutor said, it may have adversely affected their recall of the detail in the L2_I condition. Importantly though, there was no main effect for language condition on recall.

Conceptual and semantic representations of language have been suggested to be less detailed after listening to an L2 speaker, leading to adverse effects on lexical access both in terms of interlocutor speech and one's own speech (Rönnberg et al., 2008, 2013; Lev-Ari and Keysar, 2012; Lev-Ari, 2015; Lev-Ari et al., 2018). In this study we do not find strong evidence for the latter; while processing a less familiar accent may indeed be more effortful, strategies that both L1 and L2 speaker adopt during the interaction may help mitigate this effect. It is also important to note that, while research in this area has focused on L2 or so-called "foreign" or "non-native" accents, any difficulty that is due to unfamiliarity and lower intelligibility of an accent could equally apply to L1 interactions between speakers of different regional accents of the same language (Goslin et al., 2012; Lev-Ari and Keysar, 2012). It is important to disentangle subjective expectations relating to the perceived difficulty of processing L2 accents from the more general increased cognitive load that may be required when processing an unfamiliar accent. The underlying sources of this load offer an interesting window into how we store and represent speech; storing social-indexical information together with lexical information during communication with other speakers is advantageous (Goldinger, 1998) but can also incur a 'cost' when processing unfamiliar speech.

The degree to which interlocutors are at ease in this unfamiliar setting and the strategies they adopt can either alleviate or compound the effortful communication. In this study qualitative observations of the communication between our participants and the confederates suggested a naturalistic and conversational style used by the L2_A, but a more mechanical, less relaxed interaction style by the L2_I confederate despite both receiving the same training. This may either be due to differences in proficiency or in personality and may have influenced the participants' conversation style too. For instance: 1) during the communication with the L1 and L2_A confederates, both interlocutors maintained eye contact throughout the interview, more often than during the communication with the L2_I confederate; 2) the L2_A confederate acted relaxed and laughed before the first question began, while the L2_I confederate did not; 3) the L2_A confederate used interjections before asking the questions, which may have increased the naturalness of the conversation and given the participants time to get ready for the question; the L2_I confederate tended to ask questions directly. There was a tendency for L2_I confederate's answers to be shorter, but this did not prove significantly different from the answers that the other two confederates gave; 4) the L1 and L2_A confederates were more interactive in the interview, smiling at or nodding to their participants, while the L2_I confederate was more task-oriented and less interactive with their participants; 5) participants in the L2_I condition asked the confederate to repeat their question more often than in the L2_A and L1 condition. The combined effect of these differences may have led to more entrainment between the participants and L2_A speakers, albeit with an increased cost to the participants' recall of their own speech. On the other hand, participants in the L2_I condition may have attended more to the task, and conversely remembered more of their own responses to the questions. The confederates' accents in this case may have been less likely to impact on participants' encoding of their own responses into memory.

It is important to note that, regardless of the differences in recall scores in the L1 and L2_A condition, recall scores were relatively high across all three conditions. The generalization of these results needs to be considered with caution for two reasons. First, the participants in this study were mainly SLT trainees who may have already possessed the skills to be attentive to the needs of the interlocutor in order to maximize communicative success and may therefore have been more adept at adapting to the needs of the situation. Second, only one confederate was used in each condition, which might have resulted in speaker-specific effects. Nevertheless, what this suggests is that the success of communication between L1 and L2 speakers, or interactions between speakers who may not be familiar with each other's accent more generally, should not be the onus of one party, typically the speaker of the less-dominant accent. Attention and conversational strategies on the part of both interlocutors can overcome communicative challenges and ensure the success of the interaction, albeit with increased cognitive processing load and possible initial toll on the detail of the lexical and/or semantic representations of own and others' speech. Increased exposure to L2 accents has also been shown to improve the processing of these and other unencountered accents (Baese-Berk et al., 2013) in turn increasing listeners' trust in what L2 speakers say (Boduch-Grabka and Lev-Ari, 2021). This demonstrates that familiarization with diverse accents rather than expecting L2 speakers to reduce their "foreign" accent is a more equitable way forward in improving L1-L2 communication. This can be achieved on a large scale if various industries such as the media, education, and the arts made an effort to give more platform to speakers of non-dominant and non-standard varieties.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

FB and GK co-produced the literature review and background to the study and co-designed the methodology. FB oversaw the data collection and carried out the initial analyses. AK and FD supported the data analyses and carried out the statistical analyses. All authors co-wrote the discussion, contributed to the article, and approved the submitted version.

ACKNOWLEDGMENTS

We are grateful to the participants for taking part in this study and the confederates for their patience throughout the training and implementation of the research. We are also deeply grateful to Shiri Lev-Ari for her advice and insights in the early stages of the study and for generously sharing with us the story used in the Lev-Ari et al. (2018) study, as well as information on scoring and analysis.

REFERENCES

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *Int. J. Audiol.* 47, S53–S71. doi: 10.1080/14992020802301142

Andruski, J. E., Blumstein, S. E., and Burton, M. (1994). The effect ofJ. Mem. Lang. subphonetic differences on lexical access. *Cognition.* 52, 163–187. doi: 10.1016/0010-0277(94)90042-6

Baese-Berk, M. M., Bradlow, A. R., and Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *J. Acoust. Soc.* 133, EL174-EL180. doi: 10.1121/1.4789864

Boduch-Grabka, K., and Lev-Ari, S. (2021). Exposing individuals to foreign accent increases their trust in what nonnative speakers say. *Cogn. Sci.* 45:11,e13064. doi: 10.1111/cogs.13064

Bradlow, A. R., and Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*. 106, 707–729. doi: 10.1016/j.cognition.2007.04.005

Brookshire, R. H., and Nicholas, L. E. (1993). *The Discourse Comprehension Test*. Tucson, AZ: Communication Skill Builders. A Division of The Psychological Corporation.

Clarke, C. M., and Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *J. Acoust. Soc. Am*. 116, 3647–3658. doi: 10.1121/1.1815131

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Connine, C. M., Blasko, D. G., and Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition?. *J. Mem. Lang.* 32, 193–210. doi: 10.1006/jmla.1993.1011

Daneman, M., and Carpenter, P. A. (1980). Individual differences in working memory and reading. *J. Verbal Learning Verbal Behav*. 19:4, 450–466. doi: 10.1016/S0022-5371(80)90312-6

Drager, K. K. (2011). Sociophonetic variation and the lemma. *J. Phon.* 39, 694–707. doi: 10.1016/j.wocn.2011.08.005

Dunton, J., Bruce, C., and Newton, C. (2011). Investigating the impact of unfamiliar speaker accent on auditory comprehension in adults with aphasia. *Int. J. Lang. Commun. Disord*. 46, 63–73.

Francis, A. L., and Nusbaum, H. C. (2009). Effects of intelligibility on working memory demand for speech perception. *Attent, Percept, Psychophysics*. 71, 1360–1374. doi: 10.3758/APP.71.6.1360

Fuertes, J. N., Potere, J. C., and Ramirez, K. Y. (2002). Effects of speech accents on interpersonal evaluations: implications for counseling practice and research. *Cult. Divers. Ethn. Minor. Psychol.* 8, 346–356. doi: 10.1037/1099-9809.8.4.347

Garman, M. (2012). *Psycholinguistics*. Cambridge: Textbooks in Linguistics Series. Available online at: https://doi-org.libproxy.ncl.ac.uk/10.1017/CBO9781139165914 (accessed December 15, 2021).

Gaskell, M. G., and Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Lang. Cogn. Process.* 12, 613–656. doi: 10.1080/016909697386646

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Rev.* 105, 251. doi: 10.1037/0033-295X.105.2.251

Goslin, J., Duffy, H., and Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain Lang.* 122:2, 92–102. doi: 10.1016/j.bandl.2012.04.017

Hailstone, J. C., Ridgway, G. R., Bartlett, J. W., Goll, J. C., Crutch, S. J., and Warren, J. D. (2012). Accent processing in dementia. *Neuropsychologia*. 50, 2233–2244. doi: 10.1016/j.neuropsychologia.2012.05.027

Hanulíková, A., Van Alphen, P. M., Van Goch, M. M., and Weber, A. (2012). When one person's mistake is another's standard usage: the effect of foreign accent on syntactic processing. *J. Cogn. Neurosci*. 24, 878–887. doi: 10.1162/jocn_a_00103

IELTS. (2007). *The IELTS Handbook*. Cambridge: University of Cambridge Local Examinations Syndicate, The British Council, IDP Australia.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y. I., Kettermann, A., et al. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*. 87, B47–B57. doi: 10.1016/S0010-0277(02)00198-1

Lev-Ari, S. (2015). Comprehending non-native speakers: theory and evidence for adjustment in manner of processing. *Front. Psychol.* 5, 1–12. doi: 10.3389/fpsyg.2014.01546

Lev-Ari, S., Ho, E., and Keysar, B. (2018). The unforeseen consequences of interacting with non-native speakers. *Top. Cogn. Sci.* 10, 835–849. doi: 10.1111/tops.12325

Lev-Ari, S., and Keysar, B. (2012). Less-detailed representation of non-native language: Why non-native speakers' stories seem more vague. *Discourse Process.* 49, 523–538. doi: 10.1080/0163853X.2012.698493

Lindemann, S. (2002). Listening with an attitude: A model of native-speaker comprehension of non- native speakers in the United States. *Language in Society*. 31, 419–441. doi: 10.1017/S0047404502020286

Lippi-Green, R. (2012). *English With an Accent: Language, Ideology, and Discrimination in the United States*. London: Routledge. doi: 10.4324/9780203348802

Luce, P. A., and Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear Hear.* 19, 1–36. doi: 10.1097/00003446-199802000-00001

McClelland, J. L., and Elman, J. L. (1986). The TRACE model of speech perception. *Cogn. Psychol.* 18, 1–86. doi: 10.1016/0010-0285(86)90015-0

Meyer, A. S., and Schiller, N. O. (2011). *Phonetics and Phonology in Language Comprehension and Production: Differences And Similarities*. Phonology Phonetics. Available online at: https://doi-org.libproxy.ncl.ac.uk/10.1515/9783110895094 (accessed December 15, 2021).

Munro, M. J., and Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Lang. Learn*. 45, 73–97. doi: 10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., and Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Lang. Speech*. 38, 289–306. doi: 10.1177/002383099503800305

Newman, R. L., and Connolly, J. F. (2009). Electrophysiological markers of pre-lexical speech processing: evidence for bottom–up and top–down effects on spoken word processing. *Biol. Psychol*. 80, 114–121. doi: 10.1016/j.biopsycho.2008.04.008

Norris, D., McQueen, J. M., and Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *J. Exp. Psychol. Learn Mem. Cogn.* 21, 1209. doi: 10.1037/0278-7393.21.5.1209

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., et al. (2013). The ease of language understanding (ELU) model: theoretical, empirical, and clinical advances. *Front. Syst. Neurosci*. 7, 1–17. doi: 10.3389/fnsys.2013.00031

Rönnberg, J., Rudner, M., Foo, C., and Lunner, T. (2008). Cognition counts: a working memory system for ease of language understanding (ELU). *Int. J. Audiol.* 47:2, 99–105. doi: 10.1080/14992020802301167

Smithson, M., and Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol. Methods.* 11, 54–71. doi: 10.1037/1082-989X.11.1.54

Van Engen, K. J. (2015). Downstream effects of accented speech on memory. *J. Acoust. Soc. Am.* 137, 2210. doi: 10.1121/1.4920046

Van Engen, K. J., Chandrasekaran, B., and Smiljanic, R. (2012). Effects of speech clarity on recognition memory for spoken sentences. *PLoS ONE*. 7, e43753. doi: 10.1371/journal.pone.0043753

Van Engen, K. J., and Peelle, J. E. (2014). Listening effort and accented speech. *Front. Hum. Neurosci*. 8, 577. doi: 10.3389/fnhum.2014.00577

Wolter, B. (2006). Lexical network structures and L2 vocabulary acquisition: the role of L1 lexical/conceptual knowledge. *Appl. Linguist*. 27, 741–747. doi: 10.1093/applin/aml036

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 02 September 2022
DOI 10.3389/fpsyg.2022.901326

OPEN ACCESS

EDITED BY
Joseph V. Casillas,
Rutgers, The State University
of New Jersey, United States

REVIEWED BY
Pui Fong Kan,
University of Colorado Boulder,
United States
Linda Polka,
McGill University, Canada

*CORRESPONDENCE
Molly Babel
molly.babel@ubc.ca

SPECIALTY SECTION
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

RECEIVED 21 March 2022
ACCEPTED 28 July 2022
PUBLISHED 02 September 2022

CITATION
Cheung S and Babel M (2022) The own-voice benefit for word recognition in early bilinguals. *Front. Psychol.* 13:901326.
doi: 10.3389/fpsyg.2022.901326

The own-voice benefit for word recognition in early bilinguals

Sarah Cheung[1] and Molly Babel[2]*

[1]Department of Speech-Language Pathology, University of Toronto, Toronto, ON, Canada,
[2]Department of Linguistics, University of British Columbia, Vancouver, BC, Canada

The current study examines the self-voice benefit in an early bilingual population. Female Cantonese–English bilinguals produced words containing Cantonese contrasts. A subset of these minimal pairs was selected as stimuli for a perception task. Speakers' productions were grouped according to how acoustically contrastive their pronunciation of each minimal pair was and these groupings were used to design personalized experiments for each participant, featuring their own voice and the voices of others' similarly-contrastive tokens. The perception task was a two-alternative forced-choice word identification paradigm in which participants heard isolated Cantonese words, which had undergone synthesis to mask the original talker identity. Listeners were more accurate in recognizing minimal pairs produced in their own (disguised) voice than recognizing the realizations of speakers who maintain similar degrees of phonetic contrast for the same minimal pairs. Generally, individuals with larger phonetic contrasts were also more accurate in word identification for self and other voices overall. These results provide evidence for an own-voice benefit for early bilinguals. These results suggest that the phonetic distributions that undergird phonological contrasts are heavily shaped by one's own phonetic realizations.

KEYWORDS

speech perception, word recognition, bilingualism, speech production, linguistic representation

Introduction

Familiar accents and voices receive a range of processing benefits including higher recognition rates, intelligibility boosts, and increased attention in the context of competing speech (e.g., Bradlow and Bent, 2008; Adank et al., 2009; Johnsrude et al., 2013; Holmes et al., 2018). One's own voice is arguably the most familiar voice, due to our continuous exposure to it. Given that self-recognition, the ability to distinguish between the self and others, is a fundamental human capability, it is therefore unsurprising that self-referential information is processed differently from stimuli associated with others across domains (Keenan et al., 2000; Platek et al., 2004, 2006; Uddin et al., 2005; Keyes et al., 2010; Devue and Brédart, 2011; Zhao et al., 2011; Liu et al., 2019). This extends to voice processing, as researchers have not only observed that people process their own voices differently from others' voices (Hughes and Harrison, 2013; Peng et al., 2019;

Mitterer et al., 2020), but also that this difference in perception may translate into an advantage in recognizing words in self-produced speech (Eger and Reinisch, 2019).

Spoken language processing is, in a large part, shaped by experience. Infants narrow their perceptual categories based on the language varieties they are exposed to (e.g., Werker and Tees, 1984), and adults prioritize phonetic information in a language-specific manner (e.g., Johnson, 1997; Sumner et al., 2014; Schertz and Clare, 2020). Familiar languages, accents, and voices are afforded benefits in processing, and these benefits surface at different intervals in the pipeline. Concepts like *recognition* (i.e., comprehending the signal) and *encoding* (i.e., updating a representation) are different processes (Clopper et al., 2016; Todd et al., 2019) and consideration needs to be given as to whether any socially skewed or preferential encoding takes place at *perception* or *interpretation* stages (see Zheng and Samuel, 2017). In addition to unpacking the mechanisms by which preferential encoding occurs, the acoustic-auditory substance of *what* is preferentially encoded is not well predicted by theory or supported by consistent empirical results. For example, while there is evidence that familiar speech signals are preferentially encoded (e.g., Clopper et al., 2016), this does not entail that the highest frequency exemplar is the most robustly encoded (Sumner and Samuel, 2005). In some cases, early and consistent experiences shape recognition (e.g., Sumner and Samuel, 2009) and perceptual processing (Evans and Iverson, 2007), whereas in other instances, socially prestigious speech may receive a boost (Sumner and Kataoka, 2013). Familiar accents typically receive benefits, but unfamiliar accents can draw perceptual attention, making them more challenging to ignore than more familiar accents (Senior and Babel, 2018).

The aforementioned examples all relate to accent or dialect differences, but familiarity effects in spoken language are not limited to that level of abstraction. Familiarity effects also extend to individual voices. A large body of research demonstrates that familiarity with a speaker's voice eases perception (Nygaard et al., 1994; Newman et al., 2001; Perry et al., 2018). For instance, Nygaard and Pisoni (1998) showed that listeners who successfully learned the voices and names of speakers were better at identifying speech produced by the speakers they were trained on compared to unfamiliar speakers. Evidence of a familiar-talker advantage in perception has been found for young and old listeners (Yonan and Sommers, 2000; Johnsrude et al., 2013), in addition to older listeners with hearing impairments (Souza et al., 2013). Familiar-talker advantages are also found with explicit (Nygaard and Pisoni, 1998) and implicit training (Kreitewolf et al., 2017), as well as in listening conditions with a competing talker in the background (Holmes et al., 2018; Holmes and Johnsrude, 2020). Listeners show improved abilities to selectively attend to or ignore very high familiarity voices (e.g., a spouse's voice; Johnsrude et al., 2013), suggesting that a relatively fine-grained prediction is available for familiar voices. Even without awareness of speaker identity, listeners encode acoustically-specific information about words, which can result in more efficient processing if it is similar to existing representations (Creel and Tumlin, 2011).

As noted, an individual's own voice is, arguably, the voice that one has most familiarity with. Importantly, however, self-voice perception of one's own voice "sounds different" from others' because of the different mediums through which sound is physically conducted during perception. When listeners hear their own voices as they speak, sound is transmitted via both air and bone conduction (Shuster and Durrant, 2003; Reinfeldt et al., 2010). In air conduction, vibrations exit the oral cavity, travel through air and enter the ear canal, whereas in bone conduction, vibrations move through the skull bone directly to the cochlea (Stenfelt and Goode, 2005). Comparatively, when listeners hear others speak or hear their own voice in recordings, sound is conducted solely via air conduction. Despite these differences, listeners are very successful at recognizing their own productions in recordings (Xu et al., 2013). Xu et al. (2013) presented listeners with recordings of their own voices and the voices of other, familiar speakers in normal and difficult listening conditions. They found that even in high-pass filter conditions that removed acoustic information from the mean of an individual's third resonant frequency and above, listeners were able to identify their own voices. Researchers theorize that auditory familiarity with one's own voice and the association between auditory self-representation and motor representations may contribute to this self-recognition advantage (Xu et al., 2013).

Beyond an advantage in own-voice recognition, speakers monitor their own productions through auditory feedback. Delayed auditory feedback induces an increase in foreign accent for second language learners (Howell and Dworzynski, 2001) and an increase in regional accent for those who have acquired a different accent (Howell et al., 2006). This suggests that when the timing of auditory feedback is perturbed, individuals are unable to monitor their speech as effectively, resulting in a shift in their speech patterns. Real-time shifts in auditory feedback, where an individual hears resynthesized versions of their own productions that deviate from what they produced, elicits compensation to account for the synthesized acoustic shift (e.g., Houde and Jordan, 2002; Jones and Munhall, 2002; Purcell and Munhall, 2006; Katseff et al., 2012). Crucially, the magnitude of an individual's compensatory response is associated with the shifted item's position in the vowel space; shifted items that fall near a phonetic category boundary elicit a larger compensatory response (Niziolek and Guenther, 2013). Compensation for auditory feedback appears to be generally heightened for linguistically relevant dimensions (Xu et al., 2004; Chen et al., 2007; Mitsuya et al., 2011; Niziolek and Guenther, 2013).

While one's own auditory feedback is valuable to the control of motor actions in speech, do one's own productions provide a recognition advantage at the word level? Word recognition can be considered a process that serves to comprehend the speech of *others*, as, under normal contexts, an individual is aware of the linguistic message that is emitted from their own vocal tract. We are interested in how own-voice familiarity shapes the representational and recognition space for linguistic contrasts in word recognition and the acoustic-phonetic distributions that implement phonological contrasts. To test how one's own implementation of a contrast affects word recognition, an introduction of some kind of adverse listening condition is required, as identifying words in a familiar language is a fairly trivial task. Scholars have approached this from two angles – with second-language (L2) learners or first language listeners – each of which has used relatively distinct methods and landed on different conclusions.

From the L2 perspective is Eger and Reinisch (2019), who demonstrated that German-speaking learners of English were better at recognizing self-produced words in English. This suggests that L2 language learners prioritize their own realizations of phonological contrast. In a related study, Mitterer et al. (2020) show that German-speaking learners of English rate their own, in this case, vocally disguised, sentence productions as more target-like. Mitterer and colleagues offer the interpretation that it is the comprehension advantage afforded by one's own voice that supports higher ratings for self-produced sentences. However, these results for L2 language learners contrast with claims made when processing a first language. For an individual's first language, there is a reported benefit to processing the most statistically average voice over their own self-produced voice when listeners are asked to identify noise-vocoded words, a manipulation that removes fine spectral detail, but spares temporal cues and amplitude modulation (Schuerman et al., 2015, 2019). There is, however, some evidence that L1 listeners' word recognition in sentences masked with speech-shaped noise shows a benefit for self-produced sentences compared to sentences produced by others (Schuerman, 2017). Schuerman et al. (2015, 2019) suggest that listeners' preferred linguistic representations are informed by the input perceived in one's speech community — hence the improved recognition for the statistically average voice in noise-vocoded speech. They reason that own-voice preferences may only arise when listeners are aware that they are hearing their own voice, which is challenging in noise-vocoded speech. The mechanism for the own-voice benefit for L2 English learners posited by Eger and Reinisch (2019) presumes that an individual recognizes their own voice and then further perceptually adapts to their own productions.

In the current study, we test the own-voice benefit for word recognition in early bilinguals, leveraging the high levels of natural phonetic ambiguity in a heterogenous multilingual population of Cantonese–English speakers. We test whether these early bilinguals, like second language learners, show an own-voice benefit in word recognition. Moreover, we probe whether the own-voice benefit indeed hinges upon recognition of one's own voice. Following prior work (Holmes et al., 2018; Mitterer et al., 2020), some cues to talker identity are manipulated by shifting f0 and formant frequencies (using Praat; Boersma and Weenink, 2020) to limit listeners' ability to recognize their own voices. This methodology draws on the observation that manipulating these cues greatly affects the success of self-voice recognition (Xu et al., 2013).

Materials and methods

The experiment consisted of three parts: a questionnaire about multilingualism, a production task, and a perception task, all of which were completed remotely on participants' own electronic devices. All written and verbal instructions were presented in English to accommodate limited Cantonese literacy within the bilingual population at our university.

Participants

To be eligible for this study, participants were required to self-identify as female, be exposed to both Cantonese and English at or before the age of six, and minimally have the ability to carry out a basic conversation in Cantonese. Only female subjects were invited to participate to minimize between-speaker variation and to allow a more consistent vocal disguise technique (see description of audio manipulation below). Thirty-six female Cantonese-English bilinguals participated in the experiment. While all participants completed the multilingual questionnaire and the production task, the recordings of three participants obtained during the production task were excluded from the perception task due to poor recording quality and interference from background noise. In addition, two participants who completed the production task and questionnaire did not complete the perception task, resulting in 31 subjects who completed all three parts of the study. **Appendix Table A1** provides selected summary language information for the 33 participants who completed the production task and for whom a perception experiment was designed. **Appendix Table A2** contains additional demographic information about the participant population. Participants reported their languages in order of current self-assessed dominance, along with the age of acquisition of each language, and speaking, listening, and reading proficiencies on a scale from 0 (none) to 10 (perfect). The population is highly multilingual, as is typical of both Cantonese speakers in Cantonese-speaking homelands (e.g., Hong Kong, Guangzhou) and those in the Cantonese-speaking diaspora, which is the convenience sample used in the current study. For example, 27 participants report Mandarin as an additional language, and 16 report French, in addition to small numbers of individuals self-reporting knowledge of other languages. Participants' self-reported ages of acquisition indicate that Cantonese was the earliest acquired language (Median = 0, *SD* = 1.3), compared to English (Median = 3, *SD* = 1.9), and Mandarin (Median = 6, *SD* = 4) and French (Median = 9, *SD* = 2.7), the other two most attested languages amongst participants. Participants self-reported significantly higher speaking and listening proficiencies for English (speaking: *M* = 9.3, *SD* = 0.98; listening: *M* = 9.48, *SD* = 0.83) compared to Cantonese [speaking: *M* = 7.15, *SD* = 2.36; listening: *M* = 7.82, *SD* = 1.96; paired *t*-test for speaking: *t*(32) = 4.38, *p* = 0.0001; paired *t*-test for listening: *t*(32) = 4.11, *p* = 0.0003]. Mandarin was the language with the next highest self-reported proficiency across participants, though it was not a language reported by all participants, and self-reported speaking [unpaired *t*-test: *t*(54) = –2.65, *p* = 0.01] and listening [unpaired *t*-test: *t*(54) = –2.7, *p* = 0.009] skills were higher for Cantonese than Mandarin (speaking: *M* = 5.5, *SD* = 2.5; listening: *M* = 6.4, *SD* = 2.1). Participants' current place of residence was in English-dominant communities in Canada and the United States, as shown in **Appendix Table A2**.

Participants were compensated with gift cards equivalent to $5 CAD for the production task, $5 CAD for the questionnaire, and $10 CAD for the perception task. Participants were recruited through the UBC community and social media.

Materials

Multilingual language questionnaire

Participants completed an online survey that presented questions from the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007) and the Bilingual Language Profile (BLP; Gertken et al., 2014). Both resources were designed to gain a better understanding of language profiles of bilingual and multilingual speakers by including questions relating to individuals' language history, usage, attitudes and self-rated proficiency. Additionally, general questions pertaining to participants' biographical information were included in this questionnaire. This survey was administered in English.

Production stimuli

Stimuli for the production task included monosyllabic Cantonese words, presented as pictures accompanied by English translations. All pictures were hand-drawn by the researcher and presented in black and white so that no single picture was especially salient to subjects (see **Appendix Figure A2** for the complete set of visual stimuli). The word list was composed of 22 minimal pairs targeting seven segmental contrasts (see **Appendix Table A3** for the complete production word list). Three of the lexical items served as minimal pair to more than one other item, hence the number of unique words totaled 41 (and not 44) for the 22 minimal pairs. The lexical items involved word initial consonants /t͡s/, /t͡sʰ/, and /s/ and vowel contrasts /ɐ͡i/ and /e͡i/, /ɔː/ and /o͡u/, /ɐ͡i/ and /aː͡i/, /ɐ͡u/ and /aː͡u/, and /ɐ/ and /a:/. Target sounds were selected based on their presence in Cantonese and absence in English such that the selected contrasts would show variability across proficiency ranges in the Cantonese-English bilingual community. For example, three of the vowel contrasts chosen are distinguished by vowel length, a feature that is not lexically contrastive in English. The stimuli were designed to consist of all high level tone (T1) words to control for differences in tone that may cause unwanted variability in production or confusion in perception task performance. The words were chosen to be familiar to Cantonese speakers with potentially limited vocabularies due to largely using Cantonese as a home language in an English-dominant region and had meanings that could be easily represented in pictures. Pictures, as opposed to Chinese characters, were used both in the production and perception tasks to accommodate participants who have limited literacy skills.

Perception stimuli

A subset of the stimuli words used in the production task were featured in the perception task. These consisted of 13 minimal pairs featuring five vowel contrasts: /ɐ͡i/ and /e͡i/, /ɔː/ and /o͡u/, /ɐ͡i / and /aː͡i/, /ɐ͡u/ and /aː͡u/, and /ɐ/ and /a:/, which are presented in their character and Jyutping transliterations and English glosses in **Table 1**. The same pictures corresponding to these target words from the production experiment were used in the perception task. The manipulation of the audio stimuli for the perception experiment is described below.

Production task

Procedure

For the production task, participants first watched a video tutorial (made by the first author) on how to record themselves producing the list of target words. This video included a familiarization phase for participants to learn the intended referents of the picture stimuli. For each target word, participants would hear a Cantonese word and see its corresponding picture and English translation. Afterward, participants were instructed to download Praat (Boersma and Weenink, 2020) and record themselves using the built-in microphone of their personal electronic devices at a sampling frequency of 44,100 Hz. Participants accessed a .pdf file containing the picture stimuli and were asked to verbally label the target words in Cantonese, given the picture and English translation as they proceeded through the randomized list at their own pace. Each picture was shown twice to elicit two productions of each word, for a total of 82 productions. Lastly, participants were asked to verbally describe a picture

TABLE 1 *Perception Stimuli* arranged by minimal pair.

Chinese Character	English Gloss	Jyutping Romanization	Chinese Character	English Gloss	Jyutping Romanization
雞	chicken	gai1	機	machine	gei1
雞	chicken	gai1	街	street	gaai1
揮	to wave	fai1	飛	to fly	fei1
多	many	do1	刀	knife	dou1
歌	song	go1	高	tall	gou1
梳	comb	so1	鬚	beard, moustache	sou1
波	ball	bo1	煲	pot	bou1
踎	to squat	mau1	貓	cat	maau1
秋	autumn	cau1	抄	to copy	caau1
咯	cough	kat1	咭	card	kaat1
心	heart	sam1	衫	shirt	saam1
西	west	sai1	嘥	to waste	saai1
龜	turtle	gwai1	乖	well-behaved	gwaai1

Note that 雞 *chicken* is used in two minimal pairs.

of a busy park scene in Cantonese, in as much detail as they wanted. Participants saved their recordings according to their anonymous participant ID number and uploaded their recordings to Dropbox.

Segmentation

Words of the minimal pairs were segmented from recordings using Praat (Boersma and Weenink, 2020). Recordings from three participants were excluded from this process due to poor recording quality. From the productions of the remaining 33 speakers, nine speakers had at least one word excluded for a total of 15 words excluded from analyses due to incorrect labeling of the picture stimuli. The removal of one item entailed the removal of two, as the minimal pair was removed from that individual's set.

Because stimuli words were produced in isolation, word-initial stops /b/, /d/, /g/, /k/ and /k^{w}/ were identified as beginning with the stop burst, starting as an abrupt change in amplitude in the waveform and ending with the onset of quasi-periodic activity of the following vowel. The offset of the labialized voiceless velar stop /k^{w}/ was identified as a change in the waveform from a simpler periodic pattern to a more complex periodic pattern of a vowel. In this set of stimuli, the only word-final stop was /t̚/. The end boundary of this unreleased stop was identified as the same point as the end of its preceding vowel. Fricatives /s/ and /f/ were identified in waveforms as aperiodic or random patterns indicating frication noise. Affricates /t͡s/ and /t͡s^{h}/ were identified as beginning with a stop burst and ending with the offset of frication noise, signaling the end of the fricative. Aspirated alveolar affricates showed a period of high amplitude frication followed by a period of lower amplitude frication and the boundaries for aspiration were annotated using low amplitude frication as a cue. One participant produced target words intended to contain word-initial aspirated alveolar affricates with voiceless fricatives instead. For these productions, the onset and offset of the aspirated alveolar affricate /t͡s^{h}/ were marked at the same points as the beginning and end of aspiration shown in the waveform. The onset of nasals /m/, /n/ and /ŋ/ were identified at the point of a most discrete change in amplitude in the waveform. The offset of the nasal consonants in word-initial position were indicated by a sudden increase in intensity at the beginning of the following vowel. Another cue used to identify this boundary was the change from a simple waveform pattern with lower frequencies, characteristic of nasal consonants, to a more complex pattern with both high and low frequencies, characteristic of vowels. Likewise, the opposite change in intensity and opposite shift in waveform patterns indicated boundary of the word-final nasal /ŋ/. All word and sound boundaries were placed as closely as possible to zero crossings to prevent auditory distortions resulting from discontinuities at the beginnings and ends of sound intervals. Words in all 22 minimal pairs were segmented, although only the subset of words comprising 13 minimal pairs were used in the perception task. Target words were saved into their own files, while target sounds were trimmed into files with 25 ms buffers at the onset and offset of sounds in preparation for acoustic analysis.

Grouping voices

Acoustic analyses served to group minimal pairs into five groups (Groups A, B, C, D, and E) reflecting how discretely speakers produced the contrast between the two words of each minimal pair. We will refer to this measure as "contrastiveness," as it denotes the acoustic difference between target sounds in minimal pairs, but does not necessarily imply speaker proficiency or production accuracy. Because of the considerable amount of individual variation observed between minimal pairs within vowel contrasts, a given talker's group assignment was done separately for each minimal pair. This means that a speaker was not, for example, categorized as a Group A speaker, but her productions for a particular minimal pair may have been assigned to Group A, while her productions for other minimal pairs may be in another contrastiveness Group.

To determine contrastiveness we first estimated formant trajectories with samples every two seconds for each vowel using Fast Track (Barreda, 2021), a formant tracker plug-in via Praat (Boersma and Weenink, 2020). The frequency range was set at 5,000–7,000 Hz to reflect a speaker of “medium height” (Barreda, 2021), as all participants in our study were female adults.

Formant trajectories were then converted from Hertz to the Bark scale to better reflect auditory processing (Traunmüller, 1990). With the obtained Bark-scaled formant trajectories, we then performed a discrete cosine transform (DCT) which yielded three primary coefficients for F1 and F2. The three coefficients corresponded to the mean of the formant, the slope of the formant and the curvature of the slope. In addition to these six dimensions, we also measured vowel duration as a seventh dimension in which speakers could potentially show distinctiveness in production. While not all seven dimensions may be used to contrast the target vowels in our minimal pairs, we did not exclude any particular parameter to avoid making any *a priori* claims about the relative importance of these cues for contrastiveness for this bilingual population. We centered, scaled and calculated Euclidean distances for each talker’s minimal pair along all seven dimensions.

Lastly, for each minimal pair, we organized speakers according to the contrastiveness of their productions. This was done by ranking the Euclidean distances for each minimal pair and using the rankings of each to form minimal pair-specific group assignments, in which a greater Euclidean distance indicated a more distinctive production. Within each minimal pair, we formed five groups, ranging from A (most contrastive) to E (least contrastive), consisting of five to seven different voices; thus, for each minimal pair, each group had 5-7 different voices. The groups were manually adjusted to be approximately equally sized, as some talkers were missing tokens and therefore would not be presented with that particular minimal pair in their individualized perception experiment. **Figure 1** is a box-and-whisker plot presenting the phonetic distance or contrastiveness range for the productions in each of the five contrastiveness groups.

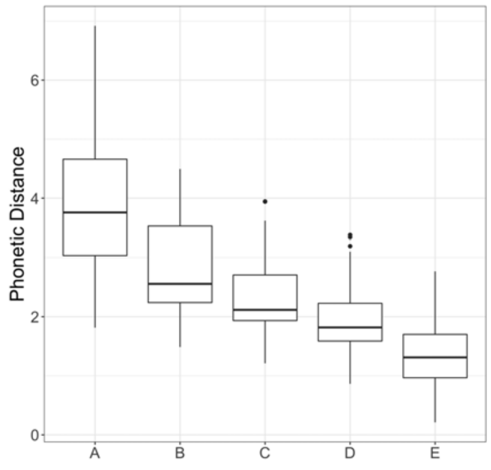

FIGURE 1
Box-and-whisker plot of phonetic distance between minimal pairs for utterances in the five contrastiveness groups.

Each subject was presented with a perception experiment, described below, featuring their own productions and the productions of other members of their contrastiveness group, for each minimal pair. Therefore, the number of different unfamiliar voices heard by each participant varied according to their group memberships.

Perception task

Audio manipulation

For the perception experiment, recordings segmented into isolated words were altered to change female voices into male-like voices using the Change-Gender function in Praat (Boersma and Weenink, 2020). This application lowered the fundamental frequency (f0) and formant frequencies of the original productions by multiplying these dimensions by factors specific to each speaker. Modulation of these parameters have been shown to influence the accuracy of self-voice recognition (Xu et al., 2013) and previous studies have successfully disguised voices using the Change-Gender function (Holmes et al., 2018; Mitterer et al., 2020). For speakers in the current study, the multiplication factors for f0 and formant frequencies ranged from 0.55 to 0.75 (mean = 0.62) and 0.79 to 0.83 (mean = 0.81) respectively. Pitch range parameters were adjusted as necessary to ensure accurate pitch tracking. Following Mitterer et al. (2020), the manipulations started with scaling the f0 by 0.59 and the formants by 0.82, which were the average manipulations made by Mitterer et al. (2020). From there, the actual values for each talker were adjusted by ear to achieve a good-sounding disguise. The specific by-talker adjustments are reported in **Appendix Table A4**. Finally, the target stimuli were RMS-amplitude normalized to 65 dB and mixed in continuous speech-shaped noise, created from the spectral profiles of the participants’ speech samples, at a signal-to-noise ratio (SNR) of +5 dB to increase the difficulty of the task. This particular SNR was determined through piloting to achieve high accuracy, but prevent ceiling performance.

