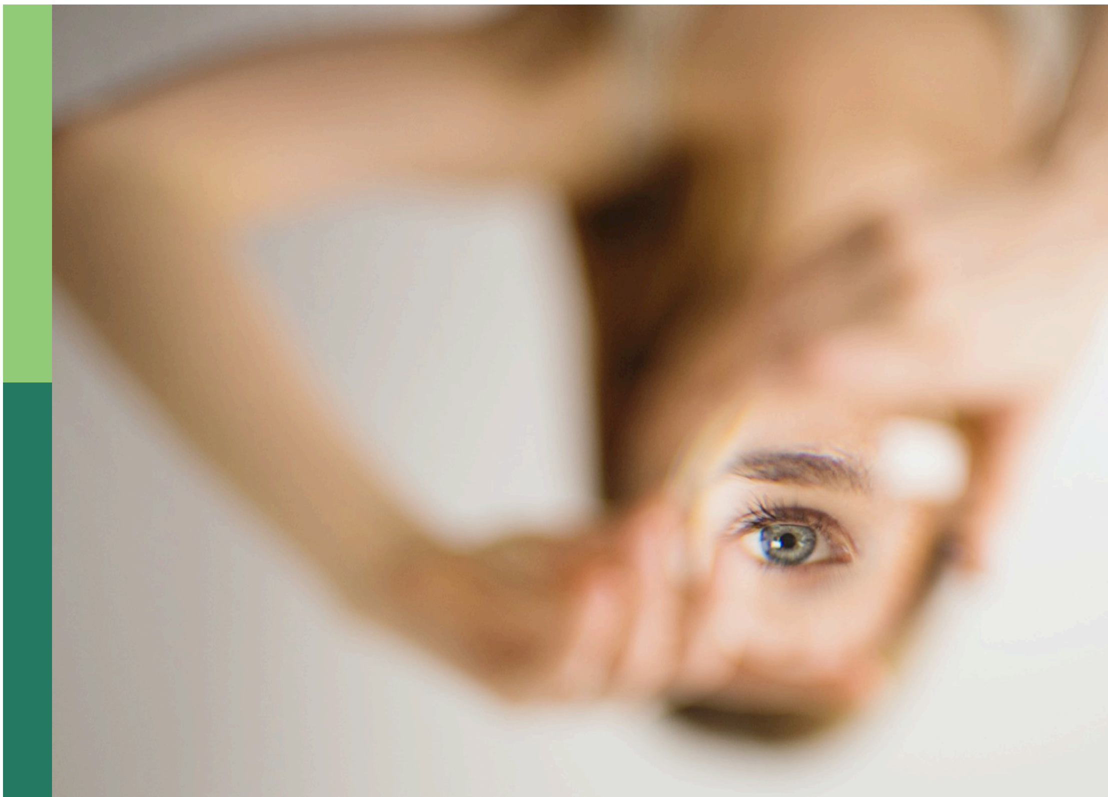
frontiers | Research Topics

Crosstalk between intonation and lexical tones: Linguistic, cognitive and neuroscience perspectives

Edited by
Hatice Zora, Annie C. Tremblay, Carlos Gussenhoven
and Fang Liu

Published in
Frontiers in Psychology
Frontiers in Neuroscience
Frontiers in Communication



ISSN 1664-8714
ISBN 978-2-8325-3301-7
DOI 10.3389/978-2-8325-3301-7

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

Crosstalk between intonation and lexical tones: Linguistic, cognitive and neuroscience perspectives

Topic editors

Hatice Zora — Max Planck Institute for Psycholinguistics, Netherlands
Annie C. Tremblay — The University of Texas at El Paso, United States
Carlos Gussenhoven — Radboud University, Netherlands
Fang Liu — University of Reading, United Kingdom

Citation

Zora, H., Tremblay, A. C., Gussenhoven, C., Liu, F., eds. (2023). *Crosstalk between intonation and lexical tones: Linguistic, cognitive and neuroscience perspectives*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3301-7

Table of contents

frontiers | Frontiers in Psychology

TYPE Editorial
PUBLISHED 08 December 2022
DOI 10.3389/fpsyg.2022.1101499

Editorial: Crosstalk between intonation and lexical tones: Linguistic, cognitive and neuroscience perspectives

Hatice Zora[1]*, Carlos Gussenhoven[2,3], Annie Tremblay[4,5] and Fang Liu[6]