Procedure

The same speakers who completed the production task were invited to complete the perception task several months later, which was administered online using jsPsych (de Leeuw, 2015).

This perception experiment was a two-alternative forced-choice lexical identification task featuring the acoustically altered recordings described above. For each trial, participants heard an isolated Cantonese word produced either by themselves or another speaker along with two pictures on the left and right sides of the screen, representing the appropriate Cantonese minimal pair. Participants were required to choose the picture corresponding to the word they heard by pressing the keys "F" or "J" for the left and right sides of the screen, respectively. Participants' responses advanced the program to the next trial. Three practice trials were provided. Audio stimuli were presented at a comfortable listening level and participants completed a headphone check prior to beginning the experiment (Woods et al., 2017). There were four repetitions of each token for a total of 560–688 trials for each participant's personalized experiment [up to 26 items (e.g., 13 minimal pairs) × a range of 5–7 speakers in each by minimal pair group × 4 repetitions of each token]. Trials were fully randomized across four blocks between which participants were offered a self-paced break. At the end of the experiment, participants were asked if they recognized their own voice throughout the experiment, to which they selected "yes" or "no" on the screen. The perception experiment was completed on participants' own electronic devices and took approximately 35–40 min to complete. Participants were asked to complete the task in a quiet place.

Results

To remove extremely fast and extremely slow responses, button presses logged under 200 ms and over 5000 ms were removed from the data, eliminating just under 2% of responses. Participants' responses on the perception task were scored as either correct or incorrect depending on whether listeners chose the picture corresponding to the intended word. These accuracy data were analyzed using a Bayesian multilevel regression model in Stan (Gabry and Češnovar, 2021) using brms (Bürkner, 2018) in R (R Core Team, 2021). The accuracy of each response (correct word identification or not) was analyzed as the dependent variable with Voice Match (other voice, own voice), Trial number (centered and scaled), and Contrastiveness Group (Groups A–E) as independent variables. Voice Match and Group, Trial and Group, and Trial and Voice Match were included as interactions. There were random slopes for Voice Match and Trial by participant. Given that most items were other voice items, Voice Match was treatment coded (with Other Voice as the reference level) and Contrastiveness Group was forward-difference coded using the coding matrices package (Venables, 2021), which compares each level in Contrastiveness Group to the adjacent level. The model family was Bernoulli and we specified weakly informative normally distributed priors that were centered at 0 for the intercept and population-level parameters. The intercept and population-level parameters had standard deviations of 5 and 2.5, respectively, following recommendations for accuracy data in Coretta (2021). Correlations used the LKJ prior with a value of 2. The models were fit with 4000 iterations (1000 warm-up) with four chains for the Hamiltonian Monte-Carlo sampling. All R-hat values were below 1.01 and Bulk ESS values were all high, suggesting the model was well mixed. The median posterior point estimates and the 95% credible interval (CrI) is reported for all parameters and interactions. An effect is considered compelling if 95% of the posterior distribution for a parameter does not include 0. An effect is considered to have weak evidence if the credible interval includes 0, but the probability of direction is at least 95%. These interpretation practices follow recommendations in Nicenboim and Vasishth (2016).

The model results are reported in **Table 2**. The intercept indicates that listeners were very good at the task, reliably identifying the intended lexical item [β = 1.66, 95% CrI = [1.32, 2.02], Pr(β > 0) = 1]. The model results provide compelling evidence for a benefit in processing one's own (disguised) voice [β = 0.23, 95% CrI = [0.06, 0.42, Pr(β > 0) = 99.5%]. This result is visualized in **Figure 2**, which presents the fitted draws from the posterior fit of the model for the own-voice effect by Contrastiveness Group.

An effect of trial suggests that listeners' accuracy improved across the course of the experiment [β = 0.07, 95% CrI = [0.01, 0.14], Pr(β > 0) = 98.22%]; the CrI for all interactions of Trial with the Contrastiveness Group contrasts overlap substantially with 0, suggesting that this cross-experiment improvement was not specific to a particular Group. The Voice Match by

TABLE 2 Summary of the posterior distribution modeling word recognition accuracy with posterior means and the 95% Credible Interval, along with the probability of direction for each effect.

Parameter	β	95% CrI	Probability of direction
Intercept	1.66	[1.32, 2.02]	100%
Voice Match (Own Voice)	0.23	[0.06, 0.42]	99.5%
Trial	0.07	[0.01, 0.14]	98.22%
Group A vs. B	−0.21	[−0.36, −0.06]	99.72%
Group B vs. C	0.27	[0.13, 0.41]	100%
Group C vs. D	0.21	[0.07, 0.34]	99.84%
Group D vs. E	0.26	[0.13, 0.39]	100%
Voice Match × Group A vs. B	0.31	[−0.09, 0.70]	93.69%
Voice Match × Group B vs. C	−0.41	[−0.77, −0.04]	98.60%
Voice Match × Group C vs. D	0.31	[−0.04, 0.68]	95.55%
Voice Match × Group D vs. E	−0.04	[−0.37, 0.28]	60.03%
Trial × Group A vs. B	0.08	[−0.06, 0.22]	86.60%
Trial × Group B vs. C	−0.04	[−0.17, 0.09]	73.05%
Trial × Group C vs. D	−0.03	[−0.15, 0.09]	68.46%
Trial × Group D vs. E	−0.04	[−0.15, 0.08]	73.08%
Voice Match × Trial	0.03	[−0.10, 0.17]	65.33%

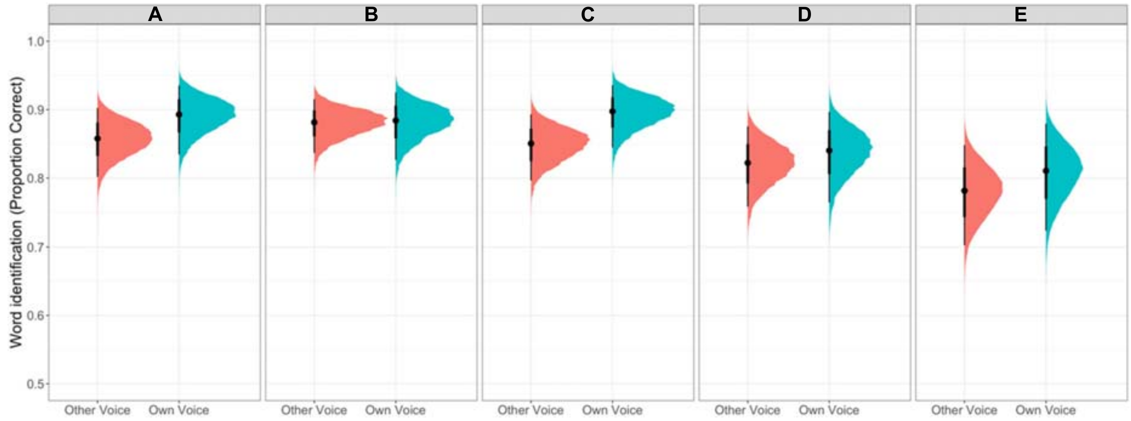

FIGURE 2
Proportion of correct responses in the perception task for the five acoustic contrastiveness groups presented as fitted draws from the posterior fit of the model. Panels **A–E** represent the five contrastiveness groups from most contrastive **(A)** to least contrastive **(E)**. Responses to both own voice and other voices are included.

Trial interaction also overlapped with 0, indicating there is no evidence that the improvement in word recognition across the course of experiment was better or worse for one's own voice or other voices.

Comparisons of adjacent Contrastiveness Groups generally present compelling evidence that higher proficiency groups perform more accurately on the word identification task [Group B vs. C: $\beta = 0.27$, 95% CrI = [0.13, 0.41], $\Pr(\beta > 0) = 100\%$; Group C vs. D: $\beta = 0.21$, 95% CrI = [0.07, 0.34]; $\Pr(\beta > 0) = 99.84\%$; Group D vs. E: $\beta = 0.26$, 95% CrI = [0.13, 0.39], $\Pr(\beta > 0) = 100\%$] with the exception of Group B outperforming Group A [$\beta = -0.21$, 95% CrI = [–0.36, –0.06], $\Pr(\beta < 0) = 99.72\%$]. Two interactions involving Voice Match and Group merit attention. There is compelling evidence for an effect that Group B showed less of an own-group advantage than Group C [$\beta = -0.41$, 95% CrI = [–0.77, –0.04], $\Pr(\beta < 0) = 98.60\%$] and there is weak evidence that Group D showed less of an effect than Group C [$\beta = 0.31$, 95% CrI = [–0.04, 0.68], $\Pr(\beta > 0) = 95.6\%$].

Discussion

This experiment tested an own-voice advantage for word recognition in Cantonese for Cantonese-English early bilinguals. Words were presented in speech-shaped noise at +5 dB SNR to make the task challenging enough to inhibit ceiling performance. Listeners were more accurate at identifying difficult vowel contrasts if they were (vocally disguised) self-produced items compared to items produced by other individuals who manifested the phonological contrast to a similar degree. This was true despite an individual's own voice being disguised, suggesting that the own-voice word recognition benefit leverages linguistic representations that exist in a normalized representational space, as opposed to relying on an exact acoustic-auditory match to one's natural acoustic patterns. Items were organized by the degree of phonetic distance for the phonological contrast into what are labeled contrastiveness groups. There was strong evidence that Group C showed more of an own-voice benefit than Group B and weak evidence that Group C showed a greater own-voice benefit than Group D. Group B was exceptional in stepping out of the anticipated order in overall accuracy. While it was generally the case that groups with higher contrastiveness performed more accurately on the word identification task, Group B out-performed Group A, the highest contrastiveness group. A possibility for why those in Group B were so outstanding may relate to imperfection in our method of calculating acoustic distance, which included acoustic dimensions that are likely not core cues to contrast, though this is speculation. We note that the overall pattern was that the own-voice benefit was robust across contrastiveness groups and word recognition accuracy decreased as contrastiveness was reduced. The contrastiveness groups relate to the degree of distinctiveness of speakers' productions, which in turn may relate to speaker proficiency. Like the finding in Eger and Reinisch (2019), however, the own-voice benefit does not seem to hinge on proficiency.

Word recognition accuracy improved over the course of the experiment with participants' own voices and other voices. Although subjects heard their own voice more often than any single other voice in the experiment, the proportion of correct responses increased across trials for all voices. Altogether, this suggests that the observed self-advantage was not simply due to listeners hearing their own voice more than other voices throughout the task. The improvement across the experiment was likely due to participants adapting to the noise, which masked the speech to inhibit ceiling performance.

Our ability to determine whether listeners explicitly heard their own voice was based on an explicit self-assessment. A subset of participants reported hearing their own voice in the experiment (n = 9), but we cannot (a) confirm that positive responses to this question were not a function of positive response bias or (b) rule out that other listeners did not implicitly hear their own voices. While we follow previous work in our implementation of the voice disguise (Holmes et al., 2018; Mitterer et al., 2020), an individual's voice identity is available in other spectral and temporal patterns. Speakers vary in terms of their unique voice profiles (Lee et al., 2019; Johnson et al., 2020) and listeners exploit different acoustic cues for talker identification (Van Lancker et al., 1985; Lavner et al., 2000). Schuerman et al. (2015, 2019) did not find support for an own-voice advantage within an individual's first language when presenting noise-vocoded speech, a type of degradation in which many spectral cues important to talker identification are severely reduced, though Schuerman (2017) finds some evidence for an own-voice benefit for word recognition in sentences for speech in noise, which better retains talker-specific information. The removal of expected cues to speaker identity does not explain the absence of an own voice-benefit in those studies, however, as voice recognition and speech recognition are separate, but connected systems (for an overview see Creel and Bregman, 2011). Listeners show an intelligibility benefit for familiar voices even when those voices are made unfamiliar, indicating that the familiarity benefit does not rely on explicit recognition of a voice (Holmes et al., 2018).

The prevalent theory in voice representation is that talkers' voices are represented according to prototypes. According to the prototype theory, each stimulus is compared to a representative or central member of its category; stimuli that better approximate the prototype will be more easily perceived as belonging to the category (Lavner et al., 2001). Under this interpretation, talker identification relies on the storage and retrieval of identities based on a set of features deviating from the prototype. As previous studies have shown, the acoustic dimensions used to characterize different voices are often talker-specific (Van Lancker et al., 1985; Lavner et al., 2000). Voices that deviate more from the prototype are perceived as more distinct and thus, the more distant a speaker's acoustic features are from the central model, the easier the speaker is to be identified (Lavner et al., 2001; Latinus et al., 2013). This may partially explain the variance in participants' self-reports of hearing their own voices in the current study despite our attempt to disguise vocal identity. Those who successfully identified themselves may have had voices that deviated more from the average template and were therefore easier to recognize. Researchers have proposed that the prototype is an average, commonly encountered, yet attractive voice (Lavner et al., 2001; Latinus et al., 2013; Lavan et al., 2019). Accordingly, this voice should be representative of the listeners' language input and environment, and people of the same linguistic community would be expected to share a similar template (Lavner et al., 2001). The implications for having a voice that approximates listeners' community prototypes with regards to a benefit in word recognition needs to be explored further. In Schuerman et al. (2015, 2019) studies, researchers identified a statistically average speaker among the subjects in their studies to represent the average of the linguistic community. When presented with noise-vocoded speech, native Dutch listeners in their studies showed better recognition of words produced by the statistically average speaker in their sample than the listeners themselves. This implies that the benefit of a prototypical voice may extend beyond the benefit of hearing one's own voice for word recognition.

The core finding in the current work is that listeners were more accurate in recognizing minimal pairs produced in their own (disguised) voice than recognizing the realizations of other speakers who maintain similar degrees of phonetic contrast for the same minimal pairs. These findings with Cantonese-English bilinguals, a population which was targeted to leverage the heterogeneity in pronunciation variation within a native speaker population, replicating and extending the findings for second language learners (Eger and Reinisch, 2019). We present evidence of an own-voice benefit for work recognition, like Eger and Reinisch, but this benefit is seen when voices were disguised and the majority of individuals did not report consciously recognizing their own masked voice.

Crucially, the own-voice advantage in word recognition suggests that the phonetic distributions that undergird phonological contrasts are heavily shaped by one's own phonetic realizations, extending the importance of self-produced items beyond real-time self-monitoring (e.g., Howell et al., 2006; Niziolek and Guenther, 2013). Online compensation for altered auditory feedback indicates that auditory self-monitoring leads to immediate, though incomplete, adjustments in speech production. Importantly, the magnitude of these adjustments is yoked to whether the auditory feedback suggests a linguistic contrast is threatened (Niziolek and Guenther, 2013). This suggests a coupled relationship between perception and production where an individual's representational space for perception and recognition align with the distributional pool available for that individual in production. Many frameworks posit some degree of connection between perception and production with theoretical models differing in terms of how parsimonious perception and production repertoires are, amongst other theoretical differences related to the actual representational space (e.g., Liberman and Mattingly, 1985; Fowler, 1996; Johnson, 1997; Goldinger and Azuma, 2004). Certainly, listeners' abilities to perceive phonetic detail is connected to their abilities to produce contrasts (e.g., Werker and Tees, 1984), but does not wholly limit it (e.g., Schouten et al., 2003). Listeners are well attuned to the distribution of phonetic variation within their speech communities, particularly when that phonetic variation has social value (e.g., Johnson et al., 1999; Hay et al., 2006; Munson et al., 2006; Szakay et al., 2016). A fully isomorphic production and perception system fails to account for how listeners adapt to novel input from other speakers without concomitantly changing their own productions (Kraljic et al., 2008). If perception and production exclusively relied on perfectly mapped mental representations, the reorganization of phonetic space or changes in the weighting of acoustic cues due to perceptual learning should also be observed in that individual's productions, but this is not well supported in the existing literature (Schertz and Clare, 2020).

What mechanism accounts for the own-voice benefit? One possibility is that the mere constant auditory exposure to one's own voice, despite the fact that an individual need not attend to their own speech for the purpose of comprehension, bestows such a high level of familiarity that it is privileged in recognition space. Alternatively, it is plausible that the way in which an individual produces a contrast is intimately tied to the way in which the contrast is realized by their most frequent interlocutors such that this manifestation of the contrast — realized by the most familiar voices and one's own — receives a recognition benefit. This explanation seems unlikely, however, given that second language learners (Eger and Reinisch, 2019) and our early bilingual population show the same own-voice benefit. A third possibility is that while, as described above, perception and production cannot be isomorphic, the yoking of an individual's speech production repertoire and that repertoire's mapping in the perceptual space is what benefits an individual's own-voice productions in recognition. This is also an interpretation offered for own-voice recognition by Xu et al. (2013), who suggest that own-voice auditory and motor representations are connected. The representation of perception and action in shared space is at the heart of the common coding hypothesis (Prinz, 1997). Assuming a shared representational space for perception and production, the common coding theory predicts that listeners compare incoming speech signals to their own productions. Therefore, in perceiving one's own voice, recognition is facilitated because the auditory signal aligns with the listeners' own productions to a greater degree. Support for this in the recognition space comes from speech-reading. Individuals are better at keyword recognition in sentences when speech-reading silent videos of themselves compared to others (Tye-Murray et al., 2013), in addition to receiving more of an audio-visual boost in noisy conditions with their own videos (Tye-Murray et al., 2015). If a shared representational space accounts for the own-voice benefit, it apparently must be part of a developmental trajectory, however, as Cooper et al. (2018) find no evidence for an own-voice benefit (or an own-mother voice) benefit for word recognition in 2.5 year olds (see also Hazan and Markham, 2004). Toddlers are better at recognizing any adult production (their own mother or a different mother) than recognizing self-produced words or words from another toddler. Infants, however, already use sensorimotor information in speech perception. English-acquiring six-month olds' abilities to perceive retroflex and dental stop contrasts is inhibited when a soother blocks tongue movement [Bruderer et al., 2015; see also Choi et al. (2019) for more evidence about the connection between sensorimotor and perceptual processing in infants]. These sets of results suggest that phonemic perception and word-level recognition have different developmental trajectories with respect to the integration of motor and auditory/acoustic information streams. Ultimately, the current study cannot adjudicate between these explanatory mechanisms, but rather provides additional evidence of an own-voice benefit in adult word recognition (Tye-Murray et al., 2013, 2015; Eger and Reinisch, 2019). Multiple threads in the literature do seem to suggest that the integration of production and perceptual representations offers promise in terms of explanatory force.

The proposed mechanism that supports an own-voice benefit in word recognition — the integration of motor and acoustic-auditory representations in the linguistic representations used for word recognition — is not intended to be unique to L2 speech processing (e.g., Eger and Reinisch, 2019) or the processing of one's less dominant language (e.g., the current work). It may simply be easier to observe the evidence of an own-voice benefit in individuals' non-dominant language(s) because it may be more error prone. Individuals' native or dominant languages also, of course, exhibit within- and cross-talker variability (e.g., Newman et al., 2001; Vaughn et al., 2019). It is important to note that while there is strong statistical evidence in support of an own-voice benefit in the current work, the effect is small. An own-voice benefit is also not mutually exclusive with a benefit for a typical voice that represents the prototype or central tendency of the local speech community (e.g., Schuerman, 2017). While listeners are highly adaptable, leveraging any available information in the signal to recognize words, it is important that work in this area use spectrally rich speech samples, as some adverse listening conditions, like noise-vocoded speech, do not encode the full array of spectral information listeners typically have access to in spoken language processing. A degraded signal may encourage listeners to engage in different processing strategies.

While the own-voice benefit for word recognition was statistically robust, some participants did appear to perform less accurately on their own voices. If some aspects of word recognition are related to community averages or prototypes, these individual differences could be accounted for by considering how distant a particular individual is from the prototype. For example, participants exemplifying a self-benefit may better approximate the prototype, while those performing worse with their own voices may deviate more from the prototype relative to other speakers in their group. This reasoning aligns with the Schuerman et al. (2015, 2019) explanation for the benefit bestowed by the statistically average voice. A shared representation for an average speaker in a heterogenous bilingual population presents a challenge, however. In multilingual speech communities where individuals vary in proficiency and language use patterns, which voices are used to form prototypes for which languages? That is, are there separate prototypes, for example, for apparent native speakers of Cantonese and apparent native speakers of English, with separate prototypes established for individuals whose voices suggest a variety of Cantonese-accented English or English-accented Cantonese? What is the representational space for a speaker who experiences speaking and listening to all of these codes in different contexts? We note the nebulous nature of this space, not to discount its importance, but rather to encourage further research that can tackle the complexities in phonetic variation that are experienced by multilingual individuals.

Our recruitment criteria specified exposure to Cantonese from an early age, at or prior to age six. This lumps very early and early acquisition and both simultaneous and sequential bilinguals all in a single group. This may ultimately not be a uniform population. Exposure to a language from birth has implications for pronunciation patterns. For example, Amengual (2019) examined the lenition rates of phrase-initial voice stops and approximants in the Spanish of simultaneous Spanish-English bilinguals, early sequential Spanish-then-English bilinguals, and late Spanish learners (with English as a first language). The simultaneous bilinguals and late learners patterned together. Given that exposure to English from birth unifies these two groups, these results suggest that early exposure to English has the potential to shape pronunciation patterns in adulthood, similar to previous suggestions for perception (e.g., Sebastián-Gallés et al., 2005). The developmental trajectory out of the sensitive period, however, is gradual, and what exactly is the appropriate age delimiter for a particular linguistic representation, pattern, or process is yet to be determined (see, for example, Flege, 1999; Werker and Tees, 2005; Cargnelutti et al., 2019).

Conclusion

Early Cantonese–English bilinguals exhibited an own-voice benefit for word recognition in Cantonese even when self-recognition of their own voice was masked by a vocal disguise. These results complement the evidence indicating an own-voice benefit in second language speakers (Eger and Reinisch, 2019). The own-voice benefit despite overt recognition of one's own voice suggests a coupled relationship between the motor representations and the multidimensional acoustic-auditory representations that support word recognition.

Data availability statement

The listener data supporting the conclusions can be made available by the authors upon request, as that is what aligns with the approved Ethics protocol.

Ethics statement

The studies involving human participants were reviewed and approved by University of British Columbia's Behavioural Research Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

Author contributions

SC and MB contributed to the conception and design of the study. SC prepared the stimuli and wrote the first draft. MB performed the statistical analyses and wrote sections of the manuscript. Both authors approved the submitted version.

Funding

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC; Funding reference number AWD-002492).

Acknowledgments

Thanks to Martin Oberg for programming assistance; participants for their time; feedback from all of the members in the Speech in Context Lab, particularly Rachel Soo, Fion Fung, and Roger Lo.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.901326/full#supplementary-material

References

Adank, P., Evans, B. G., Stuart-Smith, J., and Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *J. Exp. Psychol.* 35, 520–529. doi: 10.1037/a0013552

Amengual, M. (2019). Type of early bilingualism and its effect on the acoustic realization of allophonic variants: Early sequential and simultaneous bilinguals. *Int. J. Bilingual.* 23, 954–970. doi: 10.1177/1367006917741364

Barreda, S. (2021). Fast track: Fast (nearly) automatic formant-tracking using Praat. *Linguist. Vanguard* 7, 1379–1393. doi: 10.1515/lingvan-2020-0051

Boersma, P., and Weenink, D. (2020). *Praat: Doing phonetics by computer (6.1.21)*. Available online at: http://www.praat.org/ (accessed September 1, 2020).

Bradlow, A. R., and Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition* 106, 707–729. doi: 10.1016/j.cognition.2007.04.005

Bruderer, A. G., Danielson, D. K., Kandhadai, P., and Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proc. Natl. Acad. Sci. U.S.A.* 112, 13531–13536. doi: 10.1073/pnas.1508631112

Bürkner, P. (2018). Advanced Bayesian multilevel modeling with the R package brms. *R J.* 10, 395–411. doi: 10.32614/RJ-2018-017

Cargnelutti, E., Tomasino, B., and Fabbro, F. (2019). Language brain representation in bilinguals with different age of appropriation and proficiency of the second language: A meta-analysis of functional imaging studies. *Front. Hum. Neurosci.* 13:154. doi: 10.3389/fnhum.2019.00154

Chen, S. H., Liu, H., Xu, Y., and Larson, C. R. (2007). Voice F0 responses to pitch-shiftedvoice feedback during English speech. *J. Acoust. Soc. Am.* 121, 1157–1163. doi: 10.1121/1.2404624

Choi, D., Bruderer, A. G., and Werker, J. F. (2019). Sensorimotor influences on speech perception in pre-babbling infants: Replication and extension of Bruderer et al.(2015). *Psychonom. Bull. Rev.* 26, 1388–1399. doi: 10.3758/s13423-019-01601-0

Clopper, C. G., Tamati, T. N., and Pierrehumbert, J. B. (2016). Variation in the strength of lexical encoding across dialects. *J. Phonet.* 58, 87–103. doi: 10.1016/j.wocn.2016.06.002

Cooper, A., Fecher, N., and Johnson, E. K. (2018). Toddlers' comprehension of adult and child talkers: Adult targets versus vocal tract similarity. *Cognition* 173, 16–20. doi: 10.1016/j.cognition.2017.12.013

Coretta, S. (2021). *Github repository*. Available online at: https://github.com/stefanocoretta/bayes-regression (accessed July 15, 2021).

Creel, S. C., and Bregman, M. R. (2011). How talker identity relates to language processing. *Linguist. Lang. Comp.* 5, 190–204. doi: 10.1111/j.1749-818X.2011.00276.x

Creel, S. C., and Tumlin, M. A. (2011). On-line acoustic and semantic interpretation of talker information. *J. Mem. Lang.* 65, 264–285. doi: 10.1016/j.jml.2011.06.005

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behav. Res. Methods* 47, 1–12. doi: 10.3758/s13428-014-0458-y

Devue, C., and Brédart, S. (2011). The neural correlates of visual self-recognition. *Conscious. Cogn.* 20, 40–51. doi: 10.1016/j.concog.2010.09.007

Eger, N. A., and Reinisch, E. (2019). The impact of one's own voice and production skills on word recognition in a second language. *J. Exp. Psychol.* 45, 552–571. doi: 10.1037/xlm0000599

Evans, B. G., and Iverson, P. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. *J. Acoust. Soc. Am.* 121, 3814–3826. doi: 10.1121/1.2722209

Flege, J. E. (1999). "Age of learning and second language speech," in *Second language acquisition and the critical period hypothesis*, ed. D. Birdsong (London: Routledge), 111–142. doi: 10.4324/9781410601667-10

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *J. Acoust. Soc. Am.* 99, 1730–1741. doi: 10.1121/1.415237

Gabry, J., and Češnovar, R. (2021). *cmdstanr: R interface to 'CmdStan'*. Available online at: https://mc-stan.org/cmdstanr (accessed March 1, 2022).

Gertken, L. M., Amengual, M., and Birdsong, D. (2014). "Assessing language dominance with the bilingual language profile," in *Measuring L2 proficiency: Perspectives from SLA*, eds P. Leclercq, A. Edmonds, and H. Hilton (Bristol: Multilingual Matters), 208–225. doi: 10.21832/9781783092291-014

Goldinger, S. D., and Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonom. Bull. Rev.* 11, 716–722. doi: 10.3758/BF03196625

Hay, J., Warren, P., and Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *J. Phonet.* 34, 458–484. doi: 10.1016/j.wocn.2005.10.001

Hazan, V., and Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *J. Acoust. Soc. Am.* 116, 3108–3118. doi: 10.1121/1.1806826

Holmes, E., and Johnsrude, I. S. (2020). Speech spoken by familiar people is more resistant to interference by linguistically similar speech. *J. Exp. Psychol.* 46, 1465–1476. doi: 10.1037/xlm0000823

Holmes, E., Domingo, Y., and Johnsrude, I. S. (2018). Familiar voices are more intelligible, even if they are not recognized as familiar. *Psychol. Sci.* 29, 1575–1583. doi: 10.1177/0956797618779083

Houde, J. F., and Jordan, M. I. (2002). Sensorimotor adaptation of speech I: Compensation and adaptation. *J. Speech Lang. Hear. Res.* 45, 295–310. doi: 10.1044/1092-4388(2002/023)

Howell, P., and Dworzynski, K. (2001). Strength of German accent under altered auditory feedback. *Percept. Psychophys.* 63, 501–513. doi: 10.3758/bf03194416

Howell, P., Barry, W., and Vinson, D. (2006). Strength of British English accents in altered listening conditions. *Percept. Psychophys.* 68, 139–153. doi: 10.3758/bf03193664

Hughes, S. M., and Harrison, M. A. (2013). I like my voice better: Self-enhancement bias in perceptions of voice attractiveness. *Perception* 42, 941–949. doi: 10.1068/p7526

Johnson, K. (1997). "Speech perception without speaker normalization: An exemplar model," in *Talker Variability in Speech Processing*, eds K. Johnson and J. W. Mullennix (San Diego, CA: Academic Press), 145–165.

Johnson, K. A., Babel, M., and Fuhrman, R. A. (2020). *Bilingual acoustic voice variation is similarly structured across languages. Proceedings of Interspeech*. Available online at: https://www.isca-speech.org/archive_v0/Interspeech_2020/pdfs/3095.pdf doi: 10.21437/Interspeech.2020-3095 (accessed October 1, 2020).

Johnson, K., Strand, E. A., and D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *J. Phonet.* 27, 359–384. doi: 10.1006/jpho.1999.0100

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., and Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychol. Sci.* 24, 1995–2004. doi: 10.1177/0956797613482467

Jones, J. A., and Munhall, K. G. (2002). The role of auditory feedback during phonation: Studies of Mandarin tone production. *J. Phonet.* 30, 303–320. doi: 10.1006/jpho.2001.0160

Katseff, S., Houde, J., and Johnson, K. (2012). Partial compensation for altered auditory feedback: A trade-off with somatosensory feedback? *Lang. Speech* 55, 295–308. doi: 10.1177/0023830911417802

Keenan, J. P., Ganis, G., Freund, S., and Pascual-Leone, A. (2000). Self-face identification is increased with left hand responses. *Lateral. Asymmetr. Body Brain Cogn.* 5, 259–268. doi: 10.1080/713754382

Keyes, H., Brady, N., Reilly, R. B., and Foxe, J. J. (2010). My face or yours? Event-related potential correlates of self-face processing. *Brain Cogn.* 72, 244–254. doi: 10.1016/j.bandc.2009.09.006

Kraljic, T., Brennan, S. E., and Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition* 107, 54–81. doi: 10.1016/j.cognition.2007.07.013

Kreitewolf, J., Mathias, S. R., and von Kriegstein, K. (2017). Implicit talker training improves comprehension of auditory speech in noise. *Front. Psychol.* 8:1584. doi: 10.3389/fpsyg.2017.01584

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., and Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Curr. Biol.* 23, 1075–1080. doi: 10.1016/j.cub.2013.04.055

Lavan, N., Knight, S., and McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nat. Commun.* 10:2404. doi: 10.1038/s41467-019-10295-w

Lavner, Y., Gath, I., and Rosenhouse, J. (2000). Effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Commun.* 30, 9–26. doi: 10.1016/S0167-6393(99)00028-X

Lavner, Y., Rosenhouse, J., and Gath, I. (2001). The prototype model in speaker identification by human listeners. *Int. J. Speech Technol.* 4, 63–74. doi: 10.1023/A:1009656816383

Lee, Y., Keating, P., and Kreiman, J. (2019). Acoustic voice variation within and between speakers. *J. Acoust. Soc. Am.* 146, 1568–1579.

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36.

Liu, L., Li, W., Li, J., Lou, L., and Chen, J. (2019). Temporal features of psychological and physical self-representation: An ERP study. *Front. Psychol.* 10:785. doi: 10.3389/fpsyg.2019.00785

Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *J. Speech Lang. Hear. Res.* 50, 940–967. doi: 10.1044/1092-4388(2007/067)

Mitsuya, T., Macdonald, E. N., Purcell, D. W., and Munhall, K. G. (2011). A cross-language study of compensation in response to real-time formant perturbation. *J. Acoust. Soc. Am.* 130, 2978–2986. doi: 10.1121/1.3643826

Mitterer, H., Eger, N. A., and Reinisch, E. (2020). My English sounds better than yours: Second-language learners perceive their own accent as better than that of their peers. *PLoS One* 15:e0227643. doi: 10.1371/journal.pone.0227643

Munson, B., Jefferson, S. V., and McDonald, E. C. (2006). The influence of perceived sexual orientation on fricative identification. *J. Acoust. Soc. Am.* 119, 2427–2437. doi: 10.1121/1.2173521

Newman, R. S., Clouse, S. A., and Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *J. Acoust. Soc. Am.* 109, 1181–1196. doi: 10.1121/1.1348009

Nicenboim, B., and Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas—Part II. *Lang. Linguist. Comp.* 10, 591–613.

Niziolek, C. A., and Guenther, F. H. (2013). Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations. *J. Neurosci.* 33, 12090–12098. doi: 10.1523/JNEUROSCI.1008-13.2013

Nygaard, L. C., and Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Percept. Psychophys.* 60, 355–376. doi: 10.3758/BF03206860

Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychol. Sci.* 5, 42–46. doi: 10.1111/j.1467-9280.1994.tb00612.x

Peng, Z., Wang, Y., Meng, L., Liu, H., and Hu, Z. (2019). One's own and similar voices are more attractive than other voices. *Austral. J. Psychol.* 71, 212–222. doi: 10.1111/ajpy.12235

Perry, L. K., Mech, E. N., MacDonald, M. C., and Seidenberg, M. S. (2018). Influences of speech familiarity on immediate perception and final comprehension. *Psychonom. Bull. Rev.* 25, 431–439. doi: 10.3758/s13423-017-1297-5

Platek, S. M., Keenan, J. P., Gallup, G. G., and Mohamed, F. B. (2004). Where am I? The neurological correlates of self and other. *Cogn. Brain Res.* 19, 114–122. doi: 10.1016/j.cogbrainres.2003.11.014

Platek, S. M., Loughead, J. W., Gur, R. C., Busch, S., Ruparel, K., Phend, N., et al. (2006). Neural substrates for functionally discriminating self-face from personally familiar faces. *Hum. Brain Mapp.* 27, 91–98. doi: 10.1002/hbm.20168

Prinz, W. (1997). Perception and action planning. *Eur. J. Cogn. Psychol.* 9, 129–154.

Purcell, D. W., and Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *J. Acoust. Soc. Am.* 119, 2288–2297. doi: 10.1121/1.2173514

R Core Team (2021). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.

Reinfeldt, S., Östli, P., Håkansson, B., and Stenfelt, S. (2010). Hearing one's own voice during phoneme vocalization—Transmission by air and bone conduction. *J. Acoust. Soc. Am.* 128, 751–762. doi: 10.1121/1.3458855

Schertz, J., and Clare, E. J. (2020). Phonetic cue weighting in perception and production. *Wiley Interdiscipl. Rev. Cogn. Sci.* 11, 1–24. doi: 10.1002/wcs.1521

Schouten, B., Gerrits, E., and Van Hessen, A. (2003). The end of categorical perception as we know it. *Speech Commun.* 41, 71–80.

Schuerman, W. L. (2017). *Sensorimotor experience in speech perception.* Ph.D. thesis. Nijmegen: Radboud University Nijmegen.

Schuerman, W. L., Meyer, A., and McQueen, J. M. (2015). Do we perceive others better than ourselves? A perceptual benefit for noise-vocoded speech produced by an average speaker. *PLoS One* 10:e0129731. doi: 10.1371/journal.pone.0129731

Schuerman, W., McQueen, J. M., and Meyer, A. (2019). "Speaker statistical averageness modulates word recognition in adverse listening conditions," in *Proceedings of the 19th international congress of phonetic sciences (ICPhS 2019)*, eds S. Calhoun, P. Escudero, M. Tabain, and P. Warren (Canberra, NSW: Australasian Speech Science and Technology Association Inc), 1203–1207.

Sebastián-Gallés, N., Echeverría, S., and Bosch, L. (2005). The influence of initial exposure on lexical representation: Comparing early and simultaneous bilinguals. *J. Mem. Lang.* 52, 240–255. doi: 10.1162/jocn.2008.20004

Senior, B., and Babel, M. (2018). The role of unfamiliar accents in competing speech. *J. Acoust. Soc. Am.* 143, 931–942.

Shuster, L. I., and Durrant, J. D. (2003). Toward a better understanding of the perception of self-produced speech. *J. Commun. Disord.* 36, 1–11. doi: 10.1016/S0021-9924(02)00132-6

Souza, P., Gehani, N., Wright, R., and McCloy, D. (2013). The advantage of knowing the talker. *J. Am. Acad. Audiol.* 24, 689–700. doi: 10.3766/jaaa.24.8.6

Stenfelt, S., and Goode, R. L. (2005). Bone-conducted sound: Physiological and clinical aspects. *Otol. Neurotol.* 26, 1245–1261. doi: 10.1097/01.mao.0000187236.10842.d5

Sumner, M., and Kataoka, R. (2013). Effects of phonetically-cued talker variation on semantic encoding. *J. Acoust. Soc. Am.* 134, EL485–EL491. doi: 10.1121/1.4826151

Sumner, M., and Samuel, A. G. (2005). Perception and representation of regular variation: The case of final/t. *J. Mem. Lang.* 52, 322–338.

Sumner, M., and Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *J. Mem. Lang.* 60, 487–501. doi: 10.1097/WNR.0b013e3283263000

Sumner, M., Kim, S. K., King, E., and McGowan, K. B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Front. Psychol.* 4:1015. doi: 10.3389/fpsyg.2013.01015

Szakay, A., Babel, M., and King, J. (2016). Social categories are shared across bilinguals× lexicons. *J. Phonet.* 59, 92–109.

Todd, S., Pierrehumbert, J. B., and Hay, J. (2019). Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition* 185, 1–20. doi: 10.1016/j.cognition.2019.01.004

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *J. Acoust. Soc. Am.* 88, 97–100. doi: 10.1121/1.399849

Tye-Murray, N., Spehar, B. P., Myerson, J., Hale, S., and Sommers, M. S. (2013). Reading your own lips: Common-coding theory and visual speech perception. *Psychonom. Bull. Rev.* 20, 115–119. doi: 10.3758/s13423-012-0328-5

Tye-Murray, N., Spehar, B. P., Myerson, J., Hale, S., and Sommers, M. S. (2015). The self-advantage in visual speech processing enhances audiovisual speech recognition in noise. *Psychonom. Bull. Rev.* 22, 1048–1053. doi: 10.3758/s13423-014-0774-3

Uddin, L. Q., Kaplan, J. T., Molnar-Szakacs, I., Zaidel, E., and Iacoboni, M. (2005). Self-face recognition activates a frontoparietal "mirror" network in the right hemisphere: An event-related fMRI study. *Neuroimage* 25, 926–935. doi: 10.1016/j.neuroimage.2004.12.018

Van Lancker, D., Kreiman, J., and Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters Part I: Recognition of backward voices. *J. Phonet.* 13, 19–38. doi: 10.1016/s0095-4470(19)30723-5

Vaughn, C., Baese-Berk, M., and Idemaru, K. (2019). Re-examining phonetic variability in native and non-native speech. *Phonetica* 76, 327–358.

Venables, B. (2021). *codingMatrices: Alternative factor coding matrices for linear model formulae. R package version 0.3.3.* Available online at: https://CRAN.R-project.org/package=codingMatrices (accessed March 1, 2022).

Werker, J. F., and Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behav. Dev.* 7, 49–63.

Werker, J. F., and Tees, R. C. (2005). Speech perception as a window for understanding plasticity and commitment in language systems of the brain. *Dev. Psychobiol.* 46, 233–251. doi: 10.1002/dev.20060

Woods, K. J. P., Siegel, M. H., Traer, J., and McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attent. Percept. Psychophys.* 79, 2064–2072. doi: 10.3758/s13414-017-1361-2

Xu, M., Homae, F., Hashimoto, R., and Hagiwara, H. (2013). Acoustic cues for the recognition of self-voice and other-voice. *Front. Psychol.* 4:735. doi: 10.3389/fpsyg.2013.00735

Xu, Y., Larson, C. R., Bauer, J. J., and Hain, T. C. (2004). Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *J. Acoust. Soc. Am.* 116, 1168–1178. doi: 10.1121/1.1763952

Yonan, C. A., and Sommers, M. S. (2000). The effects of talker familiarity on spoken word identification in younger and older listeners. *Psychol. Aging* 15, 88–99.

Zhao, K., Wu, Q., Zimmer, H. D., and Fu, X. (2011). Electrophysiological correlates of visually processing subject's own name. *Neurosci. Lett.* 491, 143–147. doi: 10.1016/j.neulet.2011.01.025

Zheng, Y., and Samuel, A. G. (2017). Does seeing an Asian face make speech sound more accented? *Attent. Percept. Psychophys.* 79, 1841–1859. doi: 10.3758/s13414-017-1329-2

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 26 September 2022
DOI 10.3389/fcomm.2022.864127

OPEN ACCESS

EDITED BY
Joseph V. Casillas,
Rutgers, The State University of New Jersey, United States

REVIEWED BY
Jessie S. Nixon,
University of Tübingen, Germany
Allard Jongman,
University of Kansas, United States

*CORRESPONDENCE
Roger Yu-Hsiang Lo
roger.lo@ubc.ca

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

RECEIVED 28 January 2022
ACCEPTED 23 August 2022
PUBLISHED 26 September 2022

CITATION
Lo RY-H (2022) The dual role of post-stop fundamental frequency in the production and perception of stops in Mandarin-English bilinguals. *Front. Commun.* 7:864127.
doi: 10.3389/fcomm.2022.864127

The dual role of post-stop fundamental frequency in the production and perception of stops in Mandarin-English bilinguals

Roger Yu-Hsiang Lo*

Department of Linguistics, University of British Columbia, Vancouver, BC, Canada

In non-tonal languages with a two-way laryngeal contrast, post-stop fundamental frequency (F0) tends to vary as a function of phonological voicing in stops, and listeners use it as a cue for stop voicing. In tonal languages, F0 is the most important acoustic correlate for tone, and listeners likewise rely heavily on F0 to differentiate tones. Given this ambiguity of F0 in its ability to signal phonological voicing and tone, how do speakers of a tonal language weight it in production and perception? Relatedly, do bilingual speakers of tonal and non-tonal languages use the same weights across different language contexts? To address these questions, the cross-linguistic performances from L1 (first language) Mandarin-L2 (second language) English bilinguals dominant in Mandarin in online production and perception experiments are compared. In the production experiment, the participant read aloud Mandarin and English monosyllabic words, the onsets of which typified their two-way laryngeal contrast. For the perception experiment, which utilized a forced-choice identification paradigm, both the English and Mandarin versions shared the same target audio stimuli, comprising monosyllables whose F0 contours were modeled after Mandarin Tone 1 and Tone 4, and whose onset was always a bilabial stop. The voice onset time of the bilabial stop and the onset F0 of the nucleus were manipulated orthogonally. The production results suggest that post-stop F0 following aspirated/voiceless stops was higher than that following unaspirated/voiced stops in both Mandarin and English production. However, the F0 difference in English was larger as compared to Mandarin, indicating that participants assigned more production weight to post-stop F0 in English than in Mandarin. On the perception side, participants used post-stop F0 as a cue in perceiving stops in both English and Mandarin, with higher post-stop F0 leading to more aspirated/voiceless responses, but they allocated more weight to post-stop F0 when interpreting audio stimuli as English words than as Mandarin words. Overall, these results argue for a dual function of F0 in cueing phonological voicing in stops and lexical tone across production and perception in Mandarin. Furthermore, they suggest that bilinguals are able to dynamically adjust even a secondary cue according to different language contexts.

KEYWORDS

fundamental frequency, stop, Mandarin, English, bilingual, production, perception, cue weighting

1. Introduction

Speech sounds contrast on a multitude of continuous acoustic dimensions, with some dimensions being used as primary cues to a phonological contrast while others play a more secondary part. Following Toscano and McMurray (2010), I use the term *cue* to refer to any source of information that allows the perceiver to distinguish between different responses (e.g., the response might be whether the sound is an [i] or an [a]). An example that is often given in this connection is Lisker's (1986) finding that potential cues to word-medial voicing in English (e.g., *rapid* vs. *rabid*) include duration of the preceding vowel, duration of the closure, voice onset time (VOT), presence of vocal fold vibration during closure, burst amplitude, fundamental frequency (F0) going into and out of the closure, among others. However, the reverse—that an acoustic dimension can serve as a cue for multiple phonological contrasts—is also true but often less studied. For instance, formant frequency is not only an important cue for vowel quality, but the transition for a formant frequency band also cues the place of articulation for stop consonants (e.g., Liberman et al., 1954). Given this many-to-many mapping between phonological contrasts and acoustic dimensions, ambiguity about how speakers encode various cues for a contrast and how listeners infer potential contrasts from a cue naturally arises.

The current study explores this ambiguity from the perspective of both speech production and perception. Specifically, I am interested in (i) whether and how F0 is used by speakers of a tonal language to signal and perceive phonological *voicing* in stops, aside from lexical tone, and (ii) whether the use of F0 might be mediated by different language contexts. These two questions are addressed in tandem by comparing L1 (first language) Mandarin-L2 (second language) English bilinguals' performances in production and perception of Mandarin and English stops. The production task involves the participants reading aloud words with a stop in the onset position, while the perception part asks the participants to respond in a forced-choice identification task based on synthetic continua of both VOT and F0 values.

2. Background

2.1. Fundamental frequency as a cue to lexical tone

Similar to segments, lexical tones contrast on multiple acoustic dimensions, such as duration and intensity; however, F0 has long been established as the most important acoustic correlate for tonal distinctions, as far as Mandarin is concerned (Ohala, 1978). Indeed, the tone letters in the International Phonetic Alphabet are in their essence a discretized representation over a speaker's full pitch range, and the descriptions for lexical tones in Mandarin closely follow the F0 as they unfold over a syllable—Tone 1: high-level ˥, Tone 2: mid-rising ˧˥, Tone 3: low-dipping ˨˩˦, and Tone 4: high-falling ˥˩. Even though F0 is not the only dimension that covaries with each tone in production (Ho, 1976), and it is not the only dimension that listeners take advantage of when distinguishing tones (e.g., Blicher et al., 1990), it is the primary source that Mandarin users rely on to signal and extract information regarding tonal contrast (Gandour, 1978).

In this study, I restrict the scope to only Tone 1 and Tone 4 for both theoretical and practical considerations. On the theoretical side, Tone 1 and Tone 4 are the only two tones in Mandarin that start with the same phonological tonal register (i.e., both start with a high target), so listeners need to track the F0 trajectory, at least for the initial portion of a tonal contour initiated with a high register, to reliably tell these two tones apart. This is an important consideration for the design of the perception experiment, as will be explained in Section 2.2.2. Also, given that both Tone 1 and Tone 4 begin in the upper part of the pitch range, post-stop F0 behaviors, which will be discussed in the next section, should be more comparable across these two tones, as there is evidence suggesting that post-stop F0 is contingent on pitch height.

2.2. Fundamental frequency as a cue to stop voicing

2.2.1. Post-stop F0 in English

It has been observed that F0 in the vowel following a stop consonant tends to correlate with voicing distinctions cross-linguistically [e.g., Cantonese (Francis et al., 2006; Luo, 2018; Ren and Mok, 2021), English (House and Fairbanks, 1953; Lehiste and Peterson, 1961; Lea, 1973; Hombert, 1978; Hombert et al., 1979; Ohde, 1984; Hanson, 2009), French (Kirby and Ladd, 2016), German (Kohler, 1982), Japanese (Gao and Arai, 2018), Korean (Han and Weitzman, 1970; Jun, 1996), Mandarin (Howie, 1976; Xu and Xu, 2003; Chen, 2011; Luo, 2018; Guo, 2020), Russian (Mohr, 1971), Spanish (Dmitrieva et al., 2015), Thai (Gandour, 1974; Ewan, 1976), Xhosa (Jessen and Roux, 2002), Yoruba (Hombert, 1978)]. This phenomenon is commonly labeled as post-stop F0 perturbation, pitch skip, obstruent intrinsic F0, co-intrinsic pitch, or onset F0 perturbation. For English, whose six stops come in phonologically voiced-voiceless pairs: /b/-/p/, /d/-/t/, and /g/-/k/, it is well-established that F0 at vowel onset is significantly higher following phonologically voiceless stops than following phonologically voiced ones, regardless of the presence of actual vocal fold vibration (e.g., Abramson and Lisker, 1985; Dmitrieva et al., 2015). This type of patterning has led Kingston and Diehl (1994) to argue that post-stop F0 is not purely a result of intrinsic physiological dependencies between the articulatory and/or aerodynamic properties and the production of degrees of prevoicing or voicing delay—instead, it is at least partially the result of controlled processes referring to the phonological status of the consonant series.

The perceptual consequences of post-stop F0 to the voicing contrast are also firmly established for English: a higher post-stop F0 tends to lead to more voiceless responses than a lower F0, especially when VOT is ambiguous (Whalen et al., 1990, 1993; Francis et al., 2006). Some authors have attributed the perceptual effects of post-stop on voicing decisions to the observation that a low F0 enhances the perceptual "voicedness" of a stop by highlighting the percept of low-frequency periodic energy in the proximity of the stop release (Kingston and Diehl, 1994; Kingston et al., 2008).

2.2.2. Post-stop F0 in Mandarin

With regard to the post-stop F0 perturbation effect in Mandarin, which has six stops coming in unaspirated-aspirated pairs: /p/-/p^h/, /t/-/t^h/, and /k/-/k^h/, the existing literature depicts a mixed picture, with conflicting results across studies. Both English and Mandarin have two phonological voicing classes, with the voiced / unaspirated class typically having a short-lag VOT (under 30 ms) and the voiceless / aspirated class having a long-lag VOT (above 30 ms). Based on this similar phonetic implementation, one would expect Mandarin to pattern with English in terms of post-stop F0 effects, that is, aspirated stops should have a higher post-stop F0 than unaspirated stops. Indeed, this is the pattern found by Chen (2011) and Luo (2018). Based on read speech from 15 female native speakers of Mainland Mandarin reading monosyllabic CV words containing all six stops inserted in a carrier phrase, Luo (2018) found that aspirated stops were associated with greater F0 perturbation (i.e., a higher F0) than unaspirated stops, with a mean F0 difference in the range of 11.67 Hz and 18.35 Hz, depending on the lexical tone. With a similar experiment design to that in Luo (2018), but with gender-balanced speakers (10 females and 10 males), Chen (2011) also reached the conclusion that vowels following an aspirated stop had a higher F0 than those following an unaspirated stop in Taiwan Mandarin (for females, the difference in F0 ranged from 2 Hz and 14 Hz; for males, the range was between 2.8 Hz and 8 Hz). This general pattern was also reported in a blog post by Liberman (2014), based on the data from the Mandarin Chinese Phonetic Segmentation and Tone corpus (Yuan et al., 2014). However, as Liberman (2014) did not conduct statistical tests on this set of data, it is not yet clear if the difference was statistically significant (across genders, the mean F0 difference was between 1.5 Hz and 5.7 Hz for the /p/-/p^h/ contrast, between 1.0 Hz and 3.5 Hz for the /t/-/t^h/ contrast, and between 2.8 Hz and 7.2 Hz for the /k/-/k^h/ contrast). Rather puzzlingly, a pattern that is opposite to the above generalizations was also observed in the work by Xu and Xu (2003), where they reported that it was *unaspirated* stops that triggered a higher F0 on the onset of the following vowel (with a mean F0 difference ranging between 5 Hz and 50 Hz), using production data from seven female native speakers of Mainland Mandarin pronouncing disyllabic words containing /ta/ and /t^{h}a/ embedded in a carrier phrase. Even more interestingly, a recent work from Guo (2020), which used as stimuli tonal syllables with onsets /t/, /t^h/, or /w/ and rimes /a/ or /u/ in the four lexical tones, showed that the direction of post-stop F0 perturbation depended on the tone, such that F0 was higher following an aspirated stop only in Tone 1 and Tone 4 (i.e., tones beginning with a high register) while the opposite pattern was observed for Tone 2 and Tone 3, both of which have a low initial register.

More broadly, the issue of post-stop F0 perturbation in Mandarin is related to the debate of whether there is a trade-off between post-stop F0 and tone, and of whether the existence of tone attenuates the degree of post-stop F0 difference. While there are some studies that provide a positive answer [e.g., Gandour (1974) for Thai and Hombert (1978) for Yoruba], larger magnitudes have also been reported in tonal languages [e.g., Phuong (1981) for Northen Vietnamese, Shimizu (1994) for Thai, Xu and Xu (2003) for Mandarin, and Francis et al. (2006) for Cantonese]. In the current study, the parallel production experiments in Mandarin as well as English allow us to address this debate from a bilingual perspective. That is, the production data in Mandarin and English enables a comparison of the degree of post-stop F0 difference across a tonal and a non-tonal language within the same speaker.

The perceptual contribution of post-stop F0 to the voicing contrast in Mandarin is substantially less studied. To my knowledge, Guo (2020) is the first to systematically study whether post-stop F0 is used by Mandarin speakers as a cue when tasked to distinguish the stop voicing contrast in Mandarin. Using a two-alternative forced choice (2AFC) paradigm, Guo (2020) showed that Mandarin speakers capitalized on post-stop F0 to decode consonantal voicing information. However, the identification experiment in her study only required the listener to distinguish aspirated vs. unaspirated stops in the context of the same lexical tone (i.e., the two alternatives in the 2AFC paradigm only differed in stop voicing but shared the same lexical tone), and so it is still unclear whether Mandarin listeners continue to use post-stop F0 as a cue for voicing when they have to extract tonal information from pitch at the same time. The design of the current perception experiment addresses this problem, as explained in Section 4.3.

2.2.3. Post-stop F0 and F0 contour

Given that post-stop F0 is embedded in the global F0 trajectory that also encodes tonal and intonational information, this section briefly reviews the interaction between post-stop F0 and F0 contour in English and Mandarin. In English production, Hanson (2009) examined the effects of obstruents on F0 contour in either a high, low, or neutral pitch environment by having participants read CVm syllables in carrier sentences. She found that, in a high-pitch environment, the initial F0 contour following a voiceless stop was raised relative to the baseline /m/, but following a voiced stop, it closely approximated the baseline. In a low-pitch environment, however, both voiceless and voiced stops raised the initial F0 contour. In Mandarin production, regardless of whether aspirated stops were found to lead to a higher post-stop F0 than unaspirated ones (e.g., Chen, 2011; Luo, 2018) or otherwise (Xu and Xu, 2003), visual inspection of the F0 trajectories in these studies suggests that both aspirated and unaspirated stops raised the initial F0 contour in all lexical tonal contexts.

With respect to perception, much less is known about how F0 contour affects the perceived phonological voicing of the initial stop. It is well established that listeners of both tonal and non-tonal languages are sensitive to changes in F0 in signaling sentential intonation or lexical tone (e.g., Gandour, 1983; Ma et al., 2006; Barnes et al., 2010; Liu and Rodriguez, 2012; Xu and Mok, 2012; Dilley and Heffner, 2013; Leung and Wang, 2020). For instance, Gandour (1983) asked listeners of tonal languages (Cantonese, Mandarin, Taiwanese, Thai) and a non-tonal language (English) to make direct paired-comparison judgments of tone dissimilarity. His results revealed that the direction dimension was more important than the height dimension for listeners of a tonal language vs. a non-tonal language. Leung and Wang (2020) tested the production-perception link in three critical tonal cues—slope, curvature, and turning-point location—and two non-critical cues—mean F0 height and onset F0 height—while Mandarin listeners rated different exemplars of Tone 2. They found that statistically significant correlation was found only for critical cues. In terms of how F0 contour might bias the identification of a segment, Lehnert-LeHouillier (2007) examined German, Japanese, Spanish, and Thai listeners' identification of vowel length, using vowel continua varying orthogonally in both duration (from around 220 ms to 400 ms with a step size of about 30 ms) and F0 contour (level at 180 Hz and falling from 160 Hz to 80 Hz). She found that only Japanese listeners perceived the vowels with a falling F0 as longer; the F0 contour did not seem to have an effect for listeners of other languages. Fogerty and Humes (2012) investigated the contribution of F0, speech envelope, and temporal fine structure in consonants or vowels to overall word and sentence intelligibility. They observed that when dynamic F0 cues were flatted or removed, English listeners still obtained higher recognition scores for vowel-only (i.e., consonantal portions were masked) sentences, as compared to consonant-only (i.e., vocalic portions were masked) ones. These results suggest that dynamic F0 contour might play an important role in consonant identification. However, to the best of my knowledge, no study has systematically investigated how F0 contour alone (e.g., different F0 directions with the same onset F0 height) modulates the perception of voicing of the initial obstruent. While the current study does not set out to examine the respective contribution of post-stop F0 height and F0 contour to the perception of voicing, the potential influence of F0 contour will be addressed in Section 5.4.

2.3. Post-stop F0 at L1 production-perception interface

While there is clear evidence that post-stop F0 functions as a cue for voicing in production as well as in perception *separately*, outcomes from attempts to link the cue use *across* the two modalities remain inconclusive. More generally, based on the proposal that perceptual cue weights arise from statistical regularities in the put (e.g., Holt and Lotto, 2006; Francis et al., 2008; Toscano and McMurray, 2010), one would anticipate the relative informativeness of a cue in a speaker's productions of a contrast to be predictive of the reliance assigned to that cue in perceiving the same contrast. Theories that posit a strong and/or direct connect between production and perception, such as Motor Theory (Liberman and Mattingly, 1985) or Direct Realism (Fowler, 1986), also express such a view. However, although it is established that distributional patterns in production are exploited as cues in perception at the macro level, efforts to find correlations between use of the same cue across production and perception at the micro or individual level have been met with mixed success. For example, while Zellou (2017) found that individuals' production of anticipatory nasal coarticulation on vowels in English was correlated with their patterns of perceptual compensation, Kataoka (2011) found no significant correlation between Californians' production and perception of /u/-fronting in alveolar contexts. Zooming in on the use of post-stop F0, even as the use of post-stop F0 as a perceptual cue for stop voicing reflects the differential F0 at vowel onset in production on a population level, correlational analysis on an individual level has yet to reveal a more direct connection. For instance, the importance an English speaker assigns to post-stop F0 in production does not seem to predict the perceptual reliance of the same cue from the same individual (Shultz et al., 2012). A similar lack of relationship in post-stop F0 cue use for Spanish speakers was reported in Schertz et al. (2020). This study revisits this topic and explores whether there is a direct link between production and perception for the use of post-stop F0 in Mandarin, at both the population and individual levels.

2.4. Post-stop F0 at L2 production-perception interface

If producing and perceiving a phonological contrast means navigating between various acoustic dimensions, learning a phonological contrast in an L2 then involves adapting the weight associated with relevant dimension to approach that of native speakers of the L2 in question. The majority of work on L2 sound production and perception has put an emphasis on how L2 learners acquire foreign contrasts that rely primarily on dimensions that are not use in similar native contrasts. For instance, the difficulty for Japanese speakers to distinguish the English /r/-/l/ contrast is ascribed to the fact that this English contrast relies mainly on a difference in third formant values, whereas it is the second formant that Japanese speakers use to distinguish the categories (Miyawaki et al., 1975; Iverson et al., 2003; Lotto et al., 2004).

Another interesting line of research focuses on cases in which a first language (L1) contrast primarily relies on *more* cues than the corresponding L2 contrast. A study in this direction is Schertz et al.'s (2015) research on how L1 speakers of Korean, which uses both VOT and post-stop F0 as primary cues for its three-way stop distinction, produce and perceive the L2 English stop contrast, which relies primarily only on VOT.

The current work represents a study that is in some sense sandwiched between the two threads of research discussed above. In particular, similar to English, Mandarin relies primarily on VOT to signal its stop voicing contrast; this therefore distinguishes the case of L1 Mandarin speakers learning the L2 English stop contrast from that of L1 Japanese speakers coping with the English /r/-/l/ contrast. However, this study also deviates from Schertz et al.'s (2015) study of L1 Korean speakers in that, unlike Korean, which uses *both* VOT and F0 as primary cues for its three-way stop contrast, Mandarin only uses F0 as a secondary cue for its two-way stop contrast, but as the primary cue for its lexical tones. Crucially, for L1 speakers of a tonal language learning a non-tonal L2, F0 is an ambiguous cue that signals both tonal and non-tonal (e.g., stop voicing) contrasts in L1, but only non-tonal contrasts in L2. Examining this sort of scenario is therefore important for understanding to what extent L2 learners learn to reweight cues across phonological domains (i.e., using F0 as a dual segmental and suprasegmental cue to using it solely as a segmental cue) during L2 sound category acquisition.

In fact, the research questions raised here have been partially addressed by Guo (2020). In her study, she had a group of Mandarin-English bilinguals dominant in Mandarin produce a set of Mandarin and English words typifying stop voicings in the respective languages, and the same group of participants also took part in 2AFC perception experiments, identifying Mandarin and English words with different combinations of VOT and post-stop F0 values. Visual inspection of her production results suggests that the difference in post-stop F0 between long-lag stops and short-lag stops is smaller in Mandarin than in English, though no statistical models were used to test this observation. In perception, her results also suggest that Mandarin listeners use post-stop F0 as a cue for stop voicing in both L1 Mandarin and L2 English word identification tasks, but whether the extent with which they relied on post-stop F0 differed according to the language context was not analyzed. In this study, these caveats were addressed with a different experiment design.

Much like the link between production and perception in L1, the production-perception interface in L2 has turned out to be elusive, potentially due to more individual variability induced by more diverse L2 learning experiences. While at the broad level, the perception patterns often mirror production patterns, and vice versa, work looking for production-perception links with respect to individual cue weights has had limited luck finding correlation between the two modalities. For example, in studying L1 Korean learners' production and perception of the stop voicing contrast in English, Schertz et al. (2015) find considerable individual difference in L2 English perceptual categorization strategies in spite of the relative homogeneity of their L2 English production. In the current work, the L2 production-perception interface was also briefly examined, focusing on the use of post-stop F0 in L1 Mandarin learners' production and perception of English stops.

2.5. L1 influence on L2 cue use

Given that the target population in this study is L1 Mandarin-L2 English speakers, one would expect the usage patterns of multiple acoustic dimensions in their L2 English to be influenced by their L1 Mandarin. Such an L1-to-L2 influence can be understood in the frameworks of two major theories of L2 speech sound acquisition—the Speech Learning Model (SLM, Flege, 1995, 2007) and the Perceptual Assimilation Model's extension to L2 acquisition (PAM-L2, Best and Tyler, 2007). Both models relate the patterns of L2 sound acquisition to L1 phonology by assuming that L2 sounds are assimilated to L1 sound categories whenever possible. The difficulty of L2 sound discriminability is therefore projected from the phonetic similarity between L1 and L2 sounds, and the patterns of assimilation from L2 to L1 categories. Given that both the English and Mandarin stop contrasts make use of VOT as the primary cue, that the absence/presence of aspiration is an important indicator for phonological voicing, and that both languages have two stop categories in terms of phonological voicing, English phonemically voiced (/b, d, ɡ/) and voiceless (/p, t, k/) stops in the word-initial position will almost certainly be assimilated to Mandarin unaspirated (/p, t, k/) and aspirated stops (/p^h, t^h, k^h/), respectively. In the extreme case where English stops are processed as Mandarin stops, one would expect the participants to transfer their native Mandarin cue-weighting strategies to English, in both production and perception.

However, more recent works have also demonstrated that late L2 learners are able to fine-tune the use of various acoustic dimensions in different language contexts. For instance, Amengual (2021) examined the VOT of the English, Japanese, and Spanish /k/ in the productions of L1 English-L2 Japanese bilinguals, L1 Japanese-L2 English bilinguals, and L1 Spanish-L2 English-L3 Japanese trilingual and found that all three groups of speakers produced language-specific VOT patterns for each language, despite evidence of cross-linguistic influence. In perception, Casillas and Simonet (2018) investigated whether English beginner learners of Spanish at the early stages of their development could manifest the double phonemic boundary effect in VOT—that is, whether these bilinguals shift the perceptual VOT boundary according to the language mode they are in—and found that they were indeed able to manifest the effect, suggesting that the ability of switching between language-specific perceptual modes can be acquired later in life. It is therefore possible that the bilingual participants in this study are capable of adjusting the weight of post-stop F0 according to the language context. The production and perception experiments presented in this work allow for robust investigation of this possibility.

TABLE 1 Predicted production and perception results under difference hypotheses.

Production	
Hypothesis	**Predicted production results**
Post-stop F0 purely due to physiological / aerodynamic reasons (e.g., Ladefoged, 1967; Ohala and Ohala, 1972; Kohler, 1984) or total transfer of post-stop F0 cue use in Mandarin to English, as prediced by the SLM and PAM-L2	Post-stop F0 difference the same in Mandarin and English tokens
Post-stop F0 partially subject to active controlling (Kingston and Diehl, 1994)	The extent of post-stop F0 difference might depend on the language (i.e., larger in English than in Mandarin)
Perception	
Hypothesis	**Predicted perception results**
Transfer of the Mandarin cue-weighting strategy to English, as predicted by the SLM and PAM-L2	Post-stop F0 weights the same across Mandarin and English
Flexibility in cue use: attributing variation in post-stop F0 partially to lexical tone and partially to stop voicing in Mandarin, but only to stop voicing in English	Post-stop F0 weights depend on the language context (i.e., a higher weight in English than in Mandarin)

2.6. Goals of the current study

The use of F0 as a medium for the lexical tones in Mandarin provides an opportunity to examine whether F0 also functions as a cue for stop voicing in production—as has been found for a number of non-tonal languages—and as a cue for stop voicing in perception when Mandarin listeners also need to extract tonal information from F0. With respect to production, previous work has not converged to a definite conclusion, so the current study aims to first establish the post-stop F0 production patterns in the participating speakers. Concerning perception, while there is evidence that Mandarin listeners take advantage of post-stop F0 as a cue for stop voicing, the experiment with which this observation was made did not require the listeners to simultaneously track F0 for lexical tone, so it is therefore still an open question whether Mandarin listeners actually use post-stop F0 as a cue for stop voicing in more natural settings.

The second aim of this study is to investigate whether the use of post-stop F0 cue is sensitive to different language contexts. Capitalizing on the fact that the L1 Mandarin speakers that could be recruited in the university communities here were also L2 English speakers, one relevant question is whether Mandarin-English bilinguals use post-stop F0 cue to different extents, depending on the language "mode" they are operating in. If post-stop F0 is not solely due to physiological and/or aerodynamic reasons and is partially subject to active controlling, as postulated in Kingston and Diehl (1994), Mandarin-English bilinguals might actively, though subconsciously, suppress post-stop F0 in Mandarin because of the pressure to maintain tonal contours, which they do not have to do when speaking English. In perception, the demand to track F0 for lexical tone when perceiving Mandarin might prompt the bilingual listener to attribute variation in F0 partially to lexical tone, which makes them less likely to treat variation in post-stop F0 as an indicator for voicing. However, freed from the burden of tracking F0 for tone, as when they are perceiving English, the same listeners now have more certainty in linking the difference in post-stop F0 to consonantal voicing. These two scenarios could lead to bilinguals using the post-stop F0 cue differentially in both production and perception, which would be reflected as different cue weights for post-stop F0 that depend on the language. On the other hand, given that the bilinguals are dominant in Mandarin, they may simply import their cue-weighting strategies for Mandarin to English, as predicted by the SLM and PAM-L2, resulting in the same weight for post-stop F0, regardless of language. The hypotheses and the corresponding predicted results just described are summarized in Table 1. The conducted production and perception experiments can help distinguish between the two possibilities.

An additional aspect that is foregrounded in this study is individual variability in participants' production and perception in their L1 and L2. Specifically, the relationship between individual participants' production and perception of post-stop F0 is explored. For this purpose, individual participants' production and perceptual post-stop F0 weights in their L1 and L2 are derived first. Correlation analyses are then used to examine whether individuals' post-stop F0 weights are statistically linked either within the same modality but across languages, or within the same language but across modalities.

3. Production experiment

This experiment examined non-early Mandarin-English bilinguals' productions of Mandarin and English word-initial stops and sonorants on vowel-onset F0.

3.1. Participants

All participants were recruited from the linguistic participant pools at the University of British Columbia or the University of Toronto, and they received partial course credit for participation. A total of 103 participants completed the experiment, but only a subset of 25 L1 Mandarin-L2 English bilingual participants (14 female, 11 male; $\text{Mean}_{\text{age}} = 20.9$ years, $\text{SD}_{\text{age}} = 2.1$ years) were analyzed. The inclusion criteria are detailed below. For their production data to be considered in the analyses, a participant must satisfy all of the following criteria:

1. They completed all required experiment components;
2. They self-report as a native speaker of Mandarin;
3. They have at least one primary caretaker whose native language is Mandarin;
4. They are not simultaneous/early/childhood bilingual in Mandarin and English (i.e., they were exposed to English only after entering elementary school and did not receive their formal education in English prior to high school or university);
5. They lived in China for at least 10 years between birth and age 15.

A number of additional inclusion guidelines, which are based on their audio recording quality and their performance in the perception experiment, were applied to make sure that only high-quality data was included in the analyses. These detailed inclusion guidelines are given in Sections 3.6 and 4.4, respectively. As a preview of these additional criteria, three participants were excluded due to suboptimal recording quality, and only the data from the participants who were attentive throughout the perception experiment was included.

3.2. Stimuli

This section describes the principles behind the selection of Mandarin and English production stimuli. The same logic was used for both languages, with adaptations to accommodate the phonotactic constraints of each language.

3.2.1. Mandarin stimuli

The Mandarin stimuli consisted of 27 monosyllabic Mandarin words in isolation, as provided in Supplementary Table 1. These words had onsets that exemplified the two laryngeal categories—voiceless aspirated and voiceless unaspirated—in Mandarin, as well as the sonorants /m/, /n/, and /l/. The sonorants were included to serve as the baseline against which the phonological voicing of stops was compared. To increase the generalizability of the findings, words with stops at three places of articulation (i.e., labial, alveolar, and velar), crossed with two levels of vowel heights (high: /i/, low: /a/, embedded in /aɪ/; /aɪ/, as opposed to /a/, was used because words with /aɪ/ are phonetically more similar to the English words used in the English production counterpart; see Section 3.2.2), were included. Given that lexical tone has been reported to modulate F0 perturbation in Mandarin (Guo, 2020), and that the influence of individual lexical tones is outside the scope of the current study, only Tone 1 and Tone 4 syllables were considered. Both tones start with a high pitch register and have been found to pattern together in conditioning post-stop F0 perturbation, making their production data more comparable to each other. Note also the existence of systematic and accidental gaps that prevented a fully crossed combination of the onsets, vowels, and tones. For instance, Mandarin disallows the occurrence of a velar stop before a high front vowel, so syllables such as */k^{h}i/ and */ki/ are missing in Mandarin altogether. It is, however, accidental gaps in the language that cause */maɪ˥/, */ni˥/, etc., to be absent.

The stimuli were presented to the participants in simplified Chinese characters. Given that Mandarin has a large number of homophones that are nonetheless distinguished by different characters, each stimulus was represented with a common character so that all of them should be familiar to the participants, with the exception of *kai4* 忾, which is not a highly frequent character. To make sure that the participant knew the pronunciation of this character, its pinyin <kai4> was added to the right side of this character when presented to the participant. Care was also taken to ensure that different characters were as visually distinct as possible, to avoid the potential confound from visual priming across trials. For instance, while *pi1* could be represented with both 披 and 批, 披 was chosen because 批 shares the component 比 with another stimulus *pi4* 屁.

3.2.2. English stimuli

The English stimuli consisted of 19 monosyllabic words, as given in Supplementary Table 2. These words were selected following the same principles of stimulus section for the Mandarin tokens: the onsets typified voiceless stops, voiced stops, and sonorant at labial, alveolar, and velar places, while the vowels were either the front high vowel /i/ or the diphthong /aɪ/. When a simple combination of an onset and an open vowel did not correspond to a common English word, another common word with the same onset and nucleus but with an additional voiceless-stop coda was used as the alternative. Voiceless-stop codas, instead of other consonant classes, were used because they formed common English words. Also, for the syllable /di/, both the letter *D* and the word *deep* were used as stimuli to prevent loss of data for /di/ due to the participant not producing /di/ upon seeing *D*.

3.3. Procedure

The procedure was identical for both the Mandarin and English versions of the experiment, and the order in which the two versions were administered was counterbalanced across participants. The entire experiment took place online in response to constraints on in-person data collection due to COVID, with the participant being instructed to complete the experiment on their own computer in a quiet place. They were encouraged to use an external microphone to keep the fidelity of audio recordings as high as possible, though they could still participate using the built-in microphone on their device.

The experiment was implemented in jsPsych, version 6.1.0 (de Leeuw, 2015). The experiment started with a microphone check to ensure that the input source was set correctly, and that the recording was clear. The experimental trials commenced after three practice trials that aimed to familiarize the participant with the recording interface and experimental flow. Each stimulus was repeated three times in three blocks, respectively with a self-timed break between blocks. Stimuli were presented in a randomized order within each block. Each trial began with a plus sign at the center for 500 ms, and the recording was initiated automatically at the same time. The stimulus then appeared at the center, replacing the plus sign, and the participant was asked to read aloud the stimulus in a clear and natural manner. The trial ended with the participant clicking the "submit" button, which stopped the recording, uploaded the audio file to the server, and triggered the next trial. In the event where the participant did not click anything, the trial would terminate on its own after 10 s. The entire production experiment lasted about 15 min.

3.4. Recording annotations

All annotations and measurements were performed in Praat (Boersma and Weenink, 2021). The portion of the signal analyzed spanned from the beginning of the onset consonant to the end of the third pitch cycle of the nucleus vowel. The following guidelines were used when annotating tokens produced in either language.