[1]Department of Neurobiology of Language, Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands, [2]Department of Foreign Languages and Literatures, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, [3]Department of Language and Communication, Radboud University, Nijmegen, Netherlands, [4]Department of Linguistics, University of Kansas, Lawrence, KS, United States, [5]Department of Languages and Linguistics, The University of Texas at El Paso, El Paso, TX, United States, [6]School of Psychology and Clinical Language Sciences, University of Reading, Reading, United Kingdom

OPEN ACCESS

EDITED AND REVIEWED BY
Isabelle Peretz,
Université de Montréal, Canada

*CORRESPONDENCE
Hatice Zora
hatice.zora@mpi.nl

SPECIALTY SECTION
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Psychology

RECEIVED 17 November 2022
ACCEPTED 25 November 2022
PUBLISHED 08 December 2022

CITATION
Zora H, Gussenhoven C, Tremblay A and Liu F (2022) Editorial: Crosstalk between intonation and lexical tones: Linguistic, cognitive and neuroscience perspectives.
Front. Psychol. 13:1101499.
doi: 10.3389/fpsyg.2022.1101499

KEYWORDS

prosody, intonation, lexical tone, categorical pitch, gradient pitch

Editorial on the Research Topic
Crosstalk between intonation and lexical tones: Linguistic, cognitive and neuroscience perspectives

The interplay between categorical and continuous aspects of the speech signal remains central and yet controversial in the fields of phonetics and phonology. The division between phonological abstractions and phonetic variations has been particularly relevant to the unraveling of diverse communicative functions of pitch in the domain of prosody. Pitch influences vocal communication in two major but fundamentally different ways, and lexical and intonational tones exquisitely capture these functions. Lexical tone contrasts convey lexical meanings as well as derivational meanings at the word level and are grammatically encoded as discrete structures. Intonational tones, on the other hand, signal post-lexical meanings at the phrasal level and typically allow gradient pragmatic variations. Since categorical and gradient uses of pitch are ubiquitous and closely intertwined in their physiological and psychological processes, further research is warranted for a more detailed understanding of their structural and functional characterisations. This Research Topic addresses this matter from a wide range of perspectives, including first and second language acquisition, speech production and perception, structural and functional diversity, and working with distinct languages and experimental measures. In the following, we provide a short overview of the contributions submitted to this topic.

Behavioral investigation of tonal and intonational categoriality

Two original research articles addressed the categoriality debate of tones by expanding on existing behavioral measures. Using a Sequence Recall Task (SRT), Gussenhoven et al. tested whether a high performance in SRT indicates the lexical status of tonal information in a similar fashion to word stress. Data from speakers of non-tonal, semi-tonal, and tonal languages indicated that a tonal SRT is unlikely to discriminate between tonal and non-tonal languages due to the rich phonological nature of tone, and a number of factors affected performance, like the phonetic salience of a pitch contrast and the complexity of a language's tone system.

Rodd and Chen investigated the question whether intonation events have a categorical mental representation similar to those of segments and lexical tones by testing for a Perceptual Magnet Effect (PME). Perceived goodness and discriminability of re-synthesized productions of a Dutch rising pitch accent were evaluated by Dutch listeners. The results provided evidence for categoricalness of pitch accents, however yielding a weaker and more transient PME in pitch accents compared to the PME in lexical tones and phonemes.

Phonetic correlates of interaction between tonal functions

Phonetic correlates of interaction between lexical tone and intonation were examined by two original research articles. Zhang et al. explored how citation and neutral tones affect the perception of intonation in Mandarin. Listeners determined whether disyllabic words with citation and neutral tones formed a question or statement. The results indicated that the phonetic realizations of the neutral tone and of citation tones, realized with diverse pitch ranges and levels, determine intonation perception.

Wang et al. investigated the interaction between informative and articulatory pitch control, and specifically studied *downstep* in Mandarin and its interaction with focus and phrasing. Tonal environment, boundary strength, and focus were systematically manipulated in a production experiment. The results showed that intonation was shaped by both informative functions and articulatory constraints; downstep seems to be a phonetic effect and the interaction between focus and downstep is gradual.

Crosstalk between tone and intonation from the perspective of second language acquisition

Two original research articles studied the multifaceted function of tone from the perspective of second language (L2) acquisition. Using SRT, Kim and Tremblay examined whether listeners' use of intonational cues to a segmental contrast in the native language (L1) can facilitate the processing of an intonationally cued lexical stress contrast in the L2 by comparing Seoul Korean and French L2 learners of English. Korean listeners, who can use intonational cues to perceive segmental contrasts in their L1, outperformed French listeners, for whom segmental contrasts are not cued by intonational cues in the L1. These results suggest that cues can transfer across different types of linguistic contrasts.