1. *Beginning of stop closure voicing*: In the cases where there was prevoicing for tokens with a voiced stop in English or, very rarely, with an unaspirated stop in Mandarin, all simple periodic chunks of the waveform before the release of the onset stop were marked as stop closure voicing.
2. *Beginning of stop burst*: For tokens with a stop onset, the beginning of the burst was marked at the starting point of perturbation in the waveform.
3. *Vowel onset*: The vowel onset was operationalized as the point where the (quasi) periodic part of the vowel first crossed zero in the positive direction.
4. *End of the third pitch cycle*: Following Cole et al. (2007) and Clayards (2018), the point marking the first 3 pitch cycles as counted from vowel onset was pinned in order to derive the onset F0.

3.5. Acoustic measurements

1. *Voice Onset Time (VOT)*: In line with the typical definition, VOT is defined as the time difference between the release of the stop and the onset of voicing (pre- or post-release). Accordingly, for prevoiced tokens (i.e., those with the beginning of stop closure voicing marked) VOT took a negative value, while VOT was positive for tokens where the onset of vocalic voicing followed the stop release. Tokens where the onset of vocalic voicing coincided with the stop release had a VOT of 0 ms.
2. *Onset fundamental frequency (F0)*: This measurement was obtained by dividing 3 by the duration of the first 3 pitch cycles from vowel onset [i.e., 3 / (end of the third pitch cycle − vowel onset)]. No F0-tracking algorithm was therefore involved for this measurement.

3.6. Participant inclusion criteria

Participants whose entire recordings (i) contained excessive background noise due to their doing the experiment in a noisy place ($n = 1$), (ii) were extremely soft that made it challenging to identify acoustic landmarks for annotation ($n = 1$), or (iii) were of extremely low sampling rates ($n = 1$), were omitted from the dataset altogether. There were also three participants who attempted the experiment more than once; in such a case, only the recordings from their first experiment attempt were considered. A subset of 25 participants was then selected based on their performance in the perception experiment, as explained in Section 4.4.

3.7. Omitted data

Among the tokens produced by the 25 included participants, the following tokens were excluded from all analyses: mispronunciations (11 Mandarin and 26 English), skipped tokens (2 Mandarin and 3 English), and technical issues (2 Mandarin and 4 English, including sporadic silent periods that overlapped with stop burst and/or vowel onset). Furthermore, tokens with creaky voice at vowel onset, for which F0 estimation was therefore unreliable, were also omitted from all analyses (50 Mandarin and 33 English). Overall, 131/3,450 = 3.8% of the production tokens were excluded.

3.8. Statistical analyses

The analyses consisted of two major parts: the first part addressed whether post-stop F0 had different values across the onset types in each language, and the second part focused on the quantification of production weight for post-stop F0 in each language. All models were fitted with Bayesian mixed-effects models, using `CmdStanR` (Gabry and Češnovar, 2021), an R interface for the Stan probabilistic programming languages (Carpenter et al., 2017). Bayesian models were chosen because they return a distribution of potential values for all model parameters, making it more intuitive to assess the uncertainty associated with each parameter. In what follows, details about the statistical model employed are described.

3.8.1. Post-stop F0 models

In this set of analyses, post-stop F0 was modeled as a Gaussian linear function of a number of variables that were properties of tokens or speakers. The names of predictor variables are given **boldface**, and different levels within a variable are indicated in SMALL CAPS.

3.8.1.1. Variables

The dependent variable in all models was *z*-transformed post-stop F0. The post-stop F0 values from both Mandarin and English production were *z*-transformed within each speaker. That is, a single *z*-transformation was applied to Mandarin and English production data together for each speaker.

Four token-level predictors were considered: the **voicing** of the onset consonant, **language/tone**, the **height** of the main vowel, and the **place of articulation** (**PoA**) of the onset consonant. Forward difference coding was used for **voicing** (ASPIRATED vs. UNASPIRATED and UNASPIRATED vs. SONORANT). Helmert coding was used for **language/tone** (ENG vs. mean of MAN T1 and MAN T4, and MAN T1 vs. MAN T4). Sum coding was used for **height** (HIGH, NON-HIGH = [1, −1]) and **PoA** (LABIAL, ALVEOLAR, VELAR, with LABIAL coded with −1). To account for how each predictor affected the realization of the voicing contrast, two-way interaction terms between **voicing** and all the other predictors were also included in the model comparison process. These first-order and second-order terms therefore constituted the population-level ("fixed-effect") predictors.

For individual-level ("random-effect") predictors, by-**speaker** effects consisted of a random intercept and random slopes for all population-level predictors.

3.8.1.2. Model structure

Standardized post-stop F0 was modeled as a function of a subset of the predictor variables introduced above, using Bayesian linear mixed-effects models. All candidate models shared general specifications. Main-effect terms were included for the predictor variables selected in a particular candidate model. As mentioned above, two-way interaction terms being **voicing** and the other predictors were also considered. I did not, however, consider any three-way interactions as they are in general harder to interpret and could drastically slow down model sampling. All models also included by-speaker random intercepts, to account for variability in post-stop F0 of speakers beyond the effects of predictor variables. All possible by-speaker random slopes were also included to account for variability among speakers in the effects of predictors on post-stop F0 (Barr et al., 2013).

Each model was fitted with regularizing priors of Normal($\mu = 0, \sigma = 5$) for the intercept and all population-level parameters. An Exponential($r = 1$) distribution was used as the prior for the error term as well as for the individual-level standard deviations. Correlations among individual-level effects used the LKJ prior (Lewandowski et al., 2009) with $\xi = 1$, in order to give lower prior probability to perfect correlations. All models showed no divergent transitions and had $\hat{R}$ values close to 1 (i.e., all $\hat{R} < 1.01$), which indicates that chains were well-mixed.

3.8.1.3. Inference criteria

Evidence embedded in each model was evaluated in two ways: (i) the posterior distributions of parameters, and (ii) comparison of models of different complexities. In particular, I consider there to be strong evidence for a non-null effect if the 89% credible interval (CrI)—the narrowest interval that contains 89% of the posterior density—for the parameter does not include 0. If the 89% CrI spans 0, but the probability of the parameter not changing direction is at least 89%, I consider this to represent weak evidence for a given effect. The decision to use CrIs of 89%, as opposed to 95%, is based on Koster and McElreath (2017) and McElreath (2020), to discourage the association between a Bayesian posterior distribution and a *p*-value. Model comparison was done by means of the Bayesian leave-one-out estimate of expected log pointwise predictive density (ELPD-LOO; Vehtari et al., 2017), which aims to gauge a model's *predictive* accuracy (i.e., how close predicted values from a model are to the raw data). A higher ELPD-LOO value means that the model has a better predictive accuracy. The results from model comparison thus inform us whether a variable contributes substantially to a model's predictive power. Following Sivula et al. (2020), when the estimated absolute difference in ELPD-LOO between two models is at least 4, and 0 is not within two

TABLE 2 Candidate post-stop F0 models considered in model comparison, with their ELPD-LOO means and standard errors.

Model	ELPD-LOO mean	ELPD-LOO standard error	Predictors
M1	−3637.3	60.3	height + lang/tone
M2	−3221.8	67.3	height + lang/tone + voi
M3	−3215.5	67.3	height + lang/tone + voi + PoA
M4	−3205.5	68.2	height + lang/tone + voi + voi × height
M5	−3189.0	67.8	height + lang/tone + voi + voi × lang/tone
M6 (final)	−3173.4	68.7	height + lang/tone + voi + voi × height + voi × lang/tone
M7	−3174.3	69.2	height + lang/tone + voi + voi × height + voi × lang/tone + voi × height × lang/tone

An intercept was included in each model but is omitted here in the table to save space.

standard errors of the estimated difference, there is evidence that the two models give different predictions.

In the following sections, model parameters are reported in terms of marginal posterior means of parameters, 89% CrIs, and the probability of effect direction.

3.8.1.4. Candidate models

The construction of candidate models for model comparison relied both on prior knowledge about factors affecting post-stop F0 and on a compromise between model complexity and predictive accuracy. All the candidate models are given in Table 2. Given that vowel height is known to influence F0 ("intrinsic F0," Whalen and Levitt, 1995) and that language and lexical tone can affect F0, the base model (i.e., M1) started with the factors **height** and **language/tone**. As one of the goals is to establish whether and how post-stop F0 might be influenced by phonological voicing, further models were constructed by incrementally adding terms that involved **voicing**. For example, the comparison between M1 and M2 assessed the contribution of voicing in predictive accuracy, and comparing M2 and M4 examined the importance of the interaction between voicing and vowel height in predicting post-stop F0 values. Furthermore, a model with **PoA** as a predictor (i.e., M3) also entered into comparison to confirm that place of articulation does not cause post-stop F0 to differ. The formal specification of the final model can be found in the Supplementary material.

3.8.2. Post-stop F0 production weight model

The second set of analyses aimed to quantify the production weight associated with post-stop F0. A higher production weight means post-stop F0 is more reliable in separating different members of the contrast. Following Clayards (2018), the production weight was calculated based on the amount of overlap between the categories, which was quantified using Cohen's *d* (Cohen, 1988):

$$d = \frac{\mu_{\text{asp}} - \mu_{\text{unasp}}}{\sqrt{1/2\left(\sigma^2_{\text{asp}} + \sigma^2_{\text{unasp}}\right)}},$$

where μ_{asp} and μ_{unasp} refer to the mean F0s of the aspirated and unaspirated categories, respectively, and σ^2_{asp} and σ^2_{unasp} are the standard deviations of F0 of the aspirated and unaspirated categories, respectively.

Cohen's *d* for post-stop F0 was calculated at the population level with all speakers as a whole and at the individual level for each speaker. Only tokens produced with a positive VOT were included in the calculation, as negative VOTs were rare in the data (i.e., 9 tokens from 1 speaker in Mandarin, and 40 tokens from 5 speakers in English) and therefore were not representative of the norm of this speaker population. Additionally, rather than estimating cue weights from empirical data as in most previous work (e.g., Shultz et al., 2012; Schertz et al., 2015; Clayards, 2018), a statistical model was used to derived the weight, which allowed for uncertainty around the weight to be incorporated. For this purpose, a Bayesian mixed model was first fitted to obtain the means and standard deviations of F0 of the aspirated and unaspirated categories for the whole group and for each speaker. The model included a cross-category correlation structure and used partial pooling to estimate individual means and standard deviations. For instance, a speaker's mean post-stop F0 for the aspirated category was correlated with their mean post-stop F0 for the unaspirated category, and both mean values were informed not only by the speaker's own production data, but also by other speakers' data thanks to partial pooling. The estimated means and standard deviations were then fed to the Cohen's *d* formula above to calculate the production weight within the model. As such, the post-stop F0 weights of the entire group and for each speaker were not just a single numerical value but a *distribution* that also carried information about uncertainty. The formal specification of the model is included in the Supplementary material.

3.9. Results: Production of post-stop F0

Mean production values and standard deviations for L1 Mandarin and L2 English stops and sonorants on VOT and post-stop F0 are given in Supplementary Table 3. Distributions

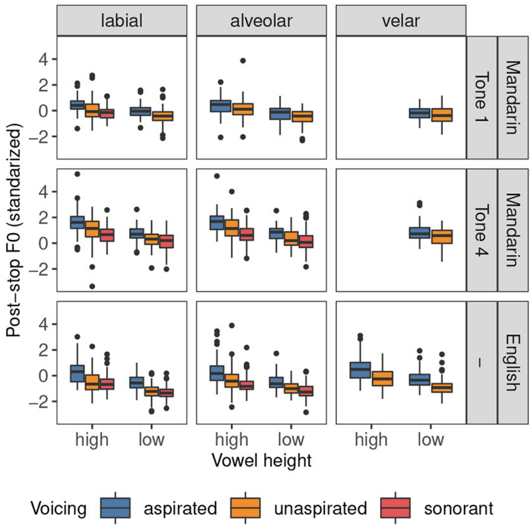

FIGURE 1
Standardized post-stop F0 values, normed by speaker, as a function of place of articulation, language/tone, vowel height, and phonological voicing.

of standardized post-stop F0 values are plotted in Figure 1. ELPD-LOO means and standard errors for the candidate models are listed in Table 2. A higher ELPD-LOO value means the model has a better predictive accuracy, so, for example, M2 makes better predictions than M1. Finally, model comparison results are summarized in Supplementary Table 4 in terms of difference in ELPD-LOO values and associated standard errors. Note that the difference score in each cell was computed by subtracting the ELPD-LOO value of the model represented in the column from the ELPD-LOO value of the model indicated in the row. For instance, the difference −415.5 came from ELPD-LOO$_{M1}$ − ELPD-LOO$_{M2}$ = (−3637.3) − (−3221.8).

The results of model comparison indeed confirmed the importance of phonological voicing in conditioning post-stop F0 (i.e., M1 vs. M2) and spoke to the importance of interaction between voicing and vowel height (i.e., M2 vs. M4), and between voicing and language/tone (i.e., M2 vs. M5). Place of articulation, however, did not seem to influence post-stop F0 (i.e., M2 vs. M3). Since no significant gain in prediction was observed past M6, M6 was selected as the best balance between model complexity and predictive performance among the models being compared. The interpretation and discussion presented below are therefore based on this model.

In presenting the results, summary statistics and visualizations derived from raw data are given first, followed by the output from the final model in terms of posterior distributions for key parameters. I first interpret population-level parameter estimates before moving on to individual-level estimates.

3.9.1. Population results

The marginal posterior distributions for population-level parameters from M6 are summarized in Table 3. As expected, both vowel height and language/tone contribute to difference in post-stop F0. Specifically, the high vowel /i/ led to a higher onset F0 (HIGH − mean height: $\bar{\beta} = 0.32$, 89% CrI = [0.27, 0.36], $p(\beta > 0) = 1.00$), and Tone 4 tended to have a higher onset F0 than Tone 1 (MAN T1 − MAN T4: $\bar{\beta} = -0.91$, 89% CrI = [−1.03, −0.79], $p(\beta < 0) = 1.00$). Also, participants' L2 English tended to have a lower onset F0, in comparison with their L1 Mandarin (ENG − (MAN T1 + MAN T4)/2: $\bar{\beta} = -0.84$, 89% CrI = [−1.02, −0.66], $p(\beta < 0) = 1.00$), which agrees with the general finding from the literature (Keating and Kuo, 2012; Lee and Sidtis, 2017). Critically, in both languages, aspirated stops had a higher post-stop F0 than unaspirated stops (ASP − UNASP: $\bar{\beta} = 0.49$, 89% CrI = [0.41, 0.56], $p(\beta > 0) = 1.00$), which in turn had a higher post-stop F0 than sonorants (UNASP − SON: $\bar{\beta} = 0.29$, 89% CrI = [0.20, 0.39], $p(\beta > 0) = 1.00$). In addition, the extent of post-stop F0 difference due to aspiration was contingent on language and tone as well, such that bilingual speakers' English tokens showed an even bigger difference than Mandarin tokens ([ASP − UNASP] × [ENG − (MAN T1 + MAN T4)/2]: $\bar{\beta} = 0.25$, 89% CrI = [0.10, 0.39], $p(\beta > 0) = 1.00$), and so did their Mandarin Tone 4 tokens in comparison with Tone 1 tokens ([ASP − UNASP] × [MAN T1 − MAN T4]: $\bar{\beta} = -0.16$, 89% CrI = [−0.28, −0.05], $p(\beta < 0) = 0.99$).

3.9.2. Individual results

The distributions for key parameters involving voicing for each participant are visualized in Figure 2. In both their Mandarin and English productions, there is strong evidence that all speakers produced a higher post-stop F0 following an aspirated stop than an unasiprated stop, as the 89% CrI is above 0 for all speakers in the [ASP − UNASP] panel in Figure 2. The [UNASP − SON] panel indicates that, for the majority of speakers (18 out of 25), the model is also confident that their onset F0 was higher adjacent to an unaspirated stop than adjacent to a sonorant. For the remaining speakers, even though their 89% CrIs span 0, their posterior means are still above 0, suggesting that, on average, their F0 patterns conform to the general trend. In terms of the post-stop F0 difference due to aspiration, about half of the speakers (13) evidently agree with the population pattern in having a bigger F0 difference in English, as indicated by their positive 89% CrIs in the [(ASP − UNASP) * LANG] panel. For the other speakers, there does not seem to be a consistent trend, as even the posterior means are going in different directions. Finally, as shown in the [(ASP − UNASP) * TONE] panel, even though only seven speakers clearly followed the observation at the population level that Tone 4 supported a more differentiated post-stop F0 distinction between aspirated and unaspirated stops, the other speakers also trend in this direction.

TABLE 3 Marginal posterior summaries for key population-level parameters from M6.

Parameter	Mean	SD	89% CrI	*p*(dir.)
intercept	0.01	0.01	[−0.01, 0.04]	$p(\beta > 0) = 0.84$
HIGH − (HIGH + LOW)/2**	0.32	0.03	[0.27, 0.36]	$p(\beta > 0) = 1.00$
ENG − (MAN T1 + MAN T4)/2**	−0.84	0.11	[−1.02, −0.66]	$p(\beta < 0) = 1.00$
MAN T1 − MAN T4**	−0.91	0.07	[−1.03, −0.79]	$p(\beta < 0) = 1.00$
ASP − UNASP**	0.49	0.05	[0.41, 0.56]	$p(\beta > 0) = 1.00$
UNASP − SON**	0.29	0.06	[0.20, 0.39]	$p(\beta > 0) = 1.00$
[ASP − UNASP] × [HIGH − (HIGH + LOW)/2]*	0.05	0.04	[−0.01, 0.12]	$p(\beta > 0) = 0.91$
[UNASP − SON] × [HIGH − (HIGH + LOW)/2]	0.04	0.04	[−0.02, 0.10]	$p(\beta > 0) = 0.86$
[ASP − UNASP] × [ENG − (MAN T1 + MAN T4)/2]**	0.25	0.09	[0.10, 0.39]	$p(\beta > 0) = 1.00$
[ASP − UNASP] × [MAN T1 − MAN T4]**	−0.16	0.07	[−0.28, −0.05]	$p(\beta < 0) = 0.99$
[UNASP − SON] × [ENG − (MAN T1 + MAN T4)/2]	−0.03	0.08	[−0.15, 0.09]	$p(\beta < 0) = 0.64$
[UNASP − SON] × [MAN T1 − MAN T4]	0.00	0.10	[−0.16, 0.15]	$p(\beta < 0) = 0.52$

The contrast coding scheme for each variable is described in Section 3.8. The parameters whose effects are judged to be strong are marked with **, and those whose effects are judged to be weak are marked with *.

3.10. Results: Production weights of post-stop F0

Standardized post-stop F0 values are plotted against raw VOT values for participants' Mandarin and English productions in Supplementary Figure 1, and the distributions of production VOT and post-stop F0 weights, expressed in terms of Cohen's *d*, at the population level are graphed in Figure 3. Although the focus on this study is on the post-stop F0 cue, for completeness, the results for the VOT weight are also reported below.

3.10.1. Population results

As can be seen in Figure 3, speakers as a group had a much higher weight for VOT than for post-stop F0, in both their Mandarin and English production. Also, regardless of language, there was more uncertainty surrounding the post-stop F0 weight than the VOT weight, as measured by the coefficient of variation (CV), which is defined as the ratio of the standard deviation to the mean (English: $\mathrm{CV_{VOT}} = 0.06$, $\mathrm{CV_{F0}} = 0.18$; Mandarin: $\mathrm{CV_{VOT}} = 0.06$, $\mathrm{CV_{F0}} = 0.17$). Contrasting the weights along the same dimension across languages, more weight was assigned to VOT in the Mandarin production (89% CrI = [6.34, 7.60]), as compared to the English production (89% CrI = [4.78, 5.82]), while the converse was true for the post-stop F0 weight: English tokens showed a heavier reliance on post-stop F0 (89% CrI = [0.70, 0.99]) than Mandarin tokens (89% CrI = [0.34, 0.54]).

3.10.2. Individual results

The reliability of each dimension for individual speakers, as estimated by Cohen's *d*, is plotted in Figure 4. Conforming to the population pattern, all speakers assigned more weight to VOT than post-stop F0 in both their Mandarin and English productions (Figure 4A). When correlating weights along the two dimensions within language, no specific correlation pattern was discernible (see Figure 4B; Mandarin: 89% CrI of $\rho_{\mathrm{VOT_{Man}},\mathrm{F0_{Man}}} = [-0.35, 0.15]$; English: 89% CrI of $\rho_{\mathrm{VOT_{Eng}},\mathrm{F0_{Eng}}} = [-0.25, 0.21]$). However, when the VOT weights were correlated across languages, a strong positive correlation was observed (89% CrI of $\rho_{\mathrm{VOT_{Man}},\mathrm{VOT_{Eng}}} = [0.40, 0.81]$), indicating that speakers who showed a larger VOT weight in Mandarin also tended to have a larger VOT weight in English (Figure 4C). In addition, for all but one speaker, VOT had more weight in their Mandarin tokens than their English tokens. For the post-stop F0 weight, most individuals (19 out of 25) echoed the population pattern in shifting their F0 weight upward when producing English tokens (Figure 4D), although there was no correlation in this cue across languages (89% CrI of $\rho_{\mathrm{F0_{Man}},\mathrm{F0_{Eng}}} = [-0.49, 0.40]$). Also notice that there was more individual variation for the post-stop F0 weight in the English production than in the Mandarin production, as indicated by a wider spread of individual weights in English than in Mandarin.

3.11. Interim discussion: Production

The Mandarin production results reported here are in line with the recent work by Guo (2020) in terms of post-stop F0: both at the population and individual levels, the vowel-onset F0 following aspirated stops was higher than that following unaspirated stops. In addition, for most speakers, vowel-onset F0 after unaspirated stops was in turn higher than that after sonorants. Similar to their Mandarin production, the participants' English production also demonstrated a difference in post-stop F0 between aspirated and unaspirated series, but

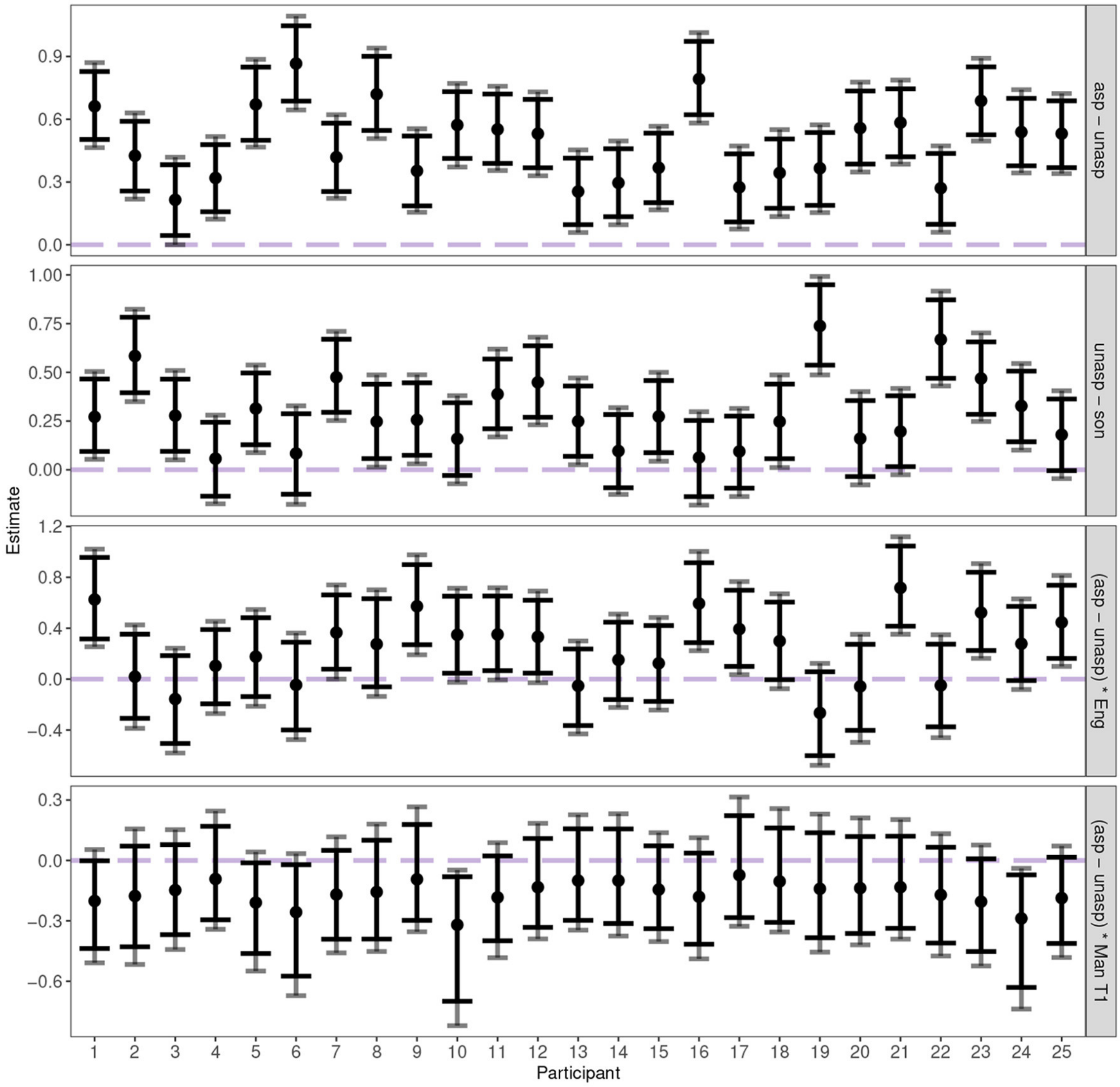

FIGURE 2
Marginal posterior summaries for key parameters involving voicing for each individual speaker. The [asp − unasp] panel shows the difference in F0 between aspirated and unaspirated stops. The [unasp − son] panel shows the difference in F0 between unaspirated stops and sonorants. The [(asp − unasp) * Eng] panel shows the further difference in F0 between aspirated and unaspirated stops in English, in comparison to Mandarin. The [(asp − unasp) * Man T1] panel shows the further difference in F0 between aspirated and unaspirated stops in Mandarin Tone 1 tokens, when compared to Tone 4 tokens. The dots denote the posterior means. The inner error bars represent 89% CrIs, and the outer error bars represent 95% CrIs.

with an even larger F0 gap, both for the speakers as a whole and for over half of the individual speakers. This pattern again agrees with what has been found in Guo (2020).

Regarding cue weighting, VOT was the most reliable dimension distinguishing aspirated from unaspirated stops in both Mandarin and English, though it seemed that VOT assumed an even higher weight in Mandarin for almost all speakers (as measured by the posterior mean). The opposite pattern was observed for the post-stop F0 weight: English induced a higher weighting in this cue for most speakers. When the weighting between the two cues was correlated within each language, however, neither an enhancing nor a trading relationship was obtained.

4. Perception experiment

The perception experiment turns to the perception of the Mandarin and English stop contrasts in the word-initial position by the same L1 Mandarin-L2 English bilinguals. The focus is on the contribution of post-stop F0 to categorization of the contrasts.

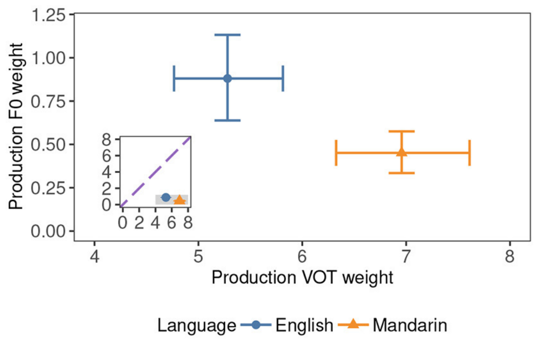

FIGURE 3
Group-level production weights for VOT and post-stop F0 in each language. The embedded plot shows the same data but with the same scale along both axes, and the dashed line is $y = x$, which represents equal production weights for both dimensions. The shaded area indicates the part enlarged in the main plot.

4.1. Participants

The same group of participants from the production experiment also took part in the perception experiment. The perceptual data analyzed here came from the same 25 participants whose production tokens were analyzed in the production experiment.

4.2. Stimuli

All stimuli were created from natural productions of the Mandarin words *bi1*, *pi1*, *bi4*, *pi4*, *yi1*, *mi1*, *mi4*, and *ni4* read by a 24-year-old male English-Mandarin speaker who speakers English as L1 but is also fluent in Mandarin. The prompts for production were words in isolation, which were presented three times to the model speaker in a randomized order. The recording was made on the Sound Devices MixPre-D audio mixer with a headset microphone. The produced syllables were then scrutinized by the author, and one token that was clear and did not have creaky quality was selected for each word as the raw tokens for manipulation.

4.2.1. Mandarin stimuli

Stimuli could be categorized into the target or filler sets, with both sets containing Tone 1 and Tone 4 syllables. The target set was composed of syllables with a bilabial stop as the onset and the high vowel [i] as the nucleus, with the VOT of the stop and the initial F0 contour of the vowel manipulated. The manipulation along the VOT and F0 dimensions is summarized in Figure 5 and explained in the following paragraphs. Bilabial stops were used because they do not have lingual targets and therefore are expected to be coarticulated to a lesser degree with the following vowel (Schertz et al., 2020). The vowel

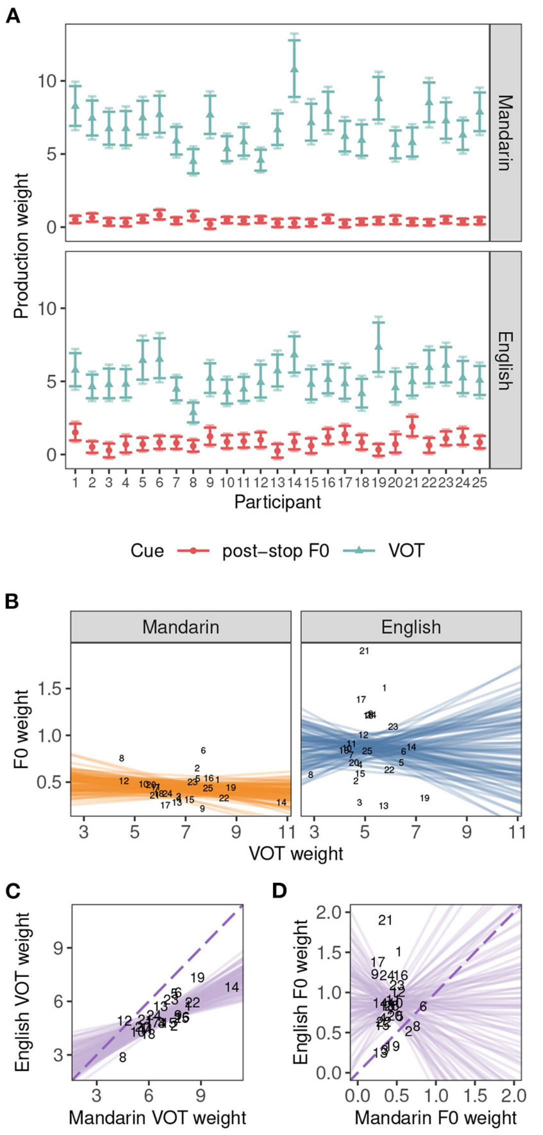

FIGURE 4
(A) Individual speakers' production weights for VOT and post-stop F0. The posterior means are represented by the dots. The 89% CrIs are represented by the inner error bars, and the 95% CrIs are represented by the outer error bars. **(B)** Post-stop F0 weights against VOT weights, separately for each language. **(C)** Production VOT weights across languages. **(D)** Production post-stop F0 weights across languages. In **(B–D)**, the solid lines represent 100 regression lines fitted with 100 posterior draws, to show the direction and uncertainty of the correlation. The dashed line is $y = x$, where the Mandarin weight equals the English weight.

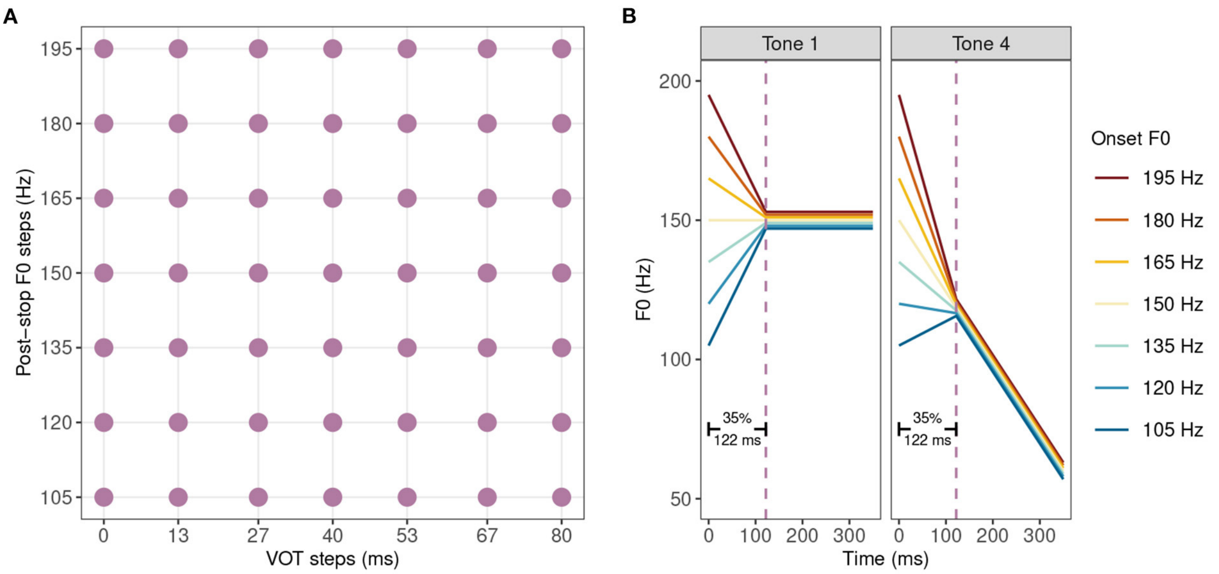

FIGURE 5
Manipulation of target stimuli for all perception experiments. **(A)** Each dot represents one stimulus, with its *x* coordinate corresponding to the VOT of the initial labial stop, and its *y* coordinate to the initial F0 of the following vowel. **(B)** Illustration of F0 trajectory manipulation for target syllables. Note the vowel duration in actual stimuli is not necessarily 350 ms due to the trade-off between VOT and vowel duration, which was also manipulated. The invariant parts across different tokens (i.e., after 122 ms) are shifted vertically in the figure for visual clarity only.

[i] was selected because its formants are more stable across time in general (Hillenbrand et al., 1995)[1]. In addition, the combination of bilabial stops with the high vowel also led to valid English lexical items *pea* and *bee*; this was critical given that the exact same stimuli were used in the English version of the experiment as well. For fillers, Mandarin words *yi1*, *yi4*, *mi1*, and *mi4* were selected because they typified other onset types than the stop.

The target syllables were created by cross-splicing the vocalic portion of the *bi1* token and the burst+aspiration portion of the *pi1* token. The detailed steps of stimulus manipulation are described below.

The first step involved creating a Tone 1 and a Tone 4 base token for downstream manipulation. The vowel duration of the *bi1* token was set to 350 ms, which is approximately the mean duration of 416.2 ms for citation Tone 1 syllables and 307.8 ms for citation Tone 4 syllables (Yang et al., 2017)[2]. The vowel duration was shifted to an ambiguous value to discourage the participant to use it as an additional cue for tone identification (e.g., Blicher et al., 1990). F0 trajectories were then manipulated to mimic natural Tone 1 and Tone 4 contours. For Tone 1, a simple pitch stylization was applied by setting both the initial and final F0 on the vowel to 150 Hz. The F0 was set to 150 Hz because this was very close to the natural Tone 1 F0 register of this particular token. Tone 4 was stylized as a linear F0 decline from 150 Hz to 60 Hz. The initial 150 Hz was to match the initial F0 value for Tone 1 while the final 60 Hz was set based on the model speaker's natural Tone 4 production. The decision to recreate Tone 4 F0 contour from a Tone 1 item, instead of using a natural Tone 4 item, was to make sure that the same intensity profile was shared and would not be a confound[3].

[1] The vowel [i] was also preferred from the perspective of VOT manipulation. Given that the starting values of the formant frequencies in the voiced part of the vowel could be substantially different depending on VOT, stimuli whose VOT values are manipulated with a "progressive cutback and replacement" approach (which was also used in this study) can have initial formant frequencies being correlated with VOT, leading formant cues to be a confound. Winn (2020) argues that since F1 of [i] is already low, the upward F1 transition common to the other vowels would be minimized, thus offering no covarying cue for VOT.

[2] These measurements are based on production of isolated monosyllables by 121 speakers (46 male and 75 female). Note that even though there seems to be a 100-ms difference between Tone 1 and Tone4, both tones have a standard deviation of about 90 ms in the syllable duration measurement, suggesting that the two tones overlap to a large extent in terms of their duration distributions.

[3] I have also attempted to create base tokens in the opposite direction: creating a Tone 1 item from a Tone 4 item. However, the resulting audio was noticeably unnatural, especially in the later portion where F0 needed to be raised from a low target of Tone 4 to a high target of Tone 1.

The second step scaled the intensity of the two base tokens to 75 dB based on the root-mean-square (RMS) amplitude. The level 75 dB was chosen because this was approximately the intensity of the raw recording. Intensity normalization was done at this step, as opposed to at a later point when actual stimuli were synthesized, because Winn (2020) cautions that "the inclusion of a lengthy aspiration portion will justifiably reduce overall RMS intensity, so equalization would result in unnatural amnlification of the syllable with voiceless onset" (p. 859). He therefore suggests that intensity amplification/attenuation should be applied before initiating VOT manipulation.

In the last step, the two intensity-equalized tokens were then modified, using a Praat script prepared by Winn (2020), to create tokens varying in VOT duration and F0 at vowel onset. The duration of VOT in the base tokens was manipulated on a 7-step series ranging from 0 ms to 80 ms. The range endpoints were meant to span the VOTs of both English and Mandarin word-initial bilabial stops while still having enough resolution. Note that negative VOT was not in the manipulated range partially because "voiced" stops in word-initial position in English are very often realized as a short-lag stop with positive VOT (Fulop and Scott, 2021) and partially because including negative values would decrease the manipulation resolution. VOT was manipulated with a progressive-cutback-and-replacement approach—that is, "the onset of a word with a voiced stop sound is progressively deleted and replaced with a roughly equivalent amount of the onset from its voiceless-onset counterpart" (Winn, 2020, p. 854)—to accommodate the observation that there tends to be an inverse relationship between VOT and duration of the following vowel (Summerfield, 1981). However, to approximate this inverse relationship in natural production, the extent of vowel shortening was not entirely commensurate with changes in VOT, that is, for every 1 ms of VOT increase, the vowel was shortened by less than 1 ms (Allen and Miller, 1999; Toscano and McMurray, 2010). The default vowel-VOT ratio of 0.65, which is the default value of Winn's (2020) script, was used for modeling this trade-off relation. The initial F0 was set at one of the seven values, from 105 Hz to 195 Hz with a step size of 15 Hz, at the beginning of the vowel. F0 then rose/fell linearly for the following 122 ms (or 35% of the vowel duration) to 150 Hz for Tone 1 stimuli and to about 118 Hz for Tone 4 stimuli. The step size was set to 15 Hz so that the difference in F0 would be large enough to be noticeable but not too large so as to distort the F0 trajectory significantly, and the temporal extent of manipulation was fixed at 35%, following the practice in Guo (2020), which was in turn based on the Mandarin production data in her study. Note that, as pointed out by one reviewer, the F0 manipulation resulted in initial F0 trajectories that differed not only in onset F0 but also in F0 contour (see Figure 5B). The F0 cue here therefore involved both F0 height and direction.

The creation of filler items roughly followed the same first two steps in creating the target items (e.g., [i˥] and [i˥˩] were created from a natural production of *yi1*), except that the filler [mi˥˩] was modified from a natural Tone 4 syllable, *mi4*, rather than being constructed from the Tone 1 syllable *mi1*. However, the tonal contour of this filler item was similarly styled to that of target Tone 4 items, to prevent this filler from standing out from the other stimuli. The rationale behind was to add acoustic variability to stimuli and therefore to encourage the participant to abstract away from low-level acoustic signals. Note, however, that this decision is not critical with regard to data analysis, as only data from target stimuli were included.

4.2.2. English stimuli

The target stimuli for the English version of the perception experiment were identical to those for the Mandarin version. The filler stimuli, on the other hand, were changed to [mi˥], [mi˥˩] (similar to English *me*), [ni˥], and [ni˥˩] (similar to English *knee*). The reason why [i˥] and [i˥˩] were not used was to avoid the use of letter *E* as one of the response options; it was preferable that all four response options were lexical items.

4.3. Procedure

In presenting the experimental procedure, I first go through the configuration and layout of response options in each trial, and then described the task involved. At a high level, the task was a forced-choice identification task, where the participant clicked on one word out of a choice of four.

4.3.1. Mandarin trial configuration

Experimental trials consisted of two trial types: targets and fillers, depending on whether the audio stimulus being played were from the target or filler set. Both trial types had as response options four Mandarin monosyllabic words. For the targets, the four response words were *pi1* 披, *pi4* 屁, *bi1* 逼, and *bi4* 闭, which differed from one another in stop voicing and lexical tone. Note that these words were also included in the production stimuli. The four options were placed at the four corners of a 600 px × 600 px square, with each option having a response area of a 50 px × 50 px square, as illustrated in Supplementary Figure 2. Furthermore, the relative positions of the four options were constrained in such a way that two words distinguished only in the voicing of onset (e.g., *pi1* vs. *bi1*) were always next to each other, so there were only 16 (4 sides × 4 possible positionings/side) possible trial option configurations. The 16 trial configurations were counterbalanced across participants at the time of testing (i.e., the counterbalance was not taken into account when participants' data was selected for analyses), and the same configuration was used throughout the course of experiment. The decision to maintain the same configuration was to prevent the participant from doing visual search, which might introduce additional cognitive load.

For the fillers, the four options were *yi1* 衣, *yi4* 意, *mi1* 咪, and *mi4* 密, which similarly differed in both onset and lexical tone. However, their positioning was not constrained in any manner, as the data collected in filler trials were not analyzed. This resulted in 24 (= 4!) possible configurations, and each participant was randomly assigned a configuration, which remained the same throughout the entire experiment.

4.3.2. English trial configuration

The experimental trials for English similarly consisted of target trials and filler trials. However, unlike the Mandarin version, the two trial types differed from each other only in the audio stimulus being played; that is, the same response layout was used for both trial types. This being the case came from the fact that English lacks lexical tone, so it was impossible to have a response layout parallel to that in the Mandarin version. The trial configuration always had as response options four English words: *pea*, *bee*, *me*, and *knee*. The four words were arranged such that *pea* and *bee* were always only one edge away from each other (and as a consequence *me* and *knee* were likewise always next to each other)—the same constraint that phonological competitors in terms of stop voicing were always adjacent to each other. This resulted in 16 possible option configurations (4 sides × 4 arrangements/side), two of which are shown in Supplementary Figure 3. These 16 configurations were counterbalanced across participants at the time of testing, and the configuration remained unaltered within an experiment session.

4.3.3. Task procedure

The experiment procedure was the same for both the Mandarin and English versions of the experiment. The whole experiment took place online and was programmed in jsPsych (de Leeuw, 2015). Participants were encouraged to use a physical mouse and to wear headphones for the experiment, though they could also do the experiment with a touchpad and/or the built-in loud speakers on their computer. The experiment started with a short hearing test, where the participant had to select the quietest tone out of three tones differing in loudness. This test was challenging to do when *not* wearing headphones. They had to respond correctly in at least five out of six trials to pass the test.

The basic procedure followed that of Experiment 1 from Dale et al. (2007). During each trial, the four options were first presented for 500 ms to remind the participant of the word at each corner. Next, a black dot, the radius of which was 5 px, appeared in the center of the screen, which the participant had to click for the audio stimulus to be immediately presented. The function of this center dot was to ensure that the mouse cursor was reset to (approximately) the center. The participant then had a 3-s period to indicate their response by clicking one of the words.

Participants had to go through three blocks, with each block having the same tokens and differing only in the order in which the tokens were presented. To have a target-to-filler ratio of about 4:1, each block contained one repetition of target stimuli and seven repetitions of filler stimuli, resulting in a total of 126 ($= 98 \times 1 + 7 \times 4$) trials in each block. Three blocks were used to achieve a compromise between having as many trials as possible and limiting the duration of the experiment under 30 min. Between blocks the participant could take a self-timed break.

4.4. Additional participant inclusion criteria

As mentioned in Section 3.1, participants' performances in the perception experiment formed a part of the inclusion criteria. The purpose is to only include participants who actually paid attention during the experiment. This criterion was operationalized by first calculating by-participant "correct" percentage of responses for each language version, separated for target and filler trials. For the target trials in the Mandarin perception experiment, a correct trial was a target trial where the participant selected as the response a word whose tone matched the tonal contour of the audio stimulus. For the filler trials in the Mandarin experiment, a correct trial was a filler trial whose selected response word corresponded exactly to the audio stimulus (e.g., selecting *yi1* for [i˥]). For the target trials in the English version of the experiment, a correct trial was a target trial whose response was either *pea* or *bee*. For the filler trials in the English experiment, a correct trial was defined as a filler trial which had *me* or *knee* as the response, taking into account the fact that the bilabial and alveolar nasal onsets in the filler stimuli were perceptually confusable. For a participant who completed both English and Mandarin perception experiments, four percentage scores were computed—% correct for targets in Mandarin perception, % correct for fillers in Mandarin perception, % correct for targets in English perception, and % correct for fillers in English perception. For each participant, an average correct percentage across the four language/trial type combinations was computed. Participants were then ranked based on the average correct percentage in a descending order, and the data from the top 25 participants was included in the analyses. A *post-hoc* analysis shows that these included participants had an average correct percentage of at least 90%.

4.5. Omitted data

For both Mandarin and English versions of the perception experiment, only the response data from the target trials were considered. Additionally, only the "correct" target trials, as defined in Section 4.4 above, were included in the analyses. Altogether, 216 (129 Tone 1 tokens and 87 Tone 4 tokens) out of

7,350 target trials were removed from the Mandarin experiment, and 59 (29 Tone 1 tokens and 21 Tone 4 tokens) out of 7,350 target trials were removed from the English experiment.

4.6. Statistical analyses

A variant of logistic regression was used to derive the perceptual weight for post-stop F0. In all the models, participants' responses were modeled as a function of VOT, post-stop F0, and tonal categories. The coefficient of the post-stop F0 variable was then used as its perceptual weight. Similar to the production models, all models were fitted with Bayesian mixed-effects models using `CmdStanR` (Gabry and Češnovar, 2021).

4.6.1. Variables

Before being fed into the analyses, the two continuous predictor variables—**VOT** and **post-stop F0**—were z-transformed with respect to the original sequence (e.g., the VOT value of 0 was consistently mapped to $[0 - \text{mean}(0, 13, 27, 40, 53, 67, 80)]/\text{sd}(0, 13, 27, 40, 53, 67, \text{and } 80) = -1.39$, regardless of listener). The variable **tone** was sum-coded with TONE 1 and TONE 4 being coded with 1 and −1, respectively. The default level for the response was always unaspirated (i.e., the unaspirated response was coded with 0, and the aspirated response was coded with 1), so a positive coefficient for a given predictor variable means that higher values of this dimension elicit more voiceless responses in listeners than lower values.

4.6.2. Model structure

Listeners' responses were assumed to be generated by a mixture of two different sources: one source was the logistic function of terms formed with the predictors, and the other was sheer randomness or guessing due to the listener not paying attention or accidentally making a mistake, that is, the response came from one of the four options being selected by chance (Kruschke, 2015). Formally, each response had a chance, γ, of being generated by the guessing process, and, with probability $1 - \gamma$, the response came from the logistic function of the predictor:

$$\text{aspirated response} \sim \text{bernoulli} \left(\gamma \cdot \frac{1}{4} + (1-\gamma) \cdot \text{logistic} \left(\beta_0 + \sum_i \beta_i x_i \right) \right).$$

Model fitting thus involved estimating the guessing probability γ along with the logistic parameters, β_i, which were taken to represent the weight given to each dimension in categorization. Bayesian hierarchical models were employed to derive a posterior probability distribution for each parameter. The full model consisted of two submodels with the same parameterization and predictors: one submodel predicted listeners' responses in the Mandarin mode while the other submodel predicted listeners' responses in the English mode, and the two submodels were tied together by correlating all logistic parameters with one another in a multinormal distribution. A guessing probability was estimated for each listener in each language mode independently. Logistic parameters were parameterized such that each was decomposed into a fixed-effect part, corresponding to the weight at the population level, and a random-effect part, representing the adjustment for each listener.

Each model used 4,000 samples across four Markov chains and was fit with a regularizing prior of Normal($\mu = 0$, $\sigma = 10$) for the fixed-effect estimates. An Exponential($r = 1$) distribution was used as the prior for listener-specific adjustments. Correlations among listener-specific adjustments used the LKJ prior with $\xi = 1$. The guessing probability for each listener in each language had a uniform prior between 0 and 1. All models showed no divergent transitions, and sampling chains were well-mixed (i.e., all $\hat{R} < 1.01$). The detailed mathematical specifications for the final model can be found in the Supplementary material.

4.6.3. Candidate models

Similar to the statistical models for production data, candidate models for perceptual performance reflected both prior knowledge and a compromise between complexity and predictive accuracy. Given that VOT is the primary cue for the stop voicing contrast in Mandarin and English, all the models in the comparison had VOT automatically included, with the simplest model containing VOT as the sole predictor. Built off this simplest models were candidates with increasing complexity introduced by terms involving post-stop F0 and tone. The full list of models considered is listed in Table 4.

4.7. Results: Perceptual weights of post-stop F0

The response patterns across different VOTs, post-stop F0s, tones, and experiment versions are shown in Figure 6. The ELPD-LOO mean and standard error for each candidate model are listed in Table 4, and the model comparison results among the candidate models are detailed in Supplementary Table 5.

Model comparison indicated the importance of post-stop F0 and tone in predicting listeners' categorization performances (M1 vs. M2 and M3 vs. M4 for post-stop F0; M1 vs. M3 and M2 vs. M4 for tone). However, including interaction terms between any pairs of the cues did not lead to substantial increase in

TABLE 4 Candidate perceptual models considered in model comparison, with their ELPD-LOO means and standard errors.

Model	ELPD-LOO mean	ELPD-LOO standard error	Predictors
M1	−1419.5	52.3	VOT
M2	−1366.2	51.3	VOT + F0
M3	−1395.4	52.0	VOT + tone
M4 (final)	−1340.5	51.4	VOT + F0 + tone
M5	−1334.6	51.5	VOT + F0 + tone + F0 × VOT
M6	−1325.4	51.7	VOT + F0 + tone + F0 × tone
M7	−1326.0	51.8	VOT + F0 + tone + F0 × VOT + F0 × tone
M8	−1327.9	52.1	VOT + F0 + tone + F0 × VOT + F0 × tone + VOT × tone

predictive accuracy. For this reason, M4 was selected as the final model, and subsequent discussion was made on the basis of M4.

4.7.1. Population results

The marginal posterior distributions for population-level effects from M4 are summarized in Table 5. All predictors, including the intercepts, had an effect on categorization. The cue of most interest here is post-stop F0, but for completeness, the results for other dimensions are also briefly discussed. On the basis of the fact that the 89% CrIs for post-stop F0 did not contain 0 in both Mandarin and English (Mandarin: 89% CrI = [0.30, 0.75]; English: 89% CrI = [0.64, 1.14]), post-stop F0 was judged to be a cue for stop voicing in both languages. However, the weight assigned to this cue was language-dependent, as evidenced by the 89% CrI of difference in post-stop F0 weights occupying only negative values (89% CrI = [−0.67, −0.04]). In particular, listeners relied on post-stop F0 more when the stimuli were presented as English words than when the exactly same stimuli were perceived as Mandarin words. The magnitude of the intercept was indicative of the location of category boundary: a positive intercept meant there were more aspirated responses in general, which translated to an early boundary within the range of values considered. This can be clearly seen in Figure 6, where the category boundary in terms of VOT (i.e., the VOT value where the proportion of aspirated responses is 0.5) occurs before the midpoint of the VOT continuum. Also, the intercept seemed stable across participants' Mandarin and English categorization performances. VOT, as expected, was the strongest cue for the voicing decision, and its weight was comparable across languages. Finally, Tone 1 stimuli seemed to trigger more aspirated responses to a similar degree in both languages.

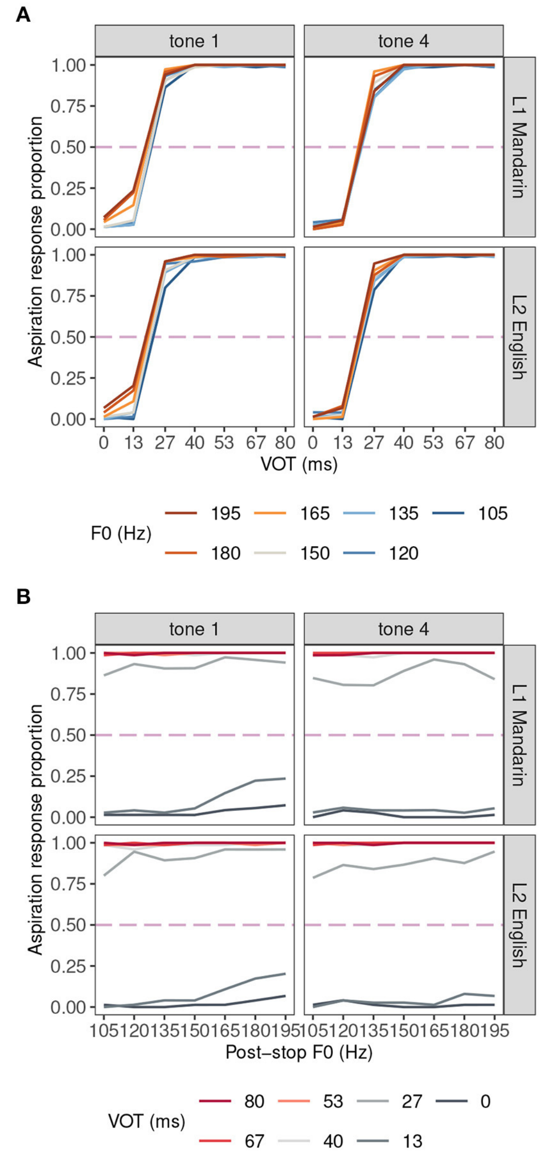

FIGURE 6
Line charts of Mandarin-English bilinguals' aggregated categorization of word-initial stops in each language, shown as a function of VOT, post-stop F0, and tone. **(A)** With VOT on the *x*-axis. **(B)** With post-stop F0 on the *x*-axis to highlight its effect on categorization.

4.7.2. Individual results

The guessing probability estimated for each listener in each language is plotted in the Supplementary Figure 4. Overall, the guessing probabilities were very low, with 24 out of 25 listeners having a mean guessing probability below 5% in either language and only one listener (i.e., participant 12) having a value of around 10% for the English task.

Individual listeners' weights for various cues, which are equal to the coefficient estimates of the corresponding acoustic dimensions, and the weight differences in these cues across

TABLE 5 Marginal posterior summary for key population-level parameters from M4.

Parameter	Mean	SD	89% CrI	p(dir.)
$\text{intercept}_{\text{Man}}$	9.14	0.68	[8.11, 10.28]	$p(\beta > 0) = 1.00$
VOT_{Man}	13.81	1.07	[12.19, 15.63]	$p(\beta > 0) = 1.00$
F0_{Man}	0.53	0.14	[0.30, 0.75]	$p(\beta > 0) = 1.00$
tone_{Man}	0.54	0.12	[0.35, 0.73]	$p(\beta > 0) = 1.00$
$\text{intercept}_{\text{Eng}}$	9.88	0.73	[8.81, 11.07]	$p(\beta > 0) = 1.00$
VOT_{Eng}	15.08	1.11	[13.42, 16.90]	$p(\beta > 0) = 1.00$
F0_{Eng}	0.89	0.16	[0.64, 1.14]	$p(\beta > 0) = 1.00$
tone_{Eng}	0.42	0.12	[0.22, 0.61]	$p(\beta > 0) = 1.00$
$\text{intercept}_{\text{Man}} - \text{intercept}_{\text{Eng}}$	−0.74	0.92	[−2.16, 0.74]	$p(\beta < 0) = 0.78$
$\text{VOT}_{\text{Man}} - \text{VOT}_{\text{Eng}}$	−1.28	1.41	[−3.43, 1.08]	$p(\beta < 0) = 0.81$
$\text{F0}_{\text{Man}} - \text{F0}_{\text{Eng}}$	−0.36	0.20	[−0.07, −0.04]	$p(\beta < 0) = 0.97$
$\text{tone}_{\text{Man}} - \text{tone}_{\text{Eng}}$	0.12	0.17	[−0.15, 0.39]	$p(\beta > 0) = 0.75$
$\rho_{\text{intercept}_{\text{Man}},\text{intercept}_{\text{Eng}}}$	0.41	0.21	[0.04, 0.74]	$p(\beta > 0) = 0.96$
$\rho_{\text{VOT}_{\text{Man}},\text{VOT}_{\text{Eng}}}$	0.52	0.19	[0.20, 0.79]	$p(\beta > 0) = 0.99$
$\rho_{\text{F0}_{\text{Man}},\text{F0}_{\text{Eng}}}$	0.34	0.28	[−0.15, 0.73]	$p(\beta > 0) = 0.88$
$\rho_{\text{tone}_{\text{Man}},\text{tone}_{\text{Eng}}}$	0.10	0.33	[−0.44, 0.62]	$p(\beta > 0) = 0.62$
$\rho_{\text{VOT}_{\text{Man}},\text{F0}_{\text{Man}}}$	−0.33	0.24	[−0.70, 0.08]	$p(\beta < 0) = 0.90$
$\rho_{\text{VOT}_{\text{Eng}},\text{F0}_{\text{Eng}}}$	−0.06	0.27	[−0.50, 0.38]	$p(\beta < 0) = 0.59$
$\rho_{\text{tone}_{\text{Man}},\text{F0}_{\text{Man}}}$	0.20	0.31	[−0.32, 0.66]	$p(\beta > 0) = 0.75$
$\rho_{\text{tone}_{\text{Eng}},\text{F0}_{\text{Eng}}}$	0.11	0.30	[−0.40, 0.59]	$p(\beta > 0) = 0.65$
$\rho_{\text{tone}_{\text{Man}}-\text{tone}_{\text{Eng}},\text{F0}_{\text{Man}}-\text{F0}_{\text{Eng}}}$	−0.11	0.34	[−0.67, 0.42]	$p(\beta < 0) = 0.63$

languages are visualized in Figure 7. Again, the results regarding the cue weight for post-stop F0 are discussed first, as it is the dimension of interest here; the results for other cues are also summarized in passing for completeness.

As shown in the [post-stop F0] panel of Figure 7A, though the 89% CrI for the post-stop F0 weight did cross 0 for some listeners, all listeners had a positive mean weight for the post-stop F0 cue for both languages, signifying that, generally speaking, the chance the aspirated response was selected went up with an increasing post-stop F0. Comparing the weights of this cue across languages (Figure 7B), all but one listener (i.e., participant 25) had a higher mean weight in English than in Mandarin; however, because of the relatively large uncertainty surrounding the estimated weight values, the 89% CrI for the *difference* between the weights still contained 0 for all participants. In spite of this "non-significant" result, the trend seemed robust and echoed the population-level pattern in terms of the direction of the effect. Another way to understand the cue is to examine whether the cue use is consistent across languages at the individual level by correlating the weights from the two language contexts. In fact, the correlation information can be directly read off from the fitted model and is summarized in the last few row in Table 5 and visualized in Figure 8. As can be seen in Figure 8C, there was a weak positive correlation of this cue across languages ($\bar{\rho} = 0.34$, 89% CrI $= [-0.15, 0.73]$, $p(\rho > 0) = 0.88$), though the 89% CrI for this correlation also spilled to the negative side, probably due to the small number of participants, which was not effective in constraining the uncertainty when the correlation was weak.

For the intercepts, which were connected with the location of category boundary, even though individual listeners varied with respect to the boundary location, the location was relatively stable within a listener, as evidenced from Figure 8A and from the positive 89% CrI of the correlation coefficient ($\bar{\rho} = 0.41$, 89% CrI $= [0.04, 0.74]$, $p(\rho > 0) = 0.96$). The same story could be stated for the VOT cue: individuals varied in a structured way, with the cue use being stable within the same individual across contexts ($\bar{\rho} = 0.52$, 89% CrI $= [0.20, 0.79]$, $p(\rho > 0) = 0.99$). As for tone, it seemed that, for most listeners (19 out of 25), the effect of Tone 1 stimuli eliciting more voiceless responses was stronger in Mandarin than in English, though the difference was not particularly big.

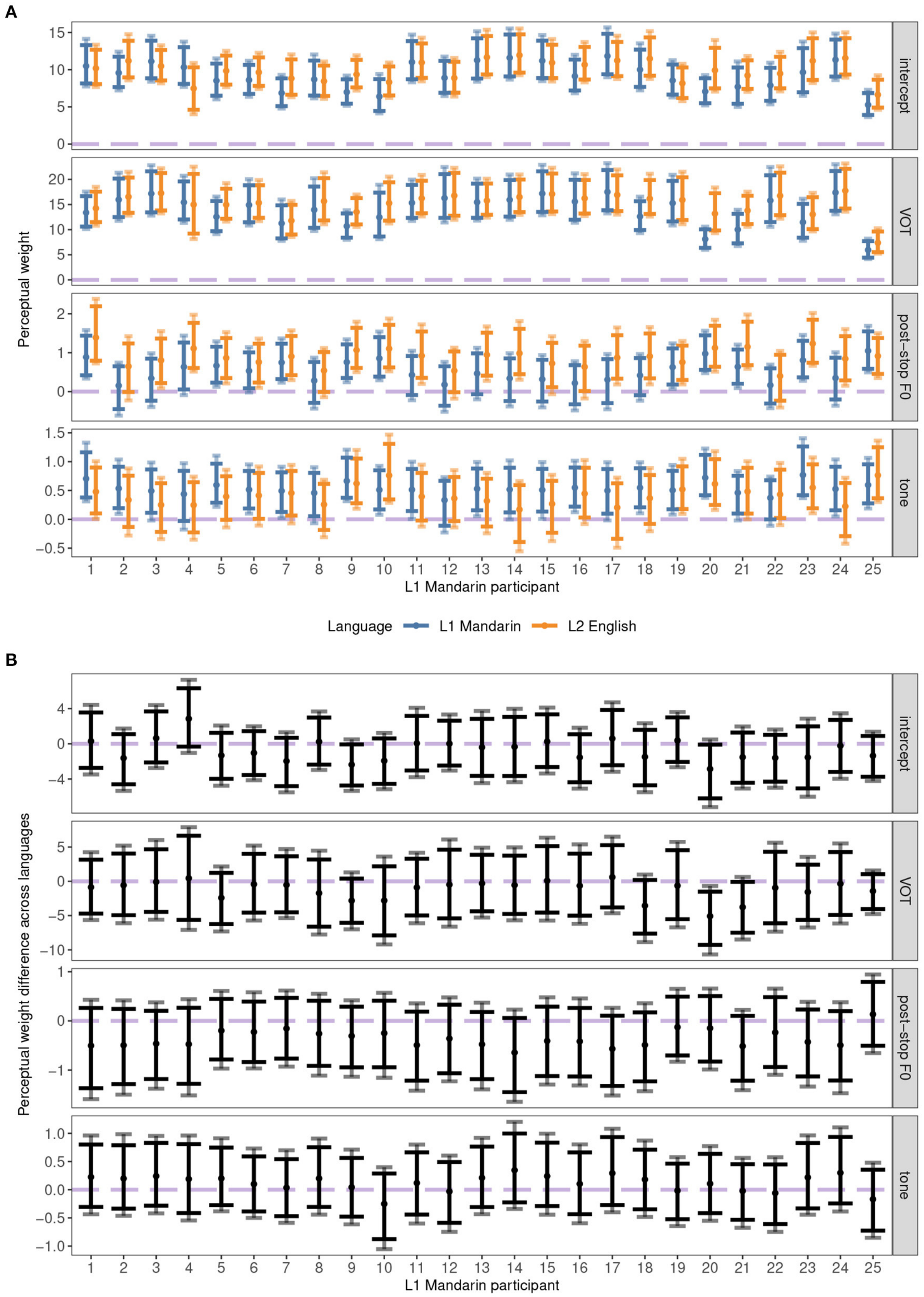

FIGURE 7
Individuals' estimated weights from the perceptual model. **(A)** Distributions of individual weights along various dimensions for Mandarin and English. **(B)** Differences in cue weights along the same dimension across languages. In both figures, posterior means are represented by the dots. The 89% CrIs are marked by the inner error bars, while the 95% CrIs are marked by the outer error bars.

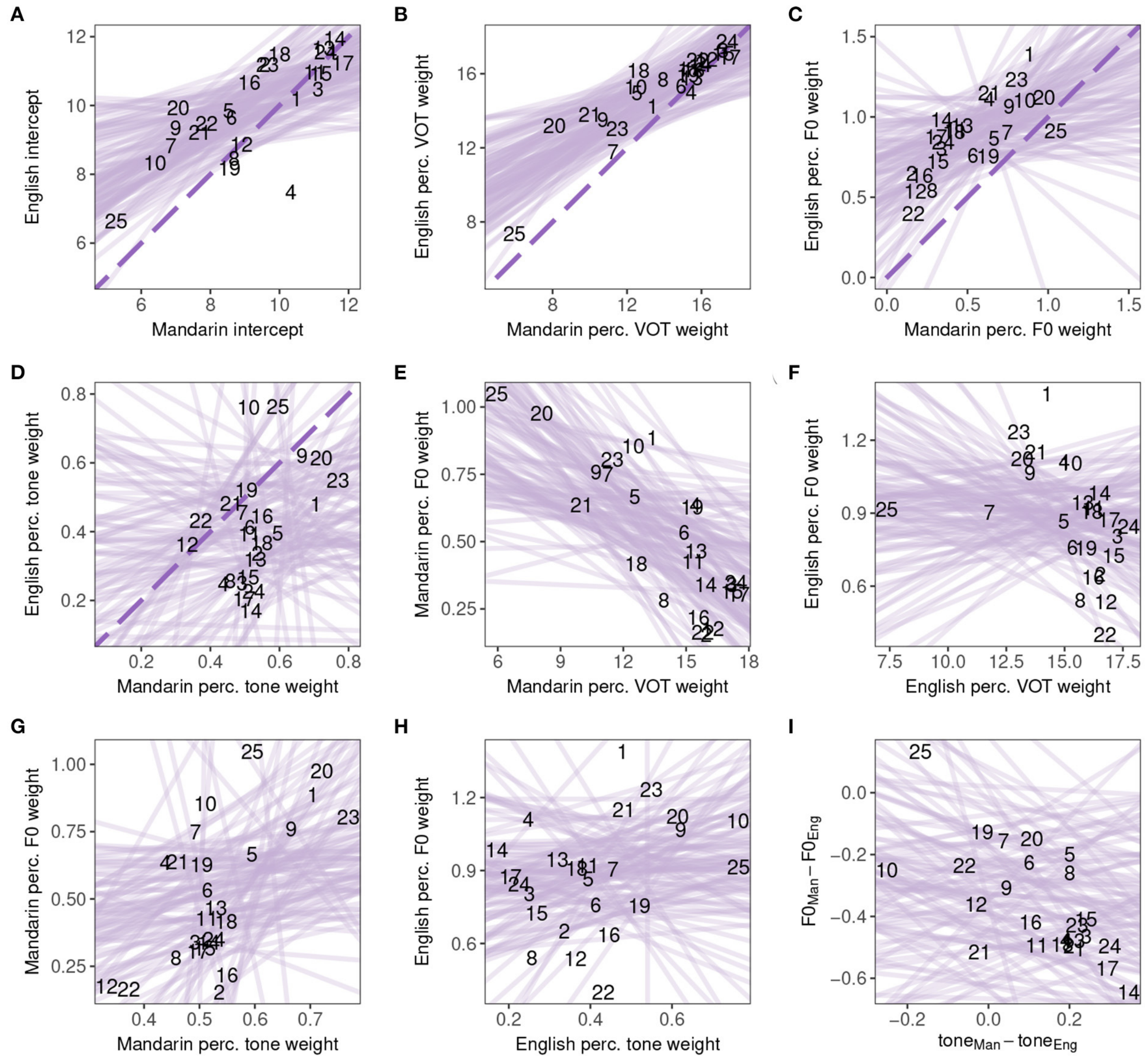

FIGURE 8

Scatter plots showing relationships (or lack thereof) between various cues. **(A)** Intercepts, which are related to category boundaries, across languages. **(B)** VOT weights across languages. **(C)** Post-stop F0 weights across languages. **(D)** Tone weights across languages. **(E)** F0 vs. VOT in Mandarin. **(F)** F0 vs. VOT in English. **(G)** F0 vs. tone in Mandarin. **(H)** F0 vs. tone in English. **(I)** Differences in post-stop F0 weights vs. differences in tone weights. Solid lines represent 100 regression lines fitted with 100 posterior draws, to show the direction and uncertainty of the correlation. The dashed line in **(A–D)** is $y = x$, where the intercept or VOT / post-stop F0 / tone weight for Mandarin equals that for English.

4.8. Comparing individual post-stop F0 weights across production and perception

Given that population-level correspondences between production and perception alone cannot be taken as evidence for a causal link—if there is a (direct or indirect) causal link between the modalities, it should surface on an individual level (Schertz et al., 2020). It is therefore expected that the weight of a given acoustic dimension on a speaker's production would predict the weight assigned to that dimension in the same speaker's perception. To test this hypothesis empirically, two models, separated for each language but otherwise sharing the same structure, were fit using both production and perception data. Each model had two submodels: one estimated individual production weights based on Cohen's *d*, and the other estimated individual perceptual weights based on the beta-coefficient for F0 in the logistic regression model. The two submodels were tied together by a common covariance matrix used to model individual-level variances. The

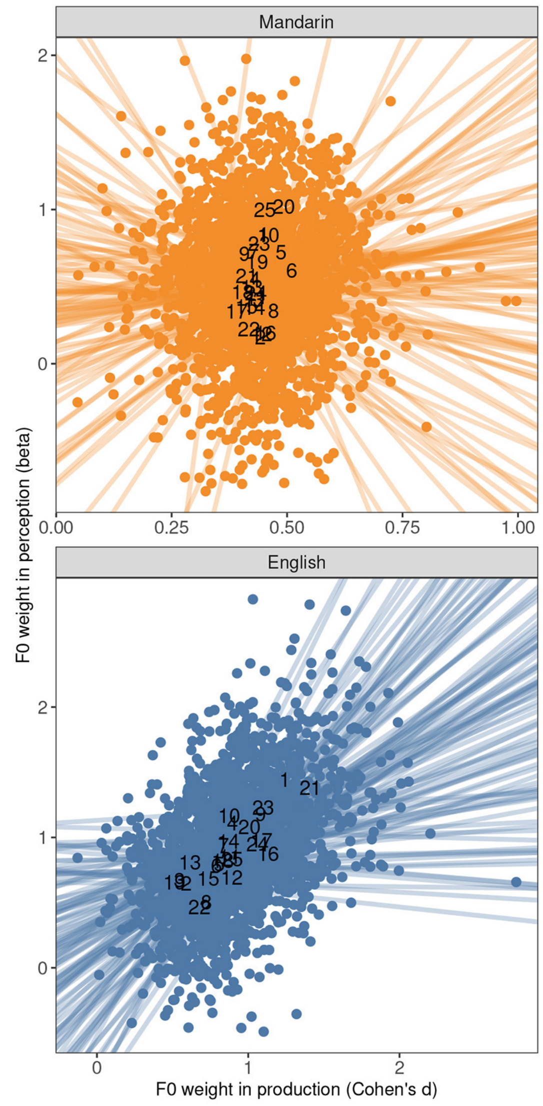

FIGURE 9
Post-stop F0 weights across perception and production in Mandarin and English. The numbers represents the posterior means for individual participants. Each dot is a posterior draw for a participant. The solid lines represent 100 regression lines fitted with 100 posterior draws, to show the direction and uncertainty of the correlation.

mathematical specification for the models can be found in the Supplementary material. Figure 9 shows individual perceptual weights plotted against the corresponding production weights. The results of correlation analyses were dependent on the language, with little evidence of correlation across modalities for Mandarin ($\bar{\rho} = 0.49$, 89% CrI = $[-2.93, 4.18]$, $p(\rho > 0) = 0.61$) but weak evidence for a positive correlation for English ($\bar{\rho} = 0.72$, 89% CrI = $[-0.15, 1.51]$, $p(\rho > 0) = 0.93$).

5. Discussion

5.1. Summary of results

The current study explores the ambiguity of F0 in Mandarin through L1 Mandarin-L2 English bilinguals' production and perception of the stop voicing contrast in their L1 and L2. The results from the conducted experiments are summarized in Table 6, which ties them back to the hypotheses and predicted results listed in Table 1, and discussed below. At the population level, these results largely echoed a recent work by Guo (2020).