Zahner-Ritter et al. used an imitation paradigm to investigate how L2 learners with a tone language as their L1 acquire pitch accents in an L2 intonation language. The authors tested the ability of Mandarin and Italian learners to imitate intonational pitch accents in German. The results indicated that experience with a tone language yields neither an advantage nor a disadvantage in the acquisition of L2 intonational pitch accents. The findings revealed instead a general cross-linguistic influence in the realization of pitch contrasts as well as improvement with higher proficiency.

Neuroscienctific evidence for structural and functional specialization of tones

Two contributions to this issue, one original research article and one hypothesis and theory article, explored the structural and functional specialization of tones using neuroscientific paradigms. Wei et al. investigated the hemispheric specialization of Chinese linguistic tones using magnetic resonance imaging and electroencephalography recordings by comparing patients with a stroke in the right temporal lobe and healthy subjects. The brain responses were lateralized in the left hemisphere for stroke patients, and in the right hemisphere for healthy individuals, indicating that the right temporal lobe is a core area for tonal processing.

Roll addressed the questionable function of the Swedish word accent contrast. Given that the two word accents do not mainly serve a lexical contrast function, they have a very low functional load from a traditional phonological contrast perspective. However, based on psychological and neurophysiological evidence and a novel analysis, the author proposes that the chief function of Swedish word accents is to predict upcoming morphological structures and facilitate lexical processing rather than being lexically distinctive.

Infant acquisition of linguistic and paralinguistic functions of pitch

The developmental course of pitch discrimination was addressed in a review article. Liu et al. discussed the lexical, intonational, and emotional functions carried by pitch, tracking how they are acquired throughout infancy. Based on a review of empirical evidence and theoretical considerations, the authors propose the Learnability Hypothesis, according to which the diverse functions of pitch are distinguished and acquired through native/environmental experiences.

Conclusion

Altogether, the articles in this Research Topic provide us with valuable information on the human disposition for stability and variability in communication, give us new insights into how (para)linguistic expressivity is built through pitch modulations, and establish new directions for future research. Key themes for further investigations include theoretical and neural network models of the interplay and integration of different tonal functions as well as a closer examination of tonal and non-tonal varieties of the same language.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

FL was supported by a European Research Council (ERC) Starting Grant (CAASD, 678733).

Acknowledgments

We would like to thank all the authors and the reviewers for their contributions to this article collection.

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.2022.845430

Intonational Cues to Segmental Contrasts in the Native Language Facilitate the Processing of Intonational Cues to Lexical Stress in the Second Language

Hyoju Kim and Annie Tremblay*

Department of Linguistics, The University of Kansas, Lawrence, KS, United States

This study examines whether second language (L2) learners' processing of an intonationally cued lexical contrast is facilitated when intonational cues signal a segmental contrast in the native language (L1). It does so by investigating Seoul Korean and French listeners' processing of intonationally cued lexical-stress contrasts in English. Neither Seoul Korean nor French has lexical stress; instead, the two languages have similar intonational systems where prominence is realized at the level of the Accentual Phrase. A critical difference between the two systems is that French has only one tonal pattern underlying the realization of the Accentual Phrase, whereas Korean has two underlying tonal patterns that depend on the laryngeal feature of the phrase-initial segment. The L and H tonal cues thus serve to distinguish segments at the lexical level in Korean but not in French; Seoul Korean listeners are thus hypothesized to outperform French listeners when processing English lexical stress realized only with (only) tonal cues (H* on the stressed syllable). Seoul Korean and French listeners completed a sequence-recall task with four-item sequences of English words that differed in intonationally cued lexical stress (experimental condition) or in word-initial segment (control condition). The results showed higher accuracy for Seoul Korean listeners than for French listeners only when processing English lexical stress, suggesting that the processing of an intonationally cued lexical contrast in the L2 is facilitated when intonational cues signal a segmental contrast in the L1. These results are interpreted within the scope of the cue-based transfer approach to L2 prosodic processing.