In both their Mandarin and English productions, the post-stop F0 following an aspirated stop tended to be higher than that following an unaspirated stop, and unaspirated stops in turn induced a higher F0 than sonorants. In addition, the extent to which post-stop F0 was differentiated between the aspirated and unaspirated categories hinged on the language and lexical tone: comparing English with Mandarin (which was represented as an average between Tone 1 and Tone 4 in this study), English supported a bigger post-stop F0 difference; contrasting Tone 1 and Tone 4 in Mandarin, Tone 4, which was realized with a higher F0 register phonetically, also sustained a slightly greater post-stop F0 distinction. The production weights for post-stop F0 across languages was also reflective of the finding above: post-stop F0 assumed a larger weight in English than in Mandarin, both at the population level and for most individuals (19 out of 25 speakers). These findings therefore support the view that post-stop F0 perturbation is not necessarily intrinsic to the articulatory system.

In perception, post-stop F0 was also used as a cue for stop voicing by the same L1 Mandarin-L2 English participants when put in either a Mandarin or an English context. However, the language context modulated the weight such that post-stop F0 carried more weight when the stimuli were presented as English words than when the same stimuli were presented as Mandarin words. This language-conditioned change in cue weighting was statistically well-supported at the population level, but, at the individual level, because of fewer data points (i.e., the same stimuli were only repeated three times for each participant), the model was less confident. Nonetheless, almost all individuals (24 out of 25) followed the population trend as far as posterior means were concerned. Overall, the patterns revealed in the perception experiment are supportive of the claim that L2 learners can adjust the use of a cue in different language contexts.

Compared across production and perception, on a population level, a higher production weight for post-stop F0 mapped to a higher perceptual weight for the same cue. This is reflected in the bilinguals' relying more on post-stop F0 to contrast stop voicing in English than in Mandarin across modalities. On an individual level, on the other hand, an individual's production weight did not reliably predict the same individual's perceptual weight,

TABLE 6 Predicted and actual production and perception results under difference hypotheses.

Production

Hypotheses	Predicted production results	Match actual results?
Post-stop F0 purely due to physiological / aerodynamic reasons (e.g., Ladefoged, 1967; Ohala and Ohala, 1972; Kohler, 1984) or total transfer of post-stop F0 cue use in Mandarin to English, as prediced by the SLM and PAM-L2	Post-stop F0 difference the same in Mandarin and English tokens	No
Post-stop F0 partially subject to active controlling (Kingston and Diehl, 1994)	The extent of post-stop F0 difference might depend on the language (i.e., larger in English than in Mandarin)	Yes. Post-stop F0 difference between aspirated and unaspirated stops was bigger in English than in Mandarin at the population level and for 19 (out of 25) speakers.

Perception

Hypotheses	Predicted perception results	Match actual results?
Transfer of the Mandarin cue-weighting strategy to English, as predicted by the SLM and PAM-L2	Post-stop F0 weights the same across Mandarin and English	No.
Flexibility in cue use: attributing variation in post-stop F0 partially to lexical tone and partially to stop voicing in Mandarin, but only to stop voicing in English	Post-stop F0 weights depend on the language context (i.e., a higher weight in English than in Mandarin)	Yes. Post-stop F0 carried more weight in English than in Mandarin at the population level. The model was less confident at the individual level, though the trend was the same as the population result for 24 out of 25 listeners.

at least for post-stop F0 with the adopted metrics. This mismatch therefore suggests at least some independence of the two modalities.

5.2. Flexibility of cue-weighting across L1 and L2

The findings from the experiments show that bilinguals, even non-early/ non-simultaneous/non-child bilinguals, are able to dynamically adjust their cue-weighting strategies in facing different language contexts in production as well as perception. Prior demonstrations on bilinguals' ability to fine-tune the use of various acoustic dimensions concerned mainly simultaneous or early bilinguals (e.g., Antoniou et al., 2010, 2012; Gonzales and Lotto, 2013; Gonzales et al., 2019). However, as reviewed in Section 2.5, more recent works have suggested that late L2 learners are also capable of such a deed. The results from this study are in line with thse recent works in that Mandarin-English bilinguals shift the post-stop F0 weight in response to the current language mode. Crucially, however, this study also demonstrates bilinguals' capability to modulate the use of a secondary cue, as opposite to just the primary cue as in previous works.

5.3. Role of tone in post-stop F0

The fact that, in production, greater post-stop F0 difference was found in Tone 4, which was realized with a higher initial pitch than Tone 1, and that, in perception, Tone 1 syllables induced more aspirated responses, points to a potential role of tone identity in conditioning post-stop F0. In fact, previous works have documented such cases in production at least. For example, as mentioned in Section 2.2.2, Guo (2020) reports that F0 following an aspirated stop is higher only in Tone 1 and Tone 4 syllables (both of which begin with a high pitch register) while F0 following an *un*aspirated is higher in Tone 2 and Tone 3 syllables (both having a low initial register). Kirby (2018) investigates the post-stop F0 effects in two other tonal languages—Thai and Vietnamese—and finds that the greatest post-stop F0 effects for Thai are present in the high-falling tone environment, though the results from Vietnamese are less clear-cut. Even in non-tonal languages, post-stop F0 difference is most prominent in high-pitch, focused conditions (Hanson, 2009; Kirby and Ladd, 2016). The enlargement of post-stop F0 difference in high-pitch contexts across tonal and non-tonal languages suggests that a general, language-independent explanation in terms of F0 control might be responsible, and more research is needed to elucidate this hypothesis.

With respect to perception, a careful inspection of Figure 6 reveals that increased aspirated responses in Tone 1 tokens resulted largely from higher post-stop F0 values in Tone 1 provoking more aspirated responses when VOT was ambiguous (i.e., when VOT was around 13 ms). A possible explanation for why Tone 1, as compared with Tone 4, led to such an effect is that it is not just the initial value of F0 that matters; the listener also tracks changes in F0 slope throughout the syllable, and such changes also contribute to the perception of F0. In the context of the current perception experiment, all Tone 1 tokens end with a tailing flat F0 contour, which might enhance the percept of the initial drop in F0, whereas the falling F0 contour in Tone 4 tokens might perceptually offset the initial drop in F0, resulting in the change in F0 being less noticeable. Another explanation is that since Tone 4 syllables tend to have a higher initial F0 in production than Tone 1 syllables, Mandarin listeners might require an acoustically higher initial F0 value in Tone 4 tokens to judge a token as starting with a high F0. Of course these speculations await more investigation.

Related to changes in F0 slope is the question, as pointed out by a reviewer, of whether the observed effect of post-stop F0 is induced by vowel-onset F0 height or by the F0 contour within the range of manipulation (i.e., from vowel onset to the 35% mark of the vowel). As can be seen in Figure 5B, the manipulation of F0 in this study conflates vowel-onset F0 height and F0 contour. For instance, for F0 manipulation in both Tone 1 and Tone 4 tokens, a higher vowel-onset F0 is associated with a more positive F0 contour. As reviewed in Section 2.2.3, both F0 height and F0 contour contribute to perception of various pitch events. It is therefore possible that both vowel-onset F0 and F0 contour drive the perception of an aspirated stop for a high post-stop F0. One possible future direction is to tease apart the respective influence of the two manipulations.

5.4. A trade-off between post-stop F0 and tone?

The fact that the post-stop F0 weight is diminished in the Mandarin context across both production and perception raises the question of whether the lost weight in post-stop F0 is transferred to other dimensions, with the most obvious candidate being tonal category. In what follows, I discuss the case with production first before moving on to perception.

The question about the existence of a trade-off between post-stop F0 and tone is tied to the debate of whether tone attenuates the degree of post-stop F0 difference. As mentioned in Section 2.2.2, whereas there are some studies that point to a positive direction (e.g., Gandour, 1974; Hombert, 1978), large magnitudes of post-stop F0 difference have also been observed in tonal languages (e.g., Phuong, 1981; Shimizu, 1994; Xu and Xu, 2003; Francis et al., 2006). In the current study, the Mandarin-English bilinguals' respective language productions do conform to the former pattern at the population level. However, not every speaker matches the population-level trend, with some speakers producing the post-stop F0 effect to a similar degree in both languages. The results presented here thus agree with Kirby's (2018) observation that attenuation of post-stop F0 effect in tone languages depends on speaker-specific implementation of laryngeal maneuvers to distinguish voicing and tone.

With respect to perception, if, as described in Section 2.6, it is indeed the case that, in interpreting the audio stimuli as Mandarin words, Mandarin-English bilinguals attribute the variation in post-stop F0 partially to the lexical tones in the language, and that in treating the stimuli as English words, they ascribe the variation to stop voicing, then it is expected the loss in post-stop F0 weight from Mandarin to English to be accompanied by an increase in tone weight. Looking at Table 5, which shows the results at the population level, it seems the loss in post-stop F0 is indeed accompanied by an increase in tone weight, though the model is not as confident in the increase in tone weight as in the decrease in post-stop F0 weight. At the individual level, the panels for post-stop F0 and tone in Figure 8 also appear to suggest that for many participants, a drop in post-stop F0 weight is compensated by a rise in tone weight, and that those who have a bigger drop tend to have a sharper rise as well (notice the apparent negative correlation in Figure 8I when the changes along these two dimensions are plotted against each other), at least as far as the posterior mean is concerned. However, the correlation coefficient estimated from posterior samples does not back up this hypothesis (as shown by the lines going into different directions in Figure 8I). Therefore, it is still inconclusive as to whether there is a trade-off relation between post-stop F0 and tone in perception.

5.5. Production-perception interface

As shown in Section 4.8, even though the use of post-stop F0 is mirrored across production and perception at the population level, the link between the two modalities at the individual level seems to be less robust. While there is weak evidence for a positive correlation between the production and perceptual weights for English, such a correlation is missing for Mandarin. This observation raises the question as to the cause of this asymmetry. One possible answer might be that post-stop F0 is an unreliable cue for phonological voicing in Mandarin. For instance, looking at Figure 7A, almost all individuals use post-stop F0 to a lesser degree as a cue for voicing in Mandarin than in English; for many, the model indicates only very weak evidence for the use of post-stop F0 as a cue. This lack of robustness in the perpetual use of post-stop F0 in Mandarin can be understood in the context of production results from previous studies. Recall from the review in Section 2.2.2 that conflicting findings have been reported regarding the direction of post-stop F0 perturbation in Mandarin. These findings might be suggestive of an inconsistent patterning between post-stop F0 and voicing in Mandarin, and/or large individual variation in this patterning due to dialects, L2 influences, etc. The net result is that Mandarin listeners learn to downweight the post-stop F0 cue as it is only marginally useful in signaling voicing. In other words, the lack of link between production and perception in Mandarin at the individual level comes about because listeners downweight the use of post-stop F0 in the face of potentially conflicting cue use in ambient speech, even though they might produce post-stop F0 in a consistent manner. This explanation is therefore in line with the proposal put forth by Beddor (2015) and Samuel and Larraza (2015) that individuals command a more flexible perceptual proficiency than their production repertoire in order to accommodate potentially large between-speaker variation.

It is worth pointing out that, among the studies that sought to establish individual-level correlation in cue use, a lack of relationship seems to be the norm. For instance, null results have been reported for VOT and F0 in English (Shultz et al., 2012), VOT, F0, closure duration, and F1 onset for English and Spanish (Schertz, 2014), or VOT, F0, and closure duration in L1 Korean and L2 English (Schertz et al., 2015), among other studies that used fairly standard paradigms similar to the one employed in this study. These studies also have in common estimating correlations from individuals' empirical mean cue weights. Such approaches disregard uncertainty surrounding the estimates, so the apparent correlation (or lack of correlation) might not be reliable. To properly account for the uncertainty requires fitting both production and perception data with a single model, and the resulting correlation might not agree with the apparent correlation based on means (M. Sonderegger, personal communication, May 20, 2022). Future research will therefore benefit from directly modeling the uncertainty.

6. Conclusion

The current work examines whether and how L1 Mandarin-L2 English bilinguals use post-stop F0 as a cue for stop voicing across production and perception in Mandarin as well as English contexts. The production results show that F0 is actively used to encode both tonal and voicing distinctions in their Mandarin tokens, and that voicing distinctions are likewise embedded with post-stop F0 in English tokens. In perception, the bilinguals are also able to extract voicing information from post-stop F0 (in the same direction as observed in production) in both languages, even when post-stop F0 is integrated in the overall pitch contour, which they need to monitor in order to identify the lexical tone. Crucially, the reliability of post-stop F0 in signaling the voicing contrast and the extent to which the bilinguals lean on post-stop F0 for voicing perceptually are language-specific, such that production and perceptual weights for post-stop F0 are greater in the English context. However, a positive correlation between production and perceptual weights at the individual level is only observed for English, but not for Mandarin. This lack of correlation in Mandarin is interpreted as reflecting Mandarin listeners' flexible perceptual strategies in response to large individual variability in the direction of post-stop F0 perturbation in Mandarin.

Data availability statement

The datasets and analysis script for this study can be found below: https://osf.io/kw8ph/.

Ethics statement

The studies involving human participants were reviewed and approved by UBC Behavioural Research Ethics Board. The participants provided their written informed consent to participate in this study.

Author contributions

The research and writing were done by RY-HL.

Acknowledgments

My many thanks go to Kathleen Currie Hall, Molly Babel, and Márton Sóskuthy for their support and feedback throughout this project.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.864127/full#supplementary-material

References

Abramson, A. S., and Lisker, L. (1985). "Relative power of cues: F0 shift versus voice timing," in *Phonetic linguistics: Essays in honor of Peter Ladefoged*, ed V. A. Fromkin (New York, NY: Academic Press), 25–33.

Allen, J. S., and Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics ofmonosyllabic words. *J. Acoust. Soc. Am.* 106, 2031–2039. doi: 10.1121/1.427949

Amengual, M. (2021). The acoustic realization of language-specific phonological categories despite dynamic cross-linguistic influence in bilingual and trilingual speech. *J. Acoust. Soc. Am.* 149, 1271–1284. doi: 10.1121/10.0003559

Antoniou, M., Best, C. T., Tyler, M. D., and Kroos, C. (2010). Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2. *J. Phon.* 38, 640–653. doi: 10.1016/j.wocn.2010.09.005

Antoniou, M., Tyler, M. D., and Best, C. T. (2012). Two ways to listen: do L2-dominant bilinguals perceive stop voicing according to language mode? *J. Phon.* 40, 582–594. doi: 10.1016/j.wocn.2012.05.005

Barnes, J., Veilleux, N., Brugos, A., and Shattuck-Hufnagel, S. (2010). "The effect of global F0 contour shape in the perception of tonal timing contrasts in American English intonation," in *Proceedings of Speech Prosody 2010* (Chicago, IL), 1–4.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Beddor, P. S. (2015). "The relation between language users' perception and production repertoires," in *Proceedings of the 18th International Congress of Phonetic Sciences* (Glasgow: University of Glasgow), 1–9.

Best, C. T., and Tyler, M. D. (2007). "Nonnative and second-language speech perception: commonalities and complementarities," in *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, eds O.-S. Bohn and M. J. Munro (Amsterdam: John Benjamins Publishing Company), 13–34.

Blicher, D. L., Diehl, R. L., and Cohen, L. B. (1990). Effects of syllable duration on the perception of the Mandarin Tone2/Tone3 distinction: evidence of auditory enhancement. *J. Phon.* 18, 37–49. doi: 10.1016/S0095-4470(19)30357-2

Boersma, P., and Weenink, D. (2021). *Praat: Doing Phonetics by Computer [Computer Program], Version 6.1.38*. Retrieved from: https://www.fon.hum.uva.nl/praat/

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: a probabilistic programming language. *J. Stat. Softw.* 76, 1–32. doi: 10.18637/jss.v076.i01

Casillas, J. V., and Simonet, M. (2018). Perceptual categorization and bilingual language modes: assessing the *double phonemic boundary* in early and late bilinguals. *J. Phon.* 71, 51–64. doi: 10.1016/j.wocn.2018.07.002

Chen, Y. (2011). How does phonology guide phonetics in segment-*F0* interaction? *J. Phon.* 39, 612–625. doi: 10.1016/j.wocn.2011.04.001

Clayards, M. (2018). Individual talker and token covariation in the production of multiple cues to stop voicing. *Phonetica* 75, 1–23. doi: 10.1159/000448809

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn.* Mahwah, NJ: Lawrence Erlbaum Associates.

Cole, J., Kim, H., Choi, H., and Hasegawa-Johnson, M. (2007). Prosodic effects on acoustic cues to stop voicing and place of articulation: evidence from Radio News speech. *J. Phon.* 35, 180–209. doi: 10.1016/j.wocn.2006.03.004

Dale, R., Kehoe, C., and Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Mem. Cogn.* 35, 15–28. doi: 10.3758/BF03195938

de Leeuw, J. R. (2015). jsPsych: a javascript library for creating behavioral experiments in a web browser. *Behav. Res. Methods* 47, 1–12. doi: 10.3758/s13428-014-0458-y

Dilley, L. C., and Heffner, C. C. (2013). The role of f0 alignment in distinguishing intonation categories: evidence from American English. *J. Speech Sci.* 3, 3–67. doi: 10.20396/joss.v3i1.15039

Dmitrieva, O., Llanos, F., Shultz, A. A., and Francis, A. L. (2015). Phonological status, not voice onset time, determines the acoustic realization of onset *F0* as a secondary voicing cue in Spanish and English. *J. Phon.* 49:77–95. doi: 10.1016/j.wocn.2014.12.005

Ewan, W. (1976). *Laryngeal behavior in speech* (Ph.D. thesis). University of California, Berkeley, Berkeley, CA.

Flege, J. E. (1995). "Second language speech learning theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research, Chapter 8*, ed W. Strange (Timonium, MD: York Press), 233–277.

Flege, J. E. (2007). "Language contact in bilingualism: Phonetic system interactions," in *Laboratory Phonology 9*, eds J. Cole and J. I. Hualde (Berlin: Mouton de Gruyter), 353–381.

Fogerty, D., and Humes, L. E. (2012). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *J. Acoust. Soc. Am.* 131, 1490–1501. doi: 10.1121/1.3676696

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *J. Phon.* 14, 3–28. doi: 10.1016/S0095-4470(19)30607-2

Francis, A. L., Ciocca, V., Wong, V. K. M., and Chan, J. K. L. (2006). Is fundamental frequency a cue to aspiration in initial stops? *J. Acoust. Soc. Am.* 120, 2884–2895. doi: 10.1121/1.2346131

Francis, A. L., Kaganovich, N., and Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *J. Acoust. Soc. Am.* 124, 1234–1251. doi: 10.1121/1.2945161

Fulop, S. A., and Scott, H. J. M. (2021). Consonant voicing in the Buckeye corpus. *J. Acoust. Soc. Am.* 149, 4190–4197. doi: 10.1121/10.0005199

Gabry, J., and Češnovar, R. (2021). *cmdstanr: R Interface to 'CmdStan'*. Available online at: https://mc-stan.org/cmdstanr; https://discourse.mc-stan.org

Gandour, J. (1974). Consonant types and tone in siamese. *J. Phon.* 2, 337–350. doi: 10.1016/S0095-4470(19)31303-8

Gandour, J. (1983). Tone perception in Far Eastern languages. *J. Phon.* 11, 149–175. doi: 10.1016/S0095-4470(19)30813-7

Gandour, J. T. (1978). "The perception of tone," in *Tone: A Linguistic Survey, Chapter 2*, ed V. A. Fromkin (New York, NY: Academic Press), 41–76.

Gao, J., and Arai, T. (2018). "F0 perturbation in a "pitch-accent" language," in *Proceedings of the Sixth International Symposium on Tonal Aspects of Languages* (Berlin), 56–60.

Gonzales, K., Byers-Heinlein, K., and Lotto, A. J. (2019). How bilinguals perceive speech depends on which language they think they're hearing. *Cognition* 182, 318–330. doi: 10.1016/j.cognition.2018.08.021

Gonzales, K., and Lotto, A. J. (2013). A Bafri, un Pafri: Bilinguals' pseudoword identifications support language-specific phonetic systems. *Psychol. Sci.* 24, 2135–2142. doi: 10.1177/0956797613486485

Guo, Y. (2020). *Production and perception of laryngeal contrasts in Mandarin and English by Mandarin speakers* (Ph.D. thesis). George Mason University, Fairfax, VA.

Han, M. S., and Weitzman, R. S. (1970). Acoustic features of Korean /P, T, K/, /p, t, k/ and /p^h, t^h, k^h/. *Phonetica* 22, 112–128. doi: 10.1159/000259311

Hanson, H. M. (2009). Effects of obstruent consonants on fundamental frequency at vowel onset in English. *J. Acoust. Soc. Am.* 1, 425–441. doi: 10.1121/1.3021306

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 97, 3099–3111. doi: 10.1121/1.411872

Ho, A. T. (1976). The acoustic variation of Mandarin tones. *Phonetica.* 33, 353–367. doi: 10.1159/000259792

Holt, L. L., and Lotto, A. J. (2006). Cue weighting in auditory categorization: implications for first and second language acquisition. *J. Acoust. Soc. Am.* 119, 3059–3071. doi: 10.1121/1.2188377

Hombert, J.-M. (1978). "Consonant types, vowel quality, and tone," in *Tone: A Linguistic Survey, Chapter 3*, ed V.A. Fromkin (New York, NY: Academic Press), 77–111.

Hombert, J.-M., Ohala, J. J., and Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language* 55, 37–58. doi: 10.2307/412518

House, A. S., and Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *J. Acoust. Soc. Am.* 25, 105–113. doi: 10.1121/1.1906982

Howie, J. M. (1976). *Acoustical Studies of Mandarin Vowels and Tones*. Cambridge: Cambridge University Press.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., et al. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87, B47-B57. doi: 10.1016/S0010-0277(02)00198-1

Jessen, M., and Roux, J. C. (2002). Voice quality differences associated with stops and clicks in Xhosa. *J. Phon.* 30, 1–52. doi: 10.1006/jpho.2001.0150

Jun, S.-A. (1996). Influence of microprosody on macroprosody: a case of phrase initial strengthening. *UCLA Working Pap. Phonet.* 92, 97–116.

Kataoka, R. (2011). *Phonetic and cognitive bases of sound change*. (Ph.D. thesis). University of California, Berkeley, Berkeley, CA.

Keating, P., and Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *J. Acoust. Soc. Am.* 132, 1050–1060. doi: 10.1121/1.4730893

Kingston, J., and Diehl, R. L. (1994). Phonetic knowledge. *Language* 70, 419–454. doi: 10.1353/lan.1994.0023

Kingston, J., Diehl, R. L., Kirk, C. J., and Castleman, W. A. (2008). On the internal perceptual structure of distinctive features: the [voice] contrast. *J. Phon.* 36, 28–54. doi: 10.1016/j.wocn.2007.02.001

Kirby, J. P. (2018). Onset pitch perturbations and the cross-linguistic implementation of voicing: evidence from tonal and non-tonal languages. *J. Phon.* 71, 326–354. doi: 10.1016/j.wocn.2018.09.009

Kirby, J. P., and Ladd, D. R. (2016). Effects of obstruent voicing on vowel F0: evidence from "true voicing" languages. *J. Acoust. Soc. Am.* 140, 2400–2411. doi: 10.1121/1.4962445

Kohler, K. J. (1982). F_0 in the production of lenis and fortis plosives. *Phonetica* 39, 199–218. doi: 10.1159/000261663

Kohler, K. J. (1984). Phonetic explanation in phonology: the feature fortis/lenis. *Phonetica* 41, 150–174. doi: 10.1159/000261721

Koster, J., and McElreath, R. (2017). Multinomial analysis of behavior: statistical methods. *Behav. Ecol. Sociobiol.* 71, 1–14. doi: 10.1007/s00265-017-2363-8

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and Stan, 2nd Edn*. London: Academic Press.

Ladefoged, P. (1967). *Three Areas of Experimental Phonetics*. Oxford: Oxford University Press.

Lea, W. A. (1973). "Segmental and suprasegmental influences on fundamental frequency contours," in *Consonant Types and Tone: Southern California Occasional Papers in linguistics no. 1*, ed L. M. Hyman (Los Angeles, CA: The Linguistics Program; University of Southern California), 16–70.

Lee, B., and Sidtis, D. V. L. (2017). The bilingual voice: vocal characteristics when speaking two languages across speech tasks. *Speech Lang. Hear.* 20, 174–185. doi: 10.1080/2050571X.2016.1273572

Lehiste, I., and Peterson, G. E. (1961). Some basic considerations in the analysis of intonation. *J. Acoust. Soc. Am.* 33, 419–425. doi: 10.1121/1.1908681

Lehnert-LeHouillier, H. (2007). "The influence of dynamic F0 on the perception of vowel duration: cross-linguistic evidence," in *Proceedings of the 16th International Congress of Phonetic Sciences* (Saarbrücken), 757–760.

Leung, K. K. W., and Wang, Y. (2020). Production-perception relationship of Mandarin tones as revealed by critical perceptual cues. *J. Acoust. Soc. Am.* 147, EL301-EL306.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *J. Multivar. Anal.* 100, 1989–2001. doi: 10.1016/j.jmva.2009.04.008

Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychol. Monogr. Gen. Appl.* 68, 1–13. doi: 10.1037/h0093673

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi: 10.1016/0010-0277(85)90021-6

Liberman, M. (2014). *Consonant Effects on F0 in Chinese [Blog Post]*. Retrieved from: https://languagelog.ldc.upenn.edu/nll/?p=12902

Lisker, L. (1986). "Voicing" in English: a catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Lang. Speech.* 29, 3–11. doi: 10.1177/002383098602900102

Liu, C., and Rodriguez, A. (2012). Categorical perception of intonation contrast: effects of listeners' language background. *J. Acoust. Soc. Am.* 131, EL427-EL433. doi: 10.1121/1.4710836

Lotto, A. J., Sato, M., and Diehl, R. L. (2004). "Mapping the task for the second language learner: the case of Japanese acquisition of /r/ and /l/," in *From Sound to Sense: 50+ Years of Discoveries in Speech Communication* (Cambridge, MA), C-181-C-186.

Luo, Q. (2018). *Consonantal effects on F0 in tonal languages* (Ph.D. thesis). Michigan State University, East Lansing, MI.

Ma, J. K.-Y., Ciocca, V., and Whitehill, T. L. (2006). Effect of intonation on Cantonese lexical tones. *J. Acoust. Soc. Am.* 120, 3978–3987. doi: 10.1121/1.2363927

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course With Examples in R and Stan*. Boca Raton, FL: CRC Press.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., and Fujimura, O. (1975). An effect of linguistic experience: the discrimination of [r] and [l] by native speakers of Japanese and English. *Percept. Psychophys.* 18, 331–340. doi: 10.3758/BF03211209

Mohr, B. (1971). Intrinsic variations in the speech signal. *Phonetica* 23, 65–93. doi: 10.1159/000259332

Ohala, J. J. (1978). "Production of tone," in *Tone: A Linguistic Survey, Chapter 1*, ed V. A. Fromkin (New York, NY: Academic Press), 5–39.

Ohala, M., and Ohala, J. (1972). The problem of aspiration in Hindi phonetics. *Ann. Bull. Res. Inst. Logopedics Phoniatr.* 6, 39–46.

Ohde, R. N. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *J. Acoust. Soc. Am.* 75, 224–230. doi: 10.1121/1.390399

Phuong, V. T. (1981). *The acoustic and perceptual nature of tone in Vietnamese* (Ph.D. thesis). Australian National University, Canberra.

Ren, X., and Mok, P. (2021). "Consonantal effects of aspiration on onset F0 in Cantonese," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 1–5.

Samuel, A. G., and Larraza, S. (2015). Does listening to non-native speech impair speech perception? *J. Phon.* 81, 51–71. doi: 10.1016/j.jml.2015.01.003

Schertz, J., Carbonell, K., and Lotto, A. J. (2020). Language specificity in phonetic cue weighting: monolingual and bilingual perception of the stop voicing contrast in English and Spanish. *Phonetica* 77, 186–208. doi: 10.1159/000497278

Schertz, J., Cho, T., Lotto, A., and Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *J. Phon.* 52:183–204. doi: 10.1016/j.wocn.2015.07.003

Schertz, J. L. (2014). *The structure and plasticity of phonetic categories across languages and modalities* (Ph.D. thesis). The University of Arizona, Tucson, AZ.

Shimizu, K. (1994). "F0 in phonation types of initial-stops," in *Proceedings of the 5th Australasian International Conference on Speech Science and Technology, Vol. 2* (Perth), 650–655.

Shultz, A. A., Francis, A. L., and Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *J. Acoust. Soc. Am.* 132, EL95-EL101. doi: 10.1121/1.4736711

Sivula, T., Magnusson, M., and Vehtari, A. (2020). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv:2008.10296 [stat.ME]*. doi: 10.48550/arxiv.2008.10296

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 1074–1095. doi: 10.1037/0096-1523.7.5.1074

Toscano, J. C., and McMurray, B. (2010). Cue integration with categories: weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cogn. Sci.* 34, 434–464. doi: 10.1111/j.1551-6709.2009.01077.x

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput*. 27:1413–1432. doi: 10.1007/s11222-016-9696-4

Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (1990). Gradient effects of fundamental frequency on stop consonant voicing judgments. *Phonetica* 47, 36–49. doi: 10.1159/000261851

Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (1993). *F0* gives voicing information even with unambiguous voice onset times. *J. Acoust. Soc. Am*. 4, 2152–2159. doi: 10.1121/1.406678

Whalen, D. H., and Levitt, A. G. (1995). The universality of intrinsic F_0 of vowels. *J. Phon*. 23, 349–366. doi: 10.1016/S0095-4470(95)80165-0

Winn, M. B. (2020). Manipulation of voice onset time in speech stimuli: a tutorial and flexible Praat script. *J. Acoust. Soc. Am*. 147, 852–866. doi: 10.1121/10.0000692

Xu, B. R., and Mok, P. (2012). "Cross-linguistic perception of intonation by Mandarin and Cantonese listeners," in *Proceedings of Speech Prosody 2012* (Shanghai), 99–102.

Xu, C. X., and Xu, Y. (2003). Effects of consonant aspiration on Mandarin tones. *J. Int. Phon. Assoc*. 33, 165–181. doi: 10.1017/S0025100303001270

Yang, J., Zhang, Y., Li, A., and Xu, L. (2017). On the duration of Mandarin tones. *Proc. Interspeech* 2017, 1407–1411. doi: 10.21437/Interspeech.2017-29

Yuan, J., Ryant, N., and Liberman, M. (2014). "Automatic phonetic segmentation in Mandarin Chinese: boundary models, glottal features and tone," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Florence: IEEE), 2539–2543.

Zellou, G. (2017). Individual differences in the production of nasal coarticulation and perceptual compensation. *J. Phon*. 61, 13–29. doi: 10.1016/j.wocn.2016.12.002

TYPE Original Research
PUBLISHED 24 November 2022
DOI 10.3389/fpsyg.2022.867353

The effects of lexical frequency and homophone neighborhood density on incomplete tonal neutralization

Yifei Bi[1]* and Yiya Chen[2,3]

[1]College of Foreign Languages, University of Shanghai for Science and Technology, Shanghai, China, [2]Leiden University Centre for Linguistics, Leiden, Netherlands, [3]Leiden Institute for Brain and Cognition, Leiden, Netherlands

We investigated the effects of lexical frequency and homophone neighborhood density on the acoustic realization of two neutralizing falling tones in Dalian Mandarin Chinese. Monosyllabic morphemes containing the target tones (Tone 1 and Tone 4) were produced by 60 native speakers from two generations (middle-aged vs. young). The duration of tone-bearing syllable rhymes, as well as the F0 curves and velocity profiles of the lexical tones were quantitatively analyzed *via* linear mixed-effects modeling and functional data analysis. Results showed no durational difference between T1 and T4. However, the F0 contours of the two falling tones were incompletely neutralized for both young and middle-aged speakers. Lexical frequency showed little effect on the incomplete tonal neutralization; there were significant differences in the turning point of the two falling tones in syllables with both high and low lexical frequency. However, homophone neighborhood density showed an effect on the incomplete neutralization between the two falling tones, reflected in significant differences in the slope and turning point of the F0 velocity profiles between the two tones carried by syllables with low density but not with high density. Moreover, homophone neighborhood density also affected the duration, the turning point of F0 curves, and velocity profiles of the T1- and T4-syllables. These results are discussed with consideration of social phonetic variations, the theory of Hypo- and Hyper-articulation (H&H), the Neighborhood Activation Model, and communication-based information-theoretic accounts. Collectively, these results broaden our understanding of the effects that lexical properties have on the acoustic details of lexical tone production and tonal sound changes.

KEYWORDS

lexical frequency, homophone neighborhood density, tonal neutralization and merger, sound change, speech production, Dalian Mandarin

OPEN ACCESS

EDITED BY
Eleanor Chodroff,
University of York, United Kingdom

REVIEWED BY
Vsevolod Kapatsinski,
University of Oregon, United States
Yuan Liang,
The Education University
of Hong Kong, Hong Kong SAR, China

*CORRESPONDENCE
Yifei Bi
yifeibi@usst.edu.cn

SPECIALTY SECTION
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

RECEIVED 01 February 2022
ACCEPTED 13 October 2022
PUBLISHED 24 November 2022

CITATION
Bi Y and Chen Y (2022) The effects of lexical frequency and homophone neighborhood density on incomplete tonal neutralization.
Front. Psychol. 13:867353.
doi: 10.3389/fpsyg.2022.867353

Introduction

The notion of neutralization presupposes the concept of contrast. Neutralization refers to sound change where a contrast that exists in a language is lost in some particular contexts in its synchronic grammar [see reviews in Yu (2011), Kubozono and Giriko (2018), and references therein]. Existing acoustic studies on neutralization have explored various factors ranging from lexical properties (e.g., word frequency and orthography) to speaker characteristics (e.g., geographical locations of the population and speaking style) (Warner et al., 2004, 2006; Kharlamov, 2014; Roettger et al., 2014; Braver and Kawahara, 2016; Nicenboim et al., 2018; Kong and Shengyi, 2019).

One interesting area of neutralization research examines the acoustic characteristic of constant devoicing. These studies focused on Indo-European languages, such as Dutch (Warner et al., 2004, 2006), German (Roettger et al., 2014; Nicenboim et al., 2018), Catalan (Dinnsen and Charles-Luce, 1984), and Russian (Kharlamov, 2014; Matsui et al., 2017). For example, Warner et al. (2004) compared the acoustic realization of both long and short vowels before the Dutch alveolar /t/ and /d/ coda. They showed that vowels preceding an underlying /d/ were significantly longer than those preceding an underlying /t/. Furthermore, following long vowels, underlying /t/ had a longer burst duration than underlying /d/. Their study exemplifies how the so-called neutralized consonants (based on impressionistic observation) may nevertheless exhibit reliable acoustic differences and show characteristics of incomplete neutralization with data elicited with a controlled experimental design.

Studies have looked at a range of factors that can possibly encourage incomplete neutralized voicing contrasts in speakers. For example, orthography has been argued to bias speakers to "artificially" hyper-articulate the /t/ vs. /d/ contrast given their different written forms (e.g., Fourakis and Iverson, 1984; Warner et al., 2004). Another commonly discussed factor is the voicing contrasts of the stimuli in the language, which presumably motivates the speakers to preserve the underlying voicing distinction between minimal pairs (in comparison to non-minimal pairs) (Port and O'Dell, 1985; Ernestus and Baayen, 2006; Kulikov, 2012). Kharlamov (2014) examined the role of orthography, phonology, and elicitation tasks (reading vs. picture naming) in the acoustic realization of Russian voicing contrasts and argued that these factors influence the incomplete neutralization of Russian voicing contrasts through different acoustic parameters. Matsui et al. (2017) further showed the importance of controlling the lexical frequency and minimal-pair effects in investigating incomplete neutralization in Russian. Despite the increasing number of studies on incomplete voicing neutralization, there is still a lack of consensus on how exactly different factors condition neutralization and a lack of research that directly examines lexical effects such as phonological neighborhood density and lexical frequency on (incomplete) voicing neutralization.

Compared with the body of quantitative studies on segmental voicing neutralization in a wide range of (Indo-European) languages, much less research has focused on neutralization at the suprasegmental level, such as lexical tones. A handful of studies have investigated lexical tonal neutralization in Cantonese (e.g., Bauer et al., 2003; Mok et al., 2013; Cheng, 2017; Liang, 2018; Lin et al., 2021). A number of studies have investigated the status of tonal contrast between two merging tonal pairs. Further research has been conducted on Standard Chinese, where researchers have investigated possible neutralization of the lexical Rising tone (LR) with the sandhi rising variant (SR) of the Low tone, which is realized with a comparable rising F0 contour as the lexical Rising tone when preceding another Low tone (Chen and Yuan, 2007; Cheng et al., 2013; Yuan and Chen, 2014; Li and Chen, 2015; Nixon et al., 2015; Lin and Hsu, 2018; Politzer-Ahles et al., 2019).

Only a few studies have investigated the possible effects of lexical properties such as word frequency on tonal neutralization/merger. For example, Yuan and Chen (2014) explored the acoustic characteristics of the LR and SR in telephone conversations and broadcast news speech. They found SR is different from LR in terms of the magnitude of the F0 rise and the time span of the F0 rise. Furthermore, they discovered that SR in the most highly frequent words (>1,000 in frequency counts of 3,431,707 words in the Xinhua newswire) showed a greater difference from the LR than in less-frequent words. Furthermore, Mok et al. (2013) investigated the effect of word (token) frequency on Cantonese tone merger. Slight differences were shown in the tone merger between high-frequency and low-frequency words. Finally, Kong and Shengyi (2019) studied the effect of frequency on tonal reduction in Standard Chinese. Their results showed that the acoustic characteristic of reduction-induced neutralized tones (and the tone-carrying syllables) correlates directly with lexical frequency. However, much is still to be understood concerning how different factors condition tonal neutralization.

In this study, we aimed to shed light on the role of lexical properties in a sound-change-related process of tonal (incomplete) neutralization. There has been increasing interest in studying the effects of lexical properties on speech production. For example, one widely studied lexical property is word frequency. High-frequency words are typically produced with a shorter duration, reduced vowels, and reduced pitch range than low-frequency words (e.g., Pluymaekers et al., 2005; Zhao and Jurafsky, 2007; Mousikou and Rastle, 2015). Another key lexical property is phonological neighborhood density. Words from high (dense) neighborhood density are typically produced faster, more accurately, and hyper-articulated compared with words from low (sparse) neighborhoods (e.g., Munson and Solomon, 2004; Wright, 2004; Baus et al., 2008;

Gahl, 2008; Dell and Gordon, 2011; Scarborough, 2013; Gahl and Strand, 2016).

It is essential to note that with studies based mainly on Indo-European languages, the phonological neighborhood is typically defined with the one-phoneme difference rule. Phonological neighbors are two words that differ in only one phoneme by substitution, deletion, or addition (Luce and Pisoni, 1998; Luce et al., 2000; Vitevitch and Luce, 2016). Kapatsinski (2006) proposed a new method and defined words as phonological neighbors if they share at least two-thirds of their total segmental string. This draws upon evidence from lexical decision tasks, naming reaction times, and familiarity rating.

Yao and Sharma (2017) have proposed that in tonal languages, such as Mandarin, to define phonological neighbors by the one-phoneme/tone difference rule: any two syllables that only differ in one phoneme or tone are phonological neighbors. Given the abundant Mandarin homophones (i.e., monosyllabic morphemes with the same segmental syllables and tone), psycholinguistic studies on Chinese spoken word production/recognition often employ the notion of homophone neighbors (e.g., Chen et al., 2009, 2016; Wang et al., 2012), including only words that share both segments and tone. This study follows that tradition and defines neighborhood as homophone neighborhood, with homophones sharing the same segmental syllable and lexical tone. The empirical base of our investigation is Dalian Mandarin given the reported ongoing sound change and (incomplete) neutralization concerning two lexical tones.

Dalian Mandarin

Dalian Mandarin is a dialect of Mandarin, mainly spoken in the urban areas of Dalian City in Northeast China, about 460 km from the capital Beijing. Dalian Mandarin belongs to the Jiao-Liao Mandarin dialect group, a major Sinitic Mandarin group. Song (1963) states that Dalian Mandarin has four lexical tones produced in isolation: T1 has a falling and slight rising F0 contour (312), T2 a rising F0 contour (34), T3 a dipping contour (213), and T4 a falling contour (53). Here, the numerical numbers represent the pitch levels/ranges, following Chao's pitch annotation system (Chao, 1968), where 1 refers to the lowest end of a speaker's pitch range, and 5 is the highest end.

Sound changes have been reported in T1 and T4 (Gao, 2007; Liu, 2009), and they are both realized with a high-falling F0 contour. Gao (2007) conducted an acoustic analysis of data data collected from three generations of (young: aged below 29, middle: aged from 50 to 59, and old: aged from 70 to 80). The results showed that the old-generation speakers produced a different citation form of T1 (411 rather than 312) and T4 (52/51 instead of 53). However, for the middle and young-generation speakers, T1 and T4 have become even closer and are transcribed to share the citation form (51). Liu (2009) also concluded that in present-day Dalian Mandarin, T1 is now a high-falling tone (51).

Figure 1 plots the average F0 contours of the four lexical tones, with each tonal contour based on 20 samples produced by a young male native speaker (born in 1990). T2 has a rising F0 contour (35) and T3 a dipping contour (213). T1 and T4 are both realized with a falling contour (51), although it is not clear to what extent they have merged.

The current study

The current work was inspired by two notable observations. First, no quantitative analysis has been conducted on whether T1 and T4 in Dalian Mandarin are completely neutralized. Regardless, impressionistically, they seem to have merged to have the same tonal identity, but likely remained incompletely neutralized. This possibility is then similar to the incomplete voicing neutralization in many Indo-European languages. If this was the case, detailed acoustic analyses and proper statistical analyses of data produced by a sufficiently large number of participants are essential for us to detect the potentially subtle tonal differences. Second, previous Dalian Mandarin studies only elicited a limited set of frequent words [i.e., the so-called vocabulary of daily uses in Gao (2007)]. We know that factors such as different lexical frequencies and phonological neighborhood density could significantly affect the retrieval and production of spoken words. These factors could also be further conditioned by speakers' age [e.g., Gordon and Kurczek (2014) on the diminished facilitation effect of neighborhood density due to aging]. Given the potential ongoing changes of these two tones, speaker age is likely to have a substantial impact on tonal realizations. The goal of this study, therefore, aimed to investigate the possible impact of speakers' age, lexical frequency, and homophone phonological neighborhood density on the (in)complete neutralization of T1 and T4 in Dalian Mandarin. Specifically, the following questions will be addressed: (1) Are the two falling tones neutralized for middle-aged and young-generation speakers of Dalian Mandarin? (2) How do lexical frequency and homophone neighborhood density affect tonal realization and neutralization?

Materials and methods

Participants

30 middle-aged (mean age: 50; SD: 3.6) and 30 young (mean age: 22; SD: 3.6) native speakers of Dalian Mandarin participated in the experiment. The participants were selected from the urban area of Dalian City, including the districts of Sha Hekou, Zhong Shan, Xi Gang, and Gan Jingzi, and self-reported to have normal vision and no history of speech disorders. Informed

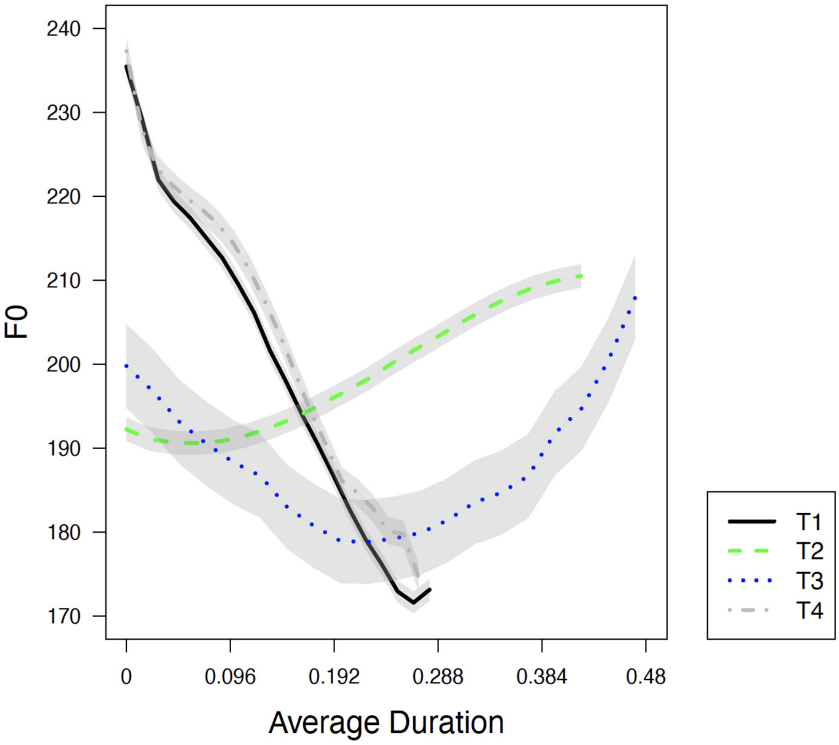

FIGURE 1
Four lexical tones in present-day Dalian Mandarin produced in isolation, with each tonal contour based on 20 samples produced by a young male native speaker. Lines represent the mean. Shaded areas stand for the standard error of the mean (±1.3 for T1, ±1.4 for T2, ±5 for T3, and ±1.6 for T4).

consent was obtained from all participants before beginning the experiment, and all participants were paid to take part.

Materials

The target stimuli included minimal T1 and T4 pairs with different lexical properties. Due to the difficulty of finding sufficient stimuli for low lexical frequency (LF) with low homophone neighborhood density (LD), no such stimuli were used in the experimental design. As a result, the stimulus sets consist of high lexical frequency (HF) and low lexical frequency (LF) syllables with high homophone neighborhood density (HD). The lexical frequency in HD and LD syllables always had HF. In total, 90 syllables were selected that had four lexical conditions: HF, LF, HD, and LD. Each lexical condition had 30 syllables, with 15 T1 and 15 T4 syllables. We used the corpus of spoken Chinese based on film subtitles by Cai and Brysbaert (2010). According to the corpus, which is based on 33,546,516 words, the logged frequency[1] of the high-frequency monosyllabic stimuli (HF) in this experiment is between 2.81 and 4.9 per million, million, with an average of 3.82. The logged frequency of the low-frequency monosyllabic stimuli (LF) is between 0.6 and 2.0, with an average of 1.55.

We used the Modern Chinese Dictionary (5th Version; The Commercial Press) to calculate homophone neighborhood density. There is no standard criterion in the literature on the specific number of homophones to define HD vs. LD. Chen et al. (2009) defined characters with more than seven homophone mates as HD, while those with fewer homophone mates as LD. Wang et al. (2012) used a threshold of nine for HD and two to eight for LD words. Yip (2002), however, has a lower threshold (i.e., six homophone mates) for HD and two for LD. In our stimuli set, we strived for the right balance between HD and LD stimuli (30 for each): for HD syllables, their number of homophone mates is above seven (and up to 84); for LD syllables, their number of homophone mates is below six (and above two). The average number of homophones for the HD and LD syllables is 19 and 4, respectively.[2] Take "班 /ban1/" (Tones 1–4 are denoted by digits) and "猜 /cai1/" as examples. The HD syllable "班 /ban1/" has 11 homophones, including 般 ('type'), 颁 ('to issue'), 斑 ('spot'), and 搬 ('to move'). They share the same

[1] This value is based on $\log_{10}$ (FREQcount + 1), in which FREQcount is the number of times the word appears in the corpus.

[2] The calculation was also double-checked with data base (http://dowls.site/) generated by Neergaard and Huang (2019).

segmental syllables and tones /ban1/ but are represented by different Chinese characters. The LD syllable "猜 /cai1/" has only one homophone, "偲 /cai1/." Moreover, "班 /ban1/" and "猜/cai1/" are both high-frequency syllables, so "班 /ban1/" was chosen as an HF syllable with HD and "猜 /cai1/" as an HF syllable with LD.

Procedure

Participants were exposed to a learning phase formed of four trials using frequent syllables (not used in the test) to familiarize them with the experimental procedure before the test. The 90 target syllables were divided into six blocks, and each block was composed of 15 trials. In the test phase, participants took a self-paced break between the blocks. The order of the trials was randomized for each participant. There was no repetition in any of the trials.

The experiment was conducted in E-prime 2.0 run on a laptop equipped with a Creative SBX-FI5.1 pro sound card. Participants were placed in a quiet room and were asked to read the stimuli in Chinese characters on the computer screen. The responses were recorded using a condenser microphone, and the recordings were stored directly on the computer's hard disk.

Data preparation

The acoustic analysis of all data was conducted in Praat (Boersma and Weenink, 2017). All the sound files were manually segmented. The onset and offset of target vowels or tone-bearing syllable rhymes (i.e., vowel or vowel with a nasal coda) determined the time intervals for extracting duration and F0 values. F0 values were sampled at 20 equidistant measurement points using a Praat script. F0 values were converted to speaker-specific z-scores to reduce cross-speaker variability for plotting and statistical analysis. Following Rose (1987), z-scores were calculated with $F0^{Zscore} = (x_i - m)/s$, where m is the mean value of x_i and s is the standard deviation, calculated per speaker.

Analysis of duration

The duration of T1- and T4-bearing syllable rhymes were modeled using linear mixed-effects (LME) in R (R Core Team, 2015), lme4 (Bates et al., 2014), and lmerTest (Kuznetsova et al., 2017). For the analysis of the effect of lexical frequency, the final model included the following fixed effects: generation (middle-aged vs. young), lexical frequency (HF vs. LF), and tone (T1 vs. T4) (without interaction). The random effects included by-subject slopes for the effects of lexical frequency and tone (without interaction) and by-item intercept. For homophone neighborhood density, the final model included the fixed effects of generation (middle-aged vs. young), homophone neighborhood density (HD vs. LD), and tone (T1 vs. T4) (without interaction). Also, the random effects included by-subject slopes for the effect of homophone neighborhood density and tone (without interaction) and by-item intercept. Note that for both analyses, the fixed factors were added stepwise, and their effects and interactions on model fits were evaluated *via* model comparisons based on log-likelihood ratios. The estimate (β), standard error *(SE)*, and *t*-values are reported below.

Analysis of F0

Functional data analysis (FDA) (Ramsay et al., 2009a; Gubian et al., 2015) was used to analyze the F0 values. FDA provides a method for analyzing a dataset that consists of entire curves with different durations. Two main procedures were conducted: smoothing with a linear time registration and a Functional Principal Component Analysis (FPCA). Smoothing (with a roughness penalty) was realized by the B-spines (De Boor, 2001). A linear time registration scaled all the smoothed curves into a normalized duration (i.e., 1), used for further FPCA analysis. FPCA provided a model for approximating the (normalized) smoothed curves using the mean curve and a number of Principal Component (PC) curves and their weights (PC scores), based on the formula $f(t) \approx \mu(t) + \sum_{j=1}^{\propto} s_j * PC_j(t)$. Here, $\mu(t)$ is the mean curve, s_j is the PC score (PCs) and $PC_j(t)$ is the corresponding PC curve.

All FDA was carried out in the FDA R package (Ramsay et al., 2009b). Apart from F0 curves, their instantaneous velocity profiles (which indicate the declining speed of the two falling tones) also reflect dynamic lexical tone articulation (Gauthier et al., 2007; Cheng et al., 2014). Therefore, both the F0 curves and the velocity profiles of the F0 values with the FDA were analyzed. A functional *t*-test was used to calculate the absolute value of the t-statistic at each sampling point of the (normalized) smoothed curves (Ramsay et al., 2009a). A functional *t*-test extends the rationale of the well-known *t*-test and can compare the means of two groups' curves within specific time domains. Moreover, the first two PC scores (s_1 and s_2), which represent most of the variation of the smoothed curves, were further used for performing LME modeling.

Different models were constructed for lexical frequency and homophone neighborhood density effects. There were four lexical frequency effect models for s_1 and s_2 in F0 curves and F0 velocity profiles, respectively. The final models included the following fixed effects: generation, lexical frequency and tone (with interaction). The random effects included by-subject slopes for the effects of lexical frequency and tone (with interaction) and by-item intercept. There were also four models for s_1 and s_2 in F0 curves and F0 velocity profiles for the homophone neighborhood density effect, respectively. The final models included the following fixed effects: generation, Homophone Neighborhood Density (HND), and tone (with interaction). The random effects included by-subject slopes for the effects of HND and tone (with interaction) and the by-item intercept. Following the advice of one reviewer, we also checked the model constructions with the function (model. selection) in the developed library MuMIn in R (Barton, 2009). Results confirmed that the final model constructions (specified previously) for rhyme duration and F0 were the best considering the AICs weights.

Results

Duration of the T1- and T4-carrying syllables

Figures 2A,B show the violin plot of the duration of the rhyme part of syllables for T1- and T4 in the four lexical conditions (HF, LF, HD, and LD) produced by middle-aged participants and young participants. For the model of lexical frequency, there was a significant main effect of generation (β = 0.25, *SE* = 0.003, $t = -10^{***}$), while for the homophone neighborhood density model, there was a significant main effect of generation (β = −0.03, *SE* = 0.003, $t = -9.9^{***}$) and homophone neighborhood density (β = −0.02, *SE* = 0.006, $t = -2.6^{*}$), which we will examine further in Section "Effects of homophone neighborhood density." It is important to note here that in both models, the lexical tone was not a significant main effect, suggesting the neutralization of T1 and T4 concerning the duration of their tone-bearing syllables.

F0 curves and velocity profiles

Functional data analysis of T1 and T4 F0 contours

Figures 3A,B show the raw F0 contours of T1 and T4 with a normalized duration from middle-aged participants and young participants. The figures show that both T1 and T4 have similar falling F0 contours and ranges (between about 140 and 260 Hz) in both generations. Based on the results of FDA, Figure 4 shows the average F0 curves for T1 and T4 in the four lexical conditions and the results of a between-participant functional *t*-test for the young participants. Figure 5 shows the average of the F0 velocity profiles for T1, and T4 in the HD and LD lexical conditions produced by young participants and their functional *t*-test statistics. For the middle-aged participants, the results were similar but are not shown in this paper. In Figures 4, 5, dotted lines represent the 0.05 critical values for the t-statistic. The higher statistic represents a more conservative critical value.

Although insights into the mean F0 curves and F0 velocity profiles of T1 and T4-based functional *t*-tests were achieved, it is important to note that functional *t*-test only considers the t-statistic of each sampling point in the smoothed curves. Therefore, the LME modeling was employed for the two principal components (PC) scores (s_1 and s_2) to investigate further the T1 and T4 F0 contours (curves and velocity profiles). In addition, LME modeling allows for variations due to individual speakers and stimulus items to be taken into account for a greater understanding of the neutralization of the two tones.

Linear mixed-effects modeling of principal component scores

We used FPCA to analyze PC scores. To show how FPCA works, the FPCA results of T1 for F0 curves and F0 velocity profiles in the HF lexical condition for the young participants were plotted in Figure 6. Each panel shown in the solid line is the mean curve μ (t). The ± curves were obtained by adding to or subtracting from μ (t) the curves (a) $\sigma(s_1)$ * $PC_1(t)$ and (b) $\sigma(s_2)$ * $PC_2(t)$. σ denotes standard deviation. PCs are numbered from 1 onwards, and the rank reflects the decreasing percentage of variance in the input data that can be explained by the PCs. As shown in Figure 6, for F0 curves (Figures 6A,B), the FPCA outputs indicate that s_1 and s_2 could explain the most variation in the HF lexical condition (77.2 and 13.8%, respectively). Figure 6A suggests that PC1 (s_1) mainly alters the slope of the F0 curves. Figure 6B suggests that PC2 (s_2) altered the turning point of the curves. The interpretations of s_1 and s_2 are consistent with Gubian (2011). The same goes for the instantaneous F0 velocity profiles (given by the slope of F0 at a single point in time), which indicates the declining speed of the two falling tones. s_1, s_2, s_3, $s_{4,}$ and s_5 could explain 30.5, 19, 17.2, 9.8, and 7.8% of the variation, respectively. Like the F0 curves, the s_1 and s_2 of the velocity profile, which account for more variance, were analyzed (Figures 6C,D). Figure 6C suggests that PC1 (s_1) mainly alters the slope of the F0 velocity profiles. Figure 6D suggests that PC2 (s_2) altered the turning point of the velocity profiles. The PC scores enable us to conduct further quantitative analysis of the effect of the two lexical factors on tonal production using LME modeling with s_1 and s_2.

We performed LME modeling with s_1 (indicating the slope) and s_2 (indicating the turning point) for F0 curves and F0 velocity profiles (i.e., the declining speed of two falling tones). Additionally, different models were fitted for lexical frequency and homophone density effects and the factors of speaker generation and lexical tonal identity. The significant results are presented in Tables 1, 2.

For F0 curves (Table 1), the dominant effect for s_1 lies in the interaction of speaker generation and tonal identity. Further details of the interaction are discussed in Section "Generational differences in the F0 characteristics of T1 and T4." For s_2, there were significant three-way interactions (Generation × Tone × Lexical frequency and Generation × Tone × Homophone neighborhood density). Therefore, separate models from the subset data about generation, tone, and lexical conditions (lexical frequency and homophone neighborhood density) were run to reveal the differences between the two falling tones for each lexical condition in each generation. The significant results are presented in Table 3.

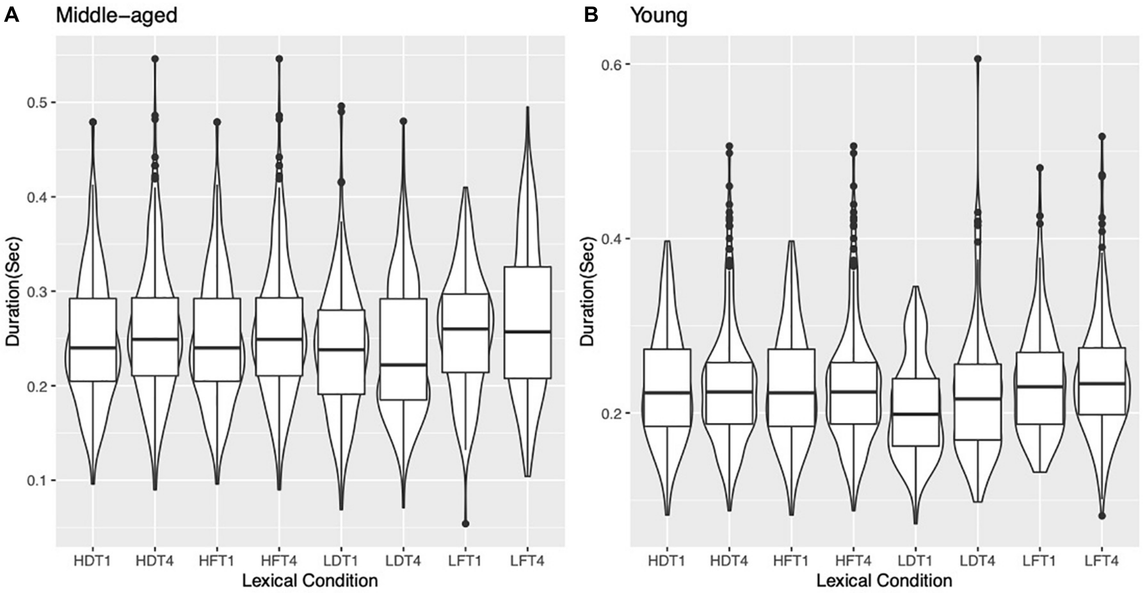

FIGURE 2
Duration of the rhyme part of syllables for T1 and T4 produced by **(A)** 30 middle-aged participants and **(B)** 30 young participants in the four lexical conditions (HF, LF, HD, and LD). For example, HFT1 represents syllables with high lexical frequency in T1.

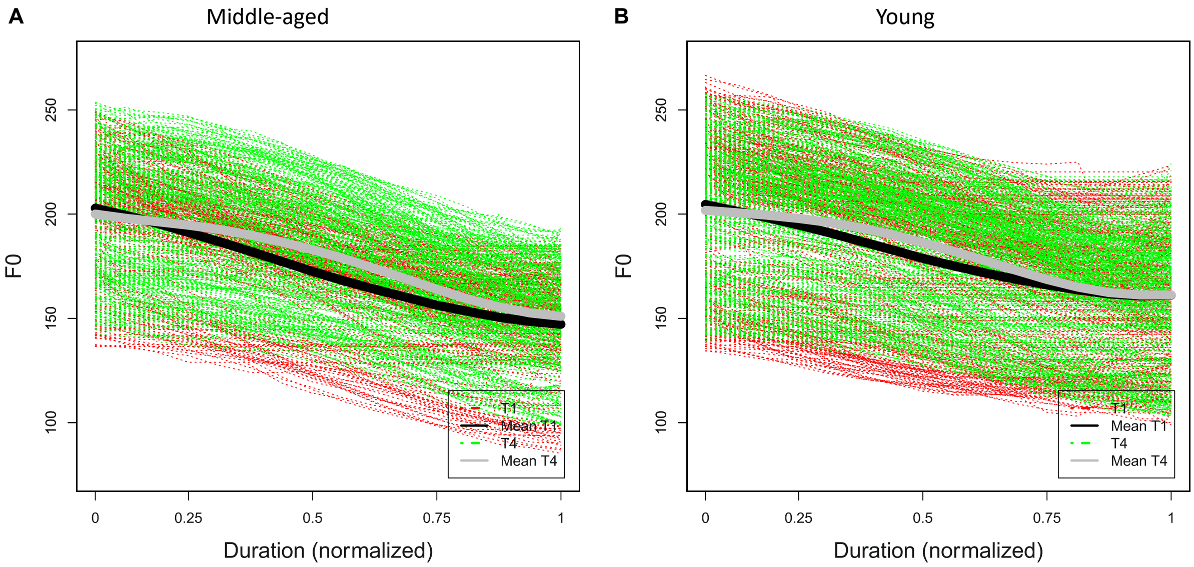

FIGURE 3
Time-normalized raw F0 contours of T1 and T4 of **(A)** 30 middle-aged participants and **(B)** 30 young participants.

From Table 3, significant differences between T1 and T4 in s_2 (corresponding to the turning point of F0 curves) can be seen for all four lexical conditions in middle-aged and young participants. Specifically, the turning point for T4 was earlier than that for T1 across generations and lexical conditions. There were also significant two-way interactions for the s_2 of F0 curves for speaker generation and lexical frequency, lexical frequency and tonal identity, and speaker generation and homophone neighborhood density. Separate models from subset data were also run to reveal the differences. Results showed that the F0 turning point of syllables with high lexical frequency was earlier than syllables with low lexical frequency for T1. However, for T4, the F0 turning point of syllables with low lexical frequency was earlier than syllables with high lexical frequency. This raises

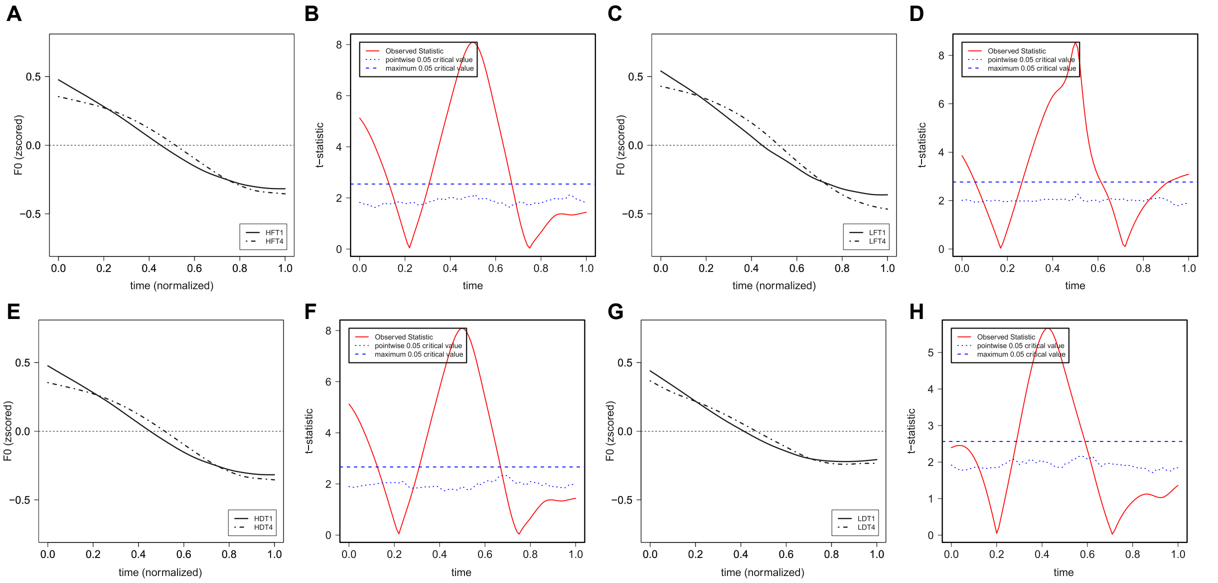

FIGURE 4
The average of the (normalized) F0 curves for T1 and T4 in the four lexical conditions produced by young participants [**(A)** HF; **(C)** LF; **(E)** HD; **(G)** LD] and their functional *t*-test statistic [**(B)** HF; **(D)** LF; **(F)** HD; **(H)** LD]. The solid lines represent the F0 curves of T1, and the dot-dash lines represent the F0 curves of T4.

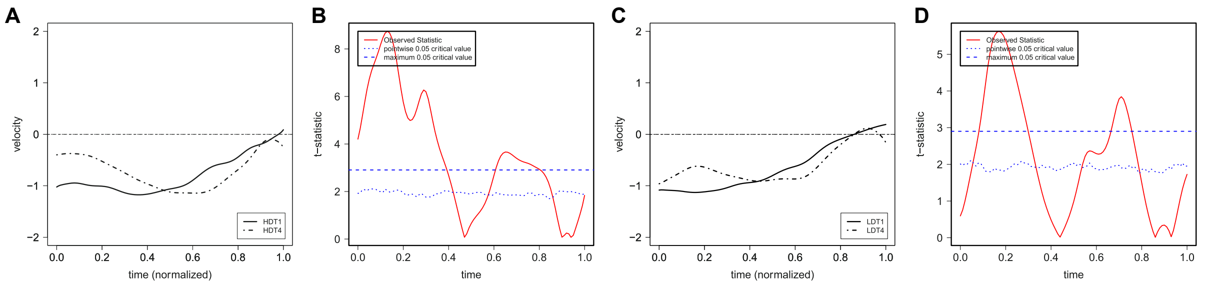

FIGURE 5
The average of the (normalized) F0 velocity profiles for T1 and T4 in the HD and LD lexical conditions produced by young participants [**(A)** HD; **(C)** LD] and their functional *t*-test statistic [**(B)** HD; **(D)** LD]. The solid lines represent the F0 curves of T1, and the dot-dash lines represent the F0 curves of T4.

questions regarding the exact effect of lexical frequency on tonal realization. Further details of the interaction between speaker generation and homophone neighborhood density are reported in Section "Effects of homophone neighborhood density."

For F0 velocity profiles (Table 2), we observe a significant interaction of lexical frequency and tone for s_2. This means that there were significant differences in the F0's turning point between the two falling tones in different lexical frequencies. The pattern is similar to the contradictory findings of F0 curves in T1 and T4. Homophone neighborhood density, on the other hand, showed significant three-way interactions with both Generation and Tone for both the s_1 and s_2 of the F0 velocity profiles. Separate models from the subset data of generation, tone, and homophone neighborhood density were conducted. The main significant results for the models are presented in Table 4.

Table 4 shows a significant difference in T1 and T4 for s_1 in the LD lexical condition among the middle-aged participants. The slope of the F0 velocity profiles for T4 was steeper than those for T1. For s_2, T1 and T4 differed again only in LD, but for both middle-aged and young participants, the turning point of T4 was earlier.

Generational differences in the F0 characteristics of T1 and T4

This section further reports the details of the incomplete neutralization of T1 and T4 by comparing how speakers of the two generations produce each of the two tones. The specific direction of the sound changes in T1 and T4 can

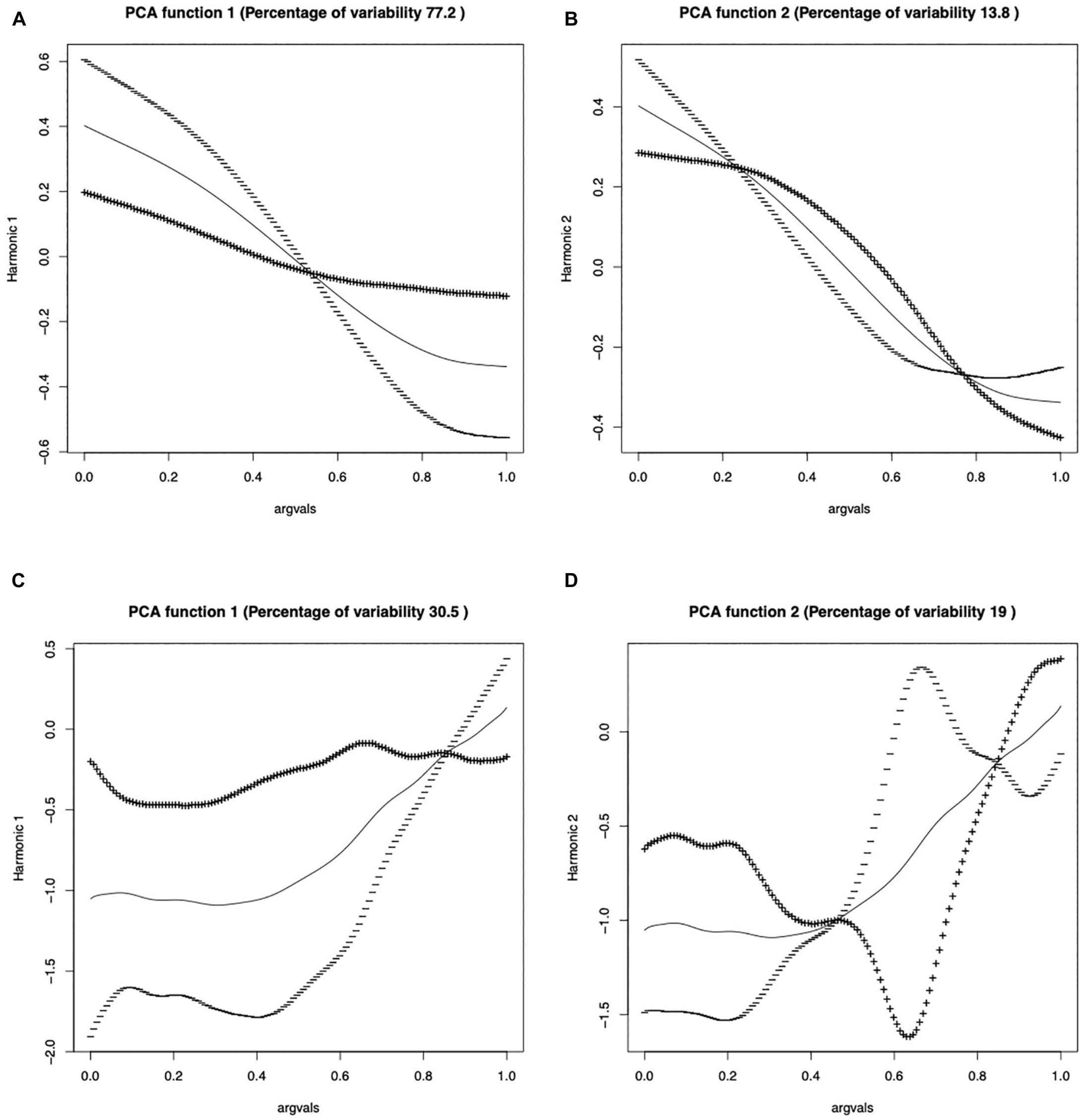

FIGURE 6
The results of FPCA for **(A)** PC1 (t) and **(B)** PC2 (t) of T1 for F0 curves [**(A)** s_1; **(B)** s_2] and F0 velocity profiles [**(C)** s_1; **(D)** s_2] in the HF lexical condition for the young participants.

also be explored. Note that for F0 curves, the dominant effect lies in the significant interaction of Generation and Tone of s_1, regardless of lexical frequency and neighborhood density. Figure 7 shows the average F0 curves of T1 and T4 and the results of functional *t*-tests between the two generations. There were significant differences in the initial and later parts of the F0 curves between the two generations in the production of both T1 and T4. Results of the LME modeling (Table 5) showed a significant difference between middle-aged speakers and young speakers for T1 and T4. Specifically, the slope (indexed *via* the s_1 of F0 curves) of both T1 and T4 was steeper for young participants than the one for middle-aged participants.

Effects of homophone neighborhood density

As stated previously, a key effect of homophone neighborhood density was found on the rhyme duration of the Tone 1- and Tone 4-carrying syllables. Furthermore, homophone neighborhood density also interacted significantly with generation and tone. These findings indicate the effect of homophone neighborhood density on lexical production, which has not been reported in the literature. Therefore, this factor will be detailed further in the proceeding section.

First, LME modeling was performed with syllable rhyme duration as the dependent variable and homophone neighborhood density as an independent variable. The by-subject slope for the effect of homophone neighborhood density

TABLE 1 Summary of linear mixed-effects modeling for F0 curves.

Lexical condition	PCs	Fixed effects	β	*SE*	*df*	*t*	*p*
Lexical frequency	s_1	Generation × Tone	0.04	0.01	3862	3.5	***
	s_2	Lexical frequency	0.05	0.01	48	4.7	***
	s_2	Tone	0.05	0.006	83	7.4	***
	s_2	Generation × Lexical frequency	−0.05	0.008	2364	−6.8	***
	s_2	Lexical frequency × Tone	−0.09	0.01	38	−6.4	***
	s_2	Generation × Tone × Lexical frequency	0.1	0.01	3255	9.1	***
Homophone neighborhood density	s_1	Generation × Tone	0.04	0.01	3175	3.1	**
	s_2	Tone	0.05	0.008	69.21	5.9	***
	s_2	Generation × Homophone neighborhood density	0.02	0.007	1752	2.5	*
	s_2	Generation × Tone × Homophone neighborhood density	−0.02	0.009	1471	−2.3	*

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE 2 Summary of linear mixed-effects modeling for F0 velocity profiles.