Keywords: speech perception, spoken word recognition, second language acquisition, Korean learners of English, French learners of English, English lexical stress

OPEN ACCESS

Edited by:
William Choi,
The University of Hong Kong,
Hong Kong SAR, China

Reviewed by:
Martin Ho Kwan Ip,
University of Pennsylvania,
United States
Seth Goss,
Emory University, United States

****Correspondence:***
Hyoju Kim
kimhj@ku.edu

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

Received: *29 December 2021*
Accepted: *22 March 2022*
Published: *14 April 2022*

Citation:
Kim H and Tremblay A (2022)
Intonational Cues to Segmental
Contrasts in the Native Language
Facilitate the Processing of
Intonational Cues to Lexical Stress in
the Second Language.
Front. Commun. 7:845430.
doi: 10.3389/fcomm.2022.845430

INTRODUCTION

In the domain of speech perception and spoken word recognition, a growing number of studies have begun to examine how second/foreign language (L2) learners perceive non-native suprasegmental contrasts (e.g., Dupoux et al., 2008; Zhang and Francis, 2010; Shport, 2015; Qin et al., 2017, 2019; Connell et al., 2018; Chan and Chang, 2019; Kim and Tremblay, 2021; Tremblay et al., 2021). One influential theoretical approach that seeks to explain the influence of the native language (L1) on the perception of L2 sound contrasts is the cue-weighting theory of speech perception (e.g., Francis et al., 2000; Francis and Nusbaum, 2002; Holt and Lotto, 2006). This theory emphasizes that speech perception is multidimensional, and acoustic cues are weighted differently not only across categories, but also across languages: Listeners from different language backgrounds hear the same acoustic stimuli differently because of the different weighting of acoustic cues in their L1. Accordingly, the cue-weighting theory stipulates that the contribution of individual acoustic cues that distinguish among phonetic categories transfers from the L1 to the L2.

For prosodic contrasts, the cue-weighting approach has focused on the functional weight of suprasegmental cues for signaling lexical information—that is, how listeners weight suprasegmental cues to lexical contrasts in the L1, and how this weighting affects the perception and processing of suprasegmental cues to prosodic contrasts in the L2. If a particular suprasegmental cue is thought to play an important role in processing lexical contrasts in the L1, it should be used to process prosodic categories in the L2; the more important a cue is in the L1, the more it is predicted to be used in the perception and processing of L2 prosodic contrasts (e.g., Qin et al., 2017; Kim and Tremblay, 2021; see also Tremblay et al., 2018). The present study further investigates how the L1 influences L2 learners' perception of prosodic contrasts, focusing on lexical stress. More specifically, this study aims to address whether listeners' use of intonational cues to a *segmental* contrast in the L1 can facilitate the processing of an intonationally cued *lexical stress* contrast in the L2.

A non-trivial body of research has found that L2 learners' perception and processing of lexical stress in English is influenced by the weighting of suprasegmental cues to lexical contrasts in the L1 (e.g., Cooper et al., 2002; Cutler et al., 2007; Zhang and Francis, 2010; Chrabaszcz et al., 2014; Lin et al., 2014; Qin et al., 2017; Connell et al., 2018; Kim and Tremblay, 2021; Tremblay et al., 2021). For instance, when processing acoustic cues to lexical stress in English, English and Mandarin listeners were reported to rely more on fundamental frequency (F0) cues than on duration or intensity cues, whereas Russian listeners relied more on duration cues than on F0 cues (Chrabaszcz et al., 2014). Russian listeners' weaker reliance on F0 and greater reliance on duration (compared to English and Mandarin listeners) were attributed to the importance of duration cues to stress contrasts in their L1. Dutch L2 learners of English also showed evidence of L1-to-L2 cue-weighting transfer: Dutch L2 learners of English were found to put greater weight on suprasegmental cues to process lexical stress compared to native English listeners, a finding that was attributed to the lower weight of vowel quality cues to lexical stress in Dutch compared to English (e.g., Cooper et al., 2002; Cutler et al., 2007; Tremblay et al., 2021). These results suggest that the weighting of suprasegmental cues to lexical stress transfers from the L1 to the L2.

Some studies have also provided evidence that listeners can transfer the use of suprasegmental cues to lexical contrasts from one type of prosodic contrast in the L1 to another in the L2 (e.g., Braun et al., 2014: perception of lexical tones by German, French, and Japanese listeners; Choi et al., 2019; Choi, 2022: perception of English lexical stress by Cantonese L2 learners of English; Kim and Tremblay, 2021: perception of English lexical stress by Gyeongsang Korean and Seoul Korean L2 learners of English; Tremblay et al., 2018: perception of intonational cues to French word-final boundaries by English and Dutch L2 learners of French; Wiener and Goss, 2019: perception of Japanese pitch accent by naïve Mandarin listeners and English L2 learners of Japanese). These studies provide preliminary evidence that those suprasegmental cues that serve important lexical functions in the L1 can be used to process different prosodic categories in the L2. To illustrate, Kim and Tremblay (2021) investigated whether Korean-speaking L2 learners of English would transfer the use of suprasegmental cues from the processing of lexical pitch accents in Korean to the processing of lexical stress in English. Gyeongsang Korean is a tonal dialect of Korean that does not have lexical stress but has lexical pitch accents, whereas Seoul Korean has neither. Gyeongsang Korean listeners were hypothesized to be more sensitive to F0 as a cue to lexical contrasts compared to Seoul Korean listeners. The results showed that Gyeongsang Korean L2 learners of English had an advantage over Seoul Korean L2 learners of English when processing intonationally cued lexical stress in English words, with duration and intensity cues not further enhancing perception in either group. Gyeongsang Korean listeners' ability to process English lexical stress was attributed to their use of F0 cues from the processing of lexical pitch accents in their L1 dialect, suggesting that suprasegmental cues that are important for distinguishing words in the L1 (i.e., F0) are used to process words in the L2. These results suggest that L2 learners whose L1 dialect does not have lexical stress can transfer the use of a suprasegmental cue (here, F0) from a different prosodic category (e.g., lexical pitch accent contrasts) to lexical stress in the L2.

An important question that arises from this research is the scope of cue weighting transfer. The cue-weighting theory of speech perception proposes that the underlying mechanism for learning speech categories or contrasts in both the L1 and the L2 is listeners' selective attention to specific acoustic dimensions, assuming that a phonetic category consists of a multidimensional structure where each dimension corresponds to a feature of the phonetic category (e.g., Iverson and Kuhl, 1995; Kuhl and Iverson, 1995; Francis and Nusbaum, 2002; Francis et al., 2008). Accordingly, the cue-based transfer approach stipulates that L2 learners' ability to attend to a particular cue in the L2 and associate it with a function that differs from that in the L1 depends on how much weight the cue has in the L1. Thus, one prediction of the theory is that the weight of a cue in the L1 will determine whether L2 learners would rely on the cue in the L2, regardless of its actual function in the L2.

Tremblay et al. (2018) provided empirical evidence that acoustic cues that serve one function in the L1 can indeed be reallocated to a different function in the L2. The authors tested whether English and Dutch L2 learners of French would differ in their use of F0 cues to word-final boundaries in the segmentation of French speech. Both English and Dutch have lexical stress contrasts, but the functional weight of F0 cues for signaling lexical identity is higher in Dutch than in English due to the lower weight of vowel quality cues to lexical stress in Dutch. Thus, it was hypothesized that Dutch listeners would show greater reliance on F0 cues than English listeners when locating word-final boundaries in French. In other words, Dutch listeners were predicted to transfer the higher functional weight of F0 cues from the processing of lexical information in the L1 (i.e., lexical stress contrasts) to the detection of word-final boundaries in the segmentation of French speech. The results of an eye-tracking experiment revealed that Dutch listeners showed greater reliance on F0 cues than English listeners when locating word-final boundaries in French. This suggests that acoustic cues that serve one function in the L1 (i.e., signaling lexical information) can be transferred to a different function in the L2 (i.e., signaling word boundaries).

Further probing the question of L1-based cue transfer, one may also ask whether cues can transfer across different *types* of linguistic contrasts (e.g., from intonationally cued segmental contrasts to intonationally cued lexical stress contrasts). If cues that have a similar function (e.g., to signal lexical information) can transfer from one prosodic contrast to another (e.g., Kim and Tremblay, 2021: from Gyeongsang Korean lexical pitch accents to English lexical stress; Qin et al., 2017: from Mandarin lexical tones to English lexical stress), then we should also expect cues to transfer from the perception of intonationally cued *segmental* contrasts to the perception of intonationally cued *lexical stress* contrasts, as these two types of contrast serve a similar function—to signal lexical information. In other words, from a cue-weighting perspective, there is no reason not to expect L1-based cue transfer to occur. Some research has provided evidence for the transfer of F0 cues across different types of contrasts. Francis and Nusbaum (2002), for example, showed that English listeners can learn to use F0 as a cue to the Korean stop contrast after short-term identification training in a