Lexical condition	PCs	Fixed effects	β	*SE*	*df*	*t*	*p*
Lexical frequency	s_2	Lexical frequency × Tone	−0.18	0.08	102	−2.2	*
Homophone neighborhood density	s_1	Generation × Homophone neighborhood density	−0.13	0.06	2571	−3.0	**
	s_1	Generation × Tone × Homophone neighborhood density	0.25	0.08	2580	3.1	**
	s_2	Homophone neighborhood density	−0.15	0.05	77	−3.1	**
	s_2	Homophone neighborhood density × Tone	0.24	0.07	78	3.5	***
	s_2	Generation × Tone × Homophone neighborhood density	−0.15	0.07	1711	−2.1	*

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE 3 Summary of linear mixed-effects modeling for F0 curves (from subset data of generation, tone, and lexical conditions).

Generation	Lexical condition	Tone	PCs	β	*SE*	*df*	*t*	*p*
Middle	HF	T1 vs. T4	s_2	0.04	0.002	53	6.4	***
	LF			−0.04	0.01	39	−2.6	*
	HD			0.04	0.009	51	4.6	***
	LD			0.05	0.01	34	4.0	***
Young	HF			0.05	0.007	53	6.4	***
	LF			0.05	0.01	34	3.7	***
	HD			0.06	0.009	38	6.2	***
	LD			0.03	0.009	29	3.7	***

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE 4 Summary of linear mixed-effects modeling for F0 velocity profiles (from subset data of generation, tone, and homophone neighborhood density).

Generation	Lexical condition	Tone	PCs	β	*SE*	*df*	*t*	*p*
Middle	LD	T1 vs. T4	s_1	−0.25	0.09	37	−2.8	**
			s_2	0.18	0.07	32	2.5	*
Young			s_2	0.25	0.09	15	2.8	**

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

and the by-item intercept were included as random effects in the final model. For young speakers, the rhyme of T1-carrying syllables showed a significant main effect of homophone neighborhood density (β = −0.02, *SE* = 0.01, *t* = −2.0*), with the duration in LD on average 20 ms shorter than HD. The rhyme duration of T4-carrying syllables was also shorter (15ms) in LD-syllables than HD-syllables (β = −0.015, *SE* = 0.007, *t* = −2.1*). Similar results were found for middle-aged speakers (T1: 17 ms shorter in LD; β = −0.017, *SE* = 0.007, *t* = −2.5*; T4: 14 ms shorter in LD; β = −0.014, *SE* = 0.006, *t* = −2.2*).

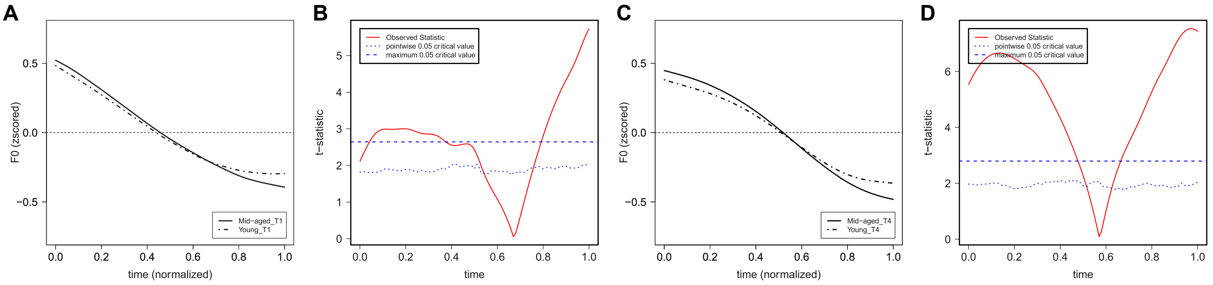

FIGURE 7
The average of the (normalized) F0 curves between the middle-aged and young participants for T1 **(A)** and T4 **(C)** and their functional *t*-test statistics in panels **(B,D)**. The solid lines represent the F0 curves of middle-aged participants, and the dot-dash lines represent the F0 curves of young participants.

TABLE 5 **Summary of linear mixed-effects modeling for F0 curves.**

Generation	F0	Tone	PCs	β	*SE*	*df*	*t*	*p*
Middle vs. Young	Curves	T1	s_1	0.04	0.01	32	3.6	**
		T4		0.07	0.03	30	2.8	**

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE 6 **Summary of linear mixed-effects modeling for F0 curves and F0 velocity profiles in the lexical condition of homophone neighborhood density for T1 and T4.**

Generation	F0	Lexical condition	Tone	PCs	β	*SE*	*df*	*t*	*p*
Middle	Velocity profiles	HD vs. LD	T1	s_2	−0.19	0.08	52	−2.2	*
			T4	s_2	0.12	0.05	52	2.2	*
Young	Curves		T1	s_2	−0.03	0.008	29	−3.1	**
	Velocity profiles				−0.12	0.06	31	−2.1	*
	Curves		T4	s_2	−0.03	0.008	29	−4.4	***
	Velocity profiles				0.27	0.07	37	4.0	***

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Second, LME modeling was performed for the FPCA results of the F0 data based on the significance tests reported earlier. Specifically, we focused on the s_2 of F0 curves and both s_1 and s_2 of F0 velocity profiles. The effect of homophone neighborhood density on T1 and T4 produced by speakers of the two generations was a leading point of interest. Separate models were run with homophone neighborhood density as the fixed effect, while the by-subject and the by-item intercept were included as random effects. Table 6 shows that there was a significant effect of neighborhood density on the s_2 (i.e., the turning point of F0 curves and F0 velocity profiles) between HD and LD for young speakers. For F0 curves, the turning point for LD was earlier than HD for both T1 and T4. For F0 velocity profiles, the turning point for LD was earlier than HD for T1 but was later than HD for T4. Significant differences were also found in the velocity profiles for the middle-aged speakers (s_2 in both T1 and T4). The turning point of F0 velocity profiles for LD was earlier than HD for both T1 and T4.

Discussion and conclusion

The current study investigated the acoustic realizations of two falling tones (i.e., T1 and T4) in Dalian Mandarin. The goal of this study was to understand the effects of (1) lexical properties (i.e., frequency and homophone neighborhood density) and (2) generation of speakers (i.e., middle-aged vs. young) on the neutralization of the two falling tones. Recordings of 45 pairs of T1 to T4 carrying syllables were elicited from 60 native participants who are from two different generations (i.e., middle-aged vs. young). The duration and F0 of the rhyme of the tone-carrying syllables were quantitatively analyzed.

Results showed no significant durational difference between the T1 and T4 tone-bearing syllable rhymes, contrary to the reports by Gao (2007) and Liu (2009). Concerning the F0 contours of the two falling tones, however, we found subtle but statistically significant differences (**Figure 3**), reflected in terms of both the F0 curves (**Figure 4**) and F0 velocity profiles

(Figure 5) of the tonal F0 contours. Generally speaking, T4 showed a consistently earlier F0 turning point than T1 for both generations of speakers. Although interactions were found among lexical properties, speaker generation, and tonal identity on the duration and F0 patterns of the two lexical tones (as discussed further below), our results confirmed that the neutralization of T1 and T4 is not complete. The general pattern of incomplete neutralization between T1 and T4 has remained rather stable across the middle-aged and young speakers.

Significant three-way interactions were found for lexical frequency, speaker generation, and tone (Table 1). Significant differences between T1 and T4 were found regardless of the frequency of the tone-carrying syllables and for both generations of speakers (Table 3). This suggests little effect of lexical frequency on the neutralization of T1–T4. Two-way interactions were also found between lexical frequency and tonal identity for the s_2 (turning point) of F0 curves and F0 velocity profiles (Tables 1, 2). For T1, the F0 turning point was earlier in syllables with high lexical frequency. For T4, the F0 turning point was earlier in syllables with low lexical frequency. It is not clear why frequency showed different effects on T1 and T4 tonal realization. Future research is needed to confirm and understand this effect.

For homophone neighborhood density, there were significant three-way interactions with speaker generation and tone (as shown in Tables 1, 2). In the low homophone neighborhood density condition, T4 showed a steeper falling slope than T1 (reflected in the s_1 of the velocity profiles) for the middle-aged speakers. Furthermore, T4 showed an earlier turning point (reflected in the s_2 of the velocity profiles) for both generations of speakers (Table 4). Focusing on individual tonal production, it was found that tones carried by syllables with low homophone neighborhood density were hyper-articulated with a longer duration than that with high homophone neighborhood density. There was also a significant effect of homophone neighborhood density on the s_2 of the F0 curves and velocity profiles between the high and low homophone neighborhood density conditions, but only for young speakers. Specifically, both T1 and T4 showed an earlier F0 turning point in the low homophone neighborhood density condition (Table 6).

Our results on the effect of speaker generation suggest that the T1 to T4 contrast in Dalian Mandarin (or their incomplete neutralization) is rather stable across the two generations. However, the different generations of speakers who took part in this study did show some differences in the specific F0 contours of the two lexical tones, suggesting ongoing changes in the F0 realization of both T1 and T4. The dominant effect lies in the significant interaction of Generation and Tone for the slope of F0 curves (s_1), regardless of lexical frequency and neighborhood density. As shown in Figure 7 and Table 5, the slope of F0 curves for both T1 and T4 were steeper for young participants than for middle-aged participants.

It is seen in the literature that a large number of speakers of Dalian mandarin are descendants of migrants from Shandong Province, whose native dialects (spoken in, e.g., Weihai and Yantai City) have falling F0 contours for both T1 and T4 cognates. It is generally accepted that the contact of these emigrants with local Dalian Mandarin speakers around 1904 (the Twentieth Century) initiated the merger between T1 and T4. Such language-contact-induced sound change is in line with the social variations and sound changes charted out in Labov (2001, 2011). In Gao (2007), we know that in Dalian Mandarin, the citation form of T1 was changed from 312 in speakers of age between 70 and 80 (old-aged speakers) to 411 in speakers of age between 50 and 59 (middle-aged speakers), and then from 411 to 51 of age below 29 (young speakers). For T4, it was changed from 53 in old-aged speakers to 52 in middle-aged speakers, and then from 52 to 51 in young speakers. In the participants of this study, both the middle-aged (mean age: 50; SD: 3.6) and young speakers (mean age: 22; SD: 3.6) realized the two tones as a falling tone (51). This suggests that the merger of the T1 and T4 has been rather stable among our two generations of speakers. From the reports in the existing literature and our data, it can be suggested that the two falling tones have neutralized gradually by approximation. That is, it is the gradual approximation of both the T1 and T4 tonal targets that has resulted in the merger of the two tones into the same falling tone contour (51), and the changes have been rather symmetrical. Garrett and Johnson (2013), however, reported contradictory results and found that sound change is typically directional and asymmetric in speech production and perception.

It is important to note that the subtle differences between T1 and T4 have remained stable across the two generations of our speakers. Given the lack of findings on the effect of speaker generation on the contrast between T1 and T4, we may conclude that the sound change reported in the literature (Gao, 2007; Liu, 2009) has already arrived at the state of incomplete neutralization among our middle-aged speakers. This suggests that while the merging of T1 and T4 has been completed by the time of their studies, the nature of the merger is incomplete neutralization. Languages, however, continually evolve, and this is also true for the two incompletely neutralized tones. Changes were observed in both T1 and T4 from the middle-aged to the young-generation speakers. In particular, the younger speakers produced both T1 and T4 with steeper F0 curves than the middle-aged speakers. Our pilot data also suggest that the acoustic features of T4 in Dalian Mandarin are similar to that of T4 in Standard Mandarin. Furthermore, native speakers of Dalian Mandarin are not able to tell the difference between the two falling tones. The change in citation form from 53 to 51 for T4 in Dalian Mandarin is probably due to the language contact with Standard Mandarin, especially for speakers from the young generation. Nevertheless, the way T1 has changed cannot be attributed to merely the influence of Standard Mandarin; otherwise, T1 and T4 should have become more different instead of being incompletely neutralized. It can be speculated that T1 became closer to the T4 instead of being closer to the T1 of Standard Mandarin because of the regional identity of the young generation. It is quite well-known that the characteristic of T1, a falling tendency, marks the regional identity of the local residents. It is likely that, speakers have therefore preferred to keep the falling F0 pattern of T1 instead of adjusting T1 to a high-level tone (as in Standard Mandarin). Future research is needed with, for example, questionnaires and interviews to verify speakers' regional identity and their preference for tonal acoustic realization.

Various factors have been discussed in the literature that condition sound changes. The effects of lexical frequency and homophone neighborhood density have been investigated in the current study. Results showed that the degree of T1–T4 neutralization does not vary as a function of lexical frequency. Specifically, no significant differences between the two falling tones were found for different lexical frequencies. In the literature on the role of lexical frequency in sound change, two different mechanisms have been posted: articulatorily-motivated and analogical changes. Articulatorily-motivated change typically affects high-frequency words first, while analogical sound change affects low-frequency words first (Phillips, 1984; Bybee, 2007). The effect of lexical frequency may also vary depending on the stage of sound change, namely, whether it is in progress or stable. For non-tonal languages, it has been claimed that the effect of lexical frequency could be the largest when the change is in progress and the smallest when the change has reached a stationary stage (Hay et al., 2015). In the case of Cantonese, which has tonal near mergers, Mok et al. (2013) found that word frequency had little impact. Cantonese tone mergers are assumed to be relatively stable. The current study echoes the findings in Cantonese and confirms the lack of lexical-frequency effect in the neutralization of tones at a relatively stable stage.

Frequency showed an effect on the acoustic realization of the individual lexical tones and their tone-carrying syllables. Specifically, for T1, the F0 turning point was found to be earlier in syllables with high lexical frequency, but for T4, the F0 turning point was earlier in syllables with low lexical frequency. This is a pattern that must be further replicated before any meaningful discussion can occur.

For the effect of homophone neighborhood density, results showed that with low homophone neighborhood density, we could observe a significant difference between the two falling tones. The slope of the F0 falling was found to be steeper in T4 than in T1. The current study is the first to examine the effect of homophone neighborhood density on tonal neutralization. Our results suggest that tones in syllables with low homophone neighborhood density tend to maintain their contrast, while those with high homophone neighborhood density may have been more completely neutralized. Further research is needed to verify this finding.

When examining homophone neighborhood density, it can be seen that the individual tones were hyper-articulated with a longer duration and a later F0 turning point in syllables with high homophone neighborhood density. To understand the patterns, it may be necessary to employ several models from the literature, namely the Neighborhood Activation Model and the communication-based accounts (e.g., Luce, 1986; Goldinger et al., 1989; Vitevitch and Sommers, 2003; Baus et al., 2008; Taler et al., 2010; Chen and Mirman, 2012; Gahl and Strand, 2016; Yao and Sharma, 2017; Arutiunian and Lopukhina, 2020; Karimi and Diaz, 2020). A crucial assumption of the Neighborhood Activation Model (NAM) is that the activation and inhibition of the target words during speech processing. During the processing of spoken words, the target word and all the other competitors (e.g., neighbors) are activated. NAM was originally proposed to model the results of word recognition. A number of studies, however, have also used the notions of activation and competition of targets and competitors modeled by the NAM to understand speech production. The number of lexical neighbors of a target word (i.e., its neighborhood density) influences the selection and production of the target word. It is easier to produce a target word with more lexical neighbors (dense phonological neighborhood density—more competitors). Typically, speech error rates, naming accuracies, and latencies have been examined to infer the effect of neighborhood density on production. In the current study, we found that the duration for HD syllables was longer than their LD counterparts. This suggests the possibility that there is an inhibitory effect of HD on tonal production. Needless to say, replication studies are needed to verify the findings and this interpretation.

Communication-based accounts support the idea that efficient communication is the main aim of language processing. If a target word is highly similar to its neighbors/competitors, it may generate high communication uncertainty. Thus, the speaker is expected to spend more time and articulatory effort to precisely produce the acoustic signal to increase the probability of it being accurately recognized. Both NAM and communication-based accounts may be related to the theory of Hypo- and Hyper-articulation (H&H) (Lindblom, 1990) in lexical production, which states that speakers produce strengthened phonetic forms when they anticipate perceptual difficulty on the part of their listeners (Buz et al., 2016). In the current study, our stimuli consist of minimal T1 and T4 pairs with different neighborhood densities. T1 and T4 share the same segmental information but with different falling tones, which are close to each other. When the target is T1, its competitors mainly include its neighbors (homophones), which share the same segmental and tonal information, as well as its minimal-paired T4 syllables (and homophone neighbors). The parallel representations of related segments and tones are expected to be activated due to tonal (incomplete) neutralization. Whether homophones and minimal pairs can be hyperarticulated or distinguished with different acoustic characteristics has been under discussion in the existing literature. For example, growing evidence shows that homophones may differ in pronunciation depending on their intended meaning. On the surface, this does not make much sense, as homophones are words that have the same phonological form but distinct meanings. However, the distinctness of the meanings may result in different pronunciations over time. There is some evidence that speakers produce homophones (e.g., bridal vs. bridle) with emotional valence appropriate to the intended meaning, leading to differences in duration and F0. Some function words with multiple meanings may also differ in duration in spontaneous speech depending on the intended meaning (e.g., Nygaard et al., 2002). The meaning of the target word matters and the homophone could be distinguished with different acoustic characteristics. Wedel et al. (2018) studied the phonetic specificity of contrastive hyper-articulation in natural speech considering minimal pairs and their results showed that cue-specific minimal pairs significantly predicted cue hyper-articulation. Therefore, considering all the competitors from meanings, minimal-paired tonal/segmental information, and listeners' expectation, homophonous targets may be hyperarticulated or realized differently with distinct acoustic characteristics in lexical production. Following these reasons, we may expect that tones with high neighborhood density should better maintain a contrast, which explains why both T1 and T4 with high neighborhood showed longer duration and a later F0 turning point. What we observed, however, is that the T1 and T4 contrast is more reliably detected in syllables with low homophone neighborhood density. Future research is needed to verify this pattern and also to tap into the effect of different ways of defining phonological connectedness (e.g., Kapatsinski, 2006) within a speaker's mental lexicon and how such networks affect speech production and sound change.

Another crucial notion in relation to the effects of frequency and neighborhood density on tonal merger, argued to condition sound change, is the functional load (FL) (Surendran and Levow, 2004; Oh et al., 2013; Wedel et al., 2013a,b; Vogel et al., 2016). FL has been suggested as an important factor in determining whether two phonemes are merged in a language (Martinet, 1933). Wedel et al. (2013a) reported the first large-scale study of the functional load hypothesis using data from sound changes in eight languages, including English, German, Dutch, and Cantonese. Results showed that the more minimal pairs defined by a phoneme pair, the less likely that phoneme pair is to have merged. Even though Cantonese was used as one of the case languages in that study, the tonal merger between T2 and T5 in Cantonese with different functional loads was not concluded in detail. Note that the way of FL calculation for a target phoneme could not be applied to our current study (due to the lack of comparable corpus). The calculation of FL in Oh et al. (2013) was followed, which focuses on the cross-language comparison of functional load for vowels, consonants, and especially tones. According to Oh et al. (2013), if a target is *pan* in T4 (pan4), its FL computation depends on *pan**, i.e., *pan* with different tones. We did a similar computation for our minimal-paired syllables of T1 and T4 and also one based on the database[3] (Neergaard and Huang, 2019). The results from both methods were consistent; syllables with high neighborhood density also have high FL, while those with low neighborhood density qualify as having low FL. The acoustic realizations of T1 and T4 with low neighborhood density (i.e., low FL), however, showed more differences than those with high neighborhood density (i.e., high FL). This is the opposite pattern from previous studies (Wedel et al., 2013b), where a higher FL was found to lead to a lower likelihood of a merger. Note that previous studies focused on phoneme contrast instead of tonal contrast, and our ways of calculating FL are not exactly the same. How to calculate FL and how exactly FL affects tonal neutralization must therefore be investigated further in order to gain a thorough understanding of FL and sound change.

In conclusion, the two falling tones in Dalian Mandarin are incompletely neutralized. The status of incomplete neutralization is relatively stable across speakers from middle-aged and young generations. Lexical frequency showed little effect on tonal neutralization, and low homophone neighborhood density helped to maintain the incompletely neutralized contrast. The effects of lexical frequency and homophone neighborhood density on tonal acoustic realization have also been investigated. It has been found that while some of the effects are predicted by existing theories of speech production and known mechanisms of sound change, the results raised more questions than they answered in regards to their effects. Further research is evidently needed to replicate these findings and to better understand the effects of lexical properties on tonal production, perception, and change.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

[3] http://dowls.site/

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of the Leiden University Centre for Linguistics, Leiden, Netherlands. The patients/participants provided their written informed consent to participate in this study.

Author contributions

YB and YC designed the experiment and revised the manuscript. YB collected, processed, analyzed the data and wrote the first draft of the manuscript with feedback from YC. Both authors approved the submitted version.

Funding

This research was supported by the Chinese Scholarship Council (CSC), Leids Universiteits Fonds (LUF) to YB, the European Research Center (to YC). YC was currently supported by the Netherlands Organization for Scientific Research (VI.C.181.040).

Acknowledgments

We thank all speakers for their participations and the No. 76 Middle School of Dalian City in Liaoning Province, China, for providing the recording booth. We also thank Wei Li for his help with data analysis.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.867353/full#supplementary-material

SUPPLEMENTARY TABLE 1
Stimuli for the production experiment.

References

Arutiunian, V., and Lopukhina, A. (2020). The effects of phonological neighborhood density in childhood word production and recognition in Russian are opposite to English. *J. Child Lang.* 47, 1244–1262. doi: 10.1017/S0305000920000112

Barton, K. (2009). *MuMIn: Multi-model inference. R package Version 0.12.2/r18.* Available online at: http://R-Forge.R-project.org/projects/mumin/

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *Lme4: Linear mixed-effects models using eigen and S4. R package Version 1.1-6.* Available online at: www.CRAN.R-project.org/package=lme4

Bauer, R. S., Kwan-hin, C., and Pak-man, C. (2003). Variation and merger of the rising tones in Hong Kong Cantonese. *Lang. Var. Change* 15, 211–225.

Baus, C., Costa, A., and Carreiras, M. (2008). Neighbourhood density and frequency effects in speech production: A case for interactivity. *Lang. Cogn. Process.* 23, 866–888.

Boersma, P., and Weenink, D. (2017). *Praat: Doing phonetics by computer [Computer program]. Version 6.0.36.* Available onine at: http://www.praat.org/ (accessed November 11, 2017).

Braver, A., and Kawahara, S. (2016). "Incomplete neutralization in Japanese monomoraic lengthening," in *Proceedings of the 2016 annual meetings on phonology* (Cambridge, MA: MIT Press).

Buz, E., Tanenhaus, M. K., and Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *J. Mem. Lang.* 89, 68–86. doi: 10.1016/j.jml.2015.12.009

Bybee, J. L. (2007). *Frequency of use and the organization of language.* Oxford: Oxford University Press.

Cai, Q., and Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One* 5:e10729. doi: 10.1371/journal.pone.0010729

Chao, Y. R. (1968). *A grammar of spoken Chinese.* Berkeley, CA: University of California Press.

Chen, H.-C., Vaid, J., and Wu, J.-T. (2009). Homophone density and phonological frequency in Chinese word recognition. *Lang. Cogn. Process.* 24, 967–982.

Chen, Q., and Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychol. Rev.* 119, 417–430. doi: 10.1037/a0027175

Chen, W.-F., Chao, P.-C., Chang, Y.-N., Hsu, C.-H., and Lee, C.-Y. (2016). Effects of orthographic consistency and homophone density on Chinese spoken word recognition. *Brain Lang.* 15, 51–62. doi: 10.1016/j.bandl.2016.04.005

Chen, Y., and Yuan, J. (2007). "A corpus study of the 3rd tone sandhi in standard chinese," in *Proceedings of the 8th annual conference of the international speech communication association (INTERSPEECH 2007)*, eds H. van Hamme and R. van Son (Baixas: International Speech Communication Association), 2749–2752.

Cheng, C., Chen, J.-Y., and Gubian, M. (2013). "Are Mandarin sandhi tone 3 and tone 2 the same or different? The results of functional data analysis," in *Proceedings of the 2013 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, Taipei.

Cheng, C., Chen, J.-Y., and Xu, Y. (2014). "An acoustic analysis of Mandarin Tone 3 sandhi elicited from an implicit priming experiment," in *Proceedings of the 2014 fourth international symposium on tonal aspects of languages*, Nijmegen.

Cheng, K. S.-K. (2017). Beginning or on-going?: 2b-3a tone change in Hong Kong Cantonese revisited. *J. Chin. Linguist.* 45, 313–343.

De Boor, C. (2001). *A practical guide to splines*. New York, NY: Springer-Verlag.

Dell, G. S., and Gordon, J. K. (2011). "Neighbors in the lexicon: Friends or foes?," in *Phonetics and phonology in language comprehension and production*, eds N. O. Schiller and A. S. Meyer (Berlin: De Gruyter Mouton), 9–38.

Dinnsen, D. A., and Charles-Luce, J. (1984). Phonological neutralization, phonetic implementation and individual differences. *J. Phon.* 12, 49–60.

Ernestus, M., and Baayen, R. H. (2006). The functionality of incomplete neutralization in Dutch: The case of past-tense formation. *Lab. Phonol.* 8, 27–49.

Fourakis, M., and Iverson, G. K. (1984). On the 'incomplete neutralization'of German final obstruents. *Phonetica* 41, 140–149.

Gahl, S. (2008). Time" and "Thyme" Are Not Homophones: The Effect of Lemma Frequency on Word Durations in Spontaneous Speech. *Language* 84, 474–496. doi: 10.1353/lan.0.0035

Gahl, S., and Strand, J. F. (2016). Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *J. Mem. Lang.* 89, 162–178. doi: 10.1016/j.jml.2015.12.006

Gao, Y. (2007). *Dalian Fangyan Shengdiao Yanjiu*. Dalian: Liaoning Normal University Press.

Garrett, A., and Johnson, K. (2013). "Phonetic bias in sound change," in *Origins of sound change: Approaches to phonologization*, ed. A. Yu (Oxford: Oxford University Press), 51–97.

Gauthier, B., Shi, R., and Xu, Y. (2007). Learning phonetic categories by tracking movements. *Cognition* 103, 80–106.

Goldinger, S. D., Luce, P. A., and Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *J. Mem. Lang.* 28, 501–518.

Gordon, J. K., and Kurczek, J. C. (2014). The ageing neighbourhood: Phonological density in naming. *Lang. Cogn. Neurosci.* 29, 326–344. doi: 10.1080/01690965.2013.837495

Gubian, M. (2011). *Functional data analysis for phonetic research. Very-large-scale phonetics research 2011*.

Gubian, M., Torreira, F., and Boves, L. (2015). Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts. *J. Phon.* 49, 16–40. doi: 10.1016/j.wocn.2014.10.001

Hay, J. B., Pierrehumbert, J. B., Walker, A. J., and LaShell, P. (2015). Tracking word frequency effects through 130 years of sound change. *Cognition* 139, 83–91. doi: 10.1016/j.cognition.2015.02.012

Kapatsinski, V. (2006). Sound similarity relations in the mental lexicon: Modeling the lexicon as a complex network. *Speech Res. Lab. Prog. Rep.* 27, 133–152.

Karimi, H., and Diaz, M. (2020). When phonological neighborhood density both facilitates and impedes: Age of acquisition and name agreement interact with phonological neighborhood during word production. *Mem. Cogn.* 48, 1061–1072. doi: 10.3758/s13421-020-01042-4

Kharlamov, V. (2014). Incomplete neutralization of the voicing contrast in word-final obstruents in Russian: Phonological, lexical, and methodological influences. *J. Phon.* 43, 47–56. doi: 10.1016/j.wocn.2014.02.002

Kong, H., and Shengyi, W. (2019). Frequency effect and neutralization of tones in Mandarin Chinese. *J. Quant. Linguist.* 26, 95–115.

Kubozono, H., and Giriko, M. (2018). *Tonal change and neutralization*, Vol. 27. Berlin: Walter de Gruyter GmbH & Co KG.

Kulikov, V. (2012). *Voicing and voice assimilation in Russian stops*. Iowa City: The University of Iowa.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13/

Labov, W. (2001). *Principles of linguistic change. Vol. 2: Social factors*. Oxford: Blackwell.

Labov, W. (2011). *Principles of linguistic change, cognitive and cultural factors*, Vol. 3. Hoboken, NJ: John Wiley & Sons.

Li, X., and Chen, Y. (2015). Representation and processing of lexical tone and tonal variants: Evidence from the mismatch negativity. *PLoS One* 10:e0143097. doi: 10.1371/journal.pone.0143097

Liang, Y. (2018). Merger and transfer: Tone variation and change of Dongguan Cantonese. *Lingua* 208, 19–30. doi: 10.1016/j.lingua.2018.03.003

Lin, Y., Yao, Y., and Luo, J. (2021). Phonetic accommodation of tone: Reversing a tone merger-in-progress via imitation. *J. Phon.* 87:101060. doi: 10.1016/j.wocn.2021.101060

Lin, Y.-J., and Hsu, Y.-Y. (2018). "Whether and how do Mandarin sandhied tone 3 and underlying tone 2 differ?" *Proceedings of the 2018 32nd Pacific Asia conference on language, information and computation*, Hong Kong. doi: 10.3389/fpsyg.2021.713665

Lindblom, B. (1990). "Explaining phonetic variation: A sketch of the H&H theory," in *Speech production and speech modelling*, eds W. J. Hardcastle and A. Marchal (Dordrecht: Springer), 403–439. doi: 10.1121/1.405815

Liu, T. (2009). *The phonology of incomplete tone merger in dalian. UC berkeley phonlab annual report*. Berkeley, CA: UC Berkeley PhonLab, 5. doi: 10.5070/P76mb5j9bm

Luce, P. A. (1986). *Neighborhoods of words in the mental Lexicon. Research on speech perception. Technical Report No. 6*. Bloomington, IN: Indiana University.

Luce, P. A., and Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear Hear.* 19, 1–36.

Luce, P. A., Goldinger, S. D., Auer, E. T., and Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Percept. Psychophys.* 62, 615–625. doi: 10.3758/BF03212113

Martinet, A. (1933). Remarques sur le système phonologique du français.*Bull. Soc. Linguist. Paris* 34, 191–202.

Matsui, M., Igarashi, Y., and Kawahara, S. (2017). Acoustic manifestation of Russian word-final devoicing in utterance-medial position. *J. Phon. Soc. Japan* 21, 1–17.

Mok, P. P., Zuo, D., and Wong, P. W. (2013). Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Lang. Var. Change* 25, 341–370.

Mousikou, P., and Rastle, K. (2015). Lexical frequency effects on articulation: A comparison of picture naming and reading aloud. *Front. Psychol.* 6:1571. doi: 10.3389/fpsyg.2015.01571

Munson, B., and Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *J. Speech Lang. Hear. Res.* 47, 1048–1058.

Neergaard, K. D., and Huang, C.-R. (2019). Constructing the Mandarin phonological network: Novel syllable inventory used to identify schematic segmentation. *Complexity* 2019:6979830.

Nicenboim, B., Roettger, T. B., and Vasishth, S. (2018). Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *J. Phonet.* 70, 39–55.

Nixon, J. S., Chen, Y., and Schiller, N. O. (2015). Multi-level processing of phonetic variants in speech production and visual word processing: Evidence from Mandarin lexical tones. *Lang. Cogn. Neurosci.* 30, 491–505. doi: 10.1080/23273798.2014.942326

Nygaard, L. C., Patel, N., and Queen, J. S. (2002). The link between prosody and meaning in the production of emotional homophones. *J. Acoust. Soc. Am.* 112, 2444–2444.

Oh, Y. M., Pellegrino, F., Coupé, C., and Marsico, E. (2013). "Cross-language comparison of functional load for vowels, consonants, and tones," in *Proceedings of the annual conference of the international speech communication association*, Brno, 3032–3036.

Phillips, B. S. (1984). Word frequency and the actuation of sound change. *Language* 320–342.

Pluymaekers, M., Ernestus, M., and Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *J. Acoust. Soc. Am.* 118, 2561–2569. doi: 10.1121/1.2011150

Politzer-Ahles, S., Connell, K., Hsu, Y.-Y., and Pan, L. (2019). "Mandarin third tone sandhi may be incompletely neutralizing in perception as well as production," *Proceedings of the 2019 19th international congress of phonetic sciences*, Melbourne.

Port, R. F., and O'Dell, M. L. (1985). Neutralization of syllable-final voicing in German. *J. Phon.* 13, 455–471.

R Core Team (2015). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Ramsay, J. O., Hooker, G., and Graves, S. (2009a). *Functional data analysis with R and matlab*. Berlin: Springer.

Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2009b). *fda: Functional data analysis. R package 2.3.8*. Available online at: https://cran.r-project.org/web/packages/fda/index.html

Roettger, T. B., Winter, B., Grawunder, S., Kirby, J., and Grice, M. (2014). Assessing incomplete neutralization of final devoicing in German. *J. Phon.* 43, 11–25. doi: 10.1016/j.wocn.2014.01.002

Rose, P. (1987). Considerations in the normalisation of the fundamental frequency of linguistic tone. *Speech Commun.* 6, 343–352. doi: 10.1016/j.cortex.2012.11.012

Scarborough, R. (2013). Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation. *J. Phon.* 41, 491–508. doi: 10.1016/j.wocn.2013.09.004

Song, X. (1963). Liaoning Yuyin Shuolue (A sketch of Liaoning Phonology). *Zhongguo Yuwen* 2, 104–114.

Surendran, D., and Levow, G.-A. (2004). "The functional load of tone in Mandarin is as high as that of vowels," in *Proceedings of the international conference speech prosody*, Nara.

Taler, V., Aaron, G. P., Steinmetz, L. G., and Pisoni, D. B. (2010). Lexical neighborhood density effects on spoken word recognition and production in healthy aging. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* 65, 551–560. doi: 10.1093/geronb/gbq039

Vitevitch, M. S., and Luce, P. A. (2016). Phonological neighborhood effects in spoken word perception and production. *Annu. Rev. Linguist.* 2, 75–94.

Vitevitch, M. S., and Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Mem. Cogn.* 31, 491–504. doi: 10.3758/bf03196091

Vogel, I., Athanasopoulou, A., and Pincus, N. (2016). Prominence, contrast, and the functional load hypothesis: An acoustic investigation. *Dimens. Phonol. Stress* 123–167.

Wang, W., Li, X., Ning, N., and Zhang, J. X. (2012). The nature of the homophone density effect: An ERP study with Chinese spoken monosyllable homophones. *Neurosci. Lett.* 516, 67–71. doi: 10.1016/j.neulet.2012.03.059

Warner, N., Good, E., Jongman, A., and Sereno, J. (2006). Orthographic vs. morphological incomplete neutralization effects. *J. Phon.* 34, 285–293. doi: 10.1016/j.wocn.2004.11.003

Warner, N., Jongman, A., Sereno, J., and Kemps, R. (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch. *J. Phon.* 32, 251–276. doi: 10.1016/S0095-4470(03)00032-9

Wedel, A., Jackson, S., and Kaplan, A. (2013a). Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Lang. Speech* 56, 395–417. doi: 10.1177/0023830913489096

Wedel, A., Kaplan, A., and Jackson, S. (2013b). High functional load inhibits phonological contrast loss: A corpus study. *Cognition* 128, 179–186. doi: 10.1016/j.cognition.2013.03.002

Wedel, A., Nelson, N., and Sharp, R. (2018). The phonetic specificity of contrastive hyperarticulation in natural speech. *J. Mem. Lang.* 100, 61–88. doi: 10.1016/j.jml.2018.01.001

Wright, R. (2004). Factors of lexical competition in vowel articulation. *Pap. Lab. Phonol.* VI, 75–87.

Yao, Y., and Sharma, B. (2017). What is in the neighborhood of a tonal syllable? Evidence from auditory lexical decision in Mandarin Chinese. *Proc. Linguist. Soc. Am.* 2, 1–14.

Yip, M. C. (2002). "Access to homophonic meanings during spoken language comprehension: Effects of context and neighborhood density," in *Proceedings of the international conference on spoken language processing*, 1665–1668.

Yu, A. C. L. (2011). "Mergers and neutralization," in *The blackwell companion to phonology*, eds M. Oostendorp, C. J. Ewen, E. Hume, and K. Rice (Oxford: Blackwell). doi: 10.1002/9781444335262.wbctp0080

Yuan, J. H., and Chen, Y. (2014). 3rd tone sandhi in standard Chinese: A corpus approach. *J. Chin. Linguist.* 42, 218–237.

Zhao, Y., and Jurafsky, D. (2007). "The effect of lexical frequency on tone production," in *Proceedings of the 2007 16th international congress of phonetic sciences*, Saarbrücken.

Frontiers in Communication

Investigates the power of communication across culture and society

A cross-disciplinary journal that advances our understanding of the global communication revolution and its relevance across social, economic and cultural spheres.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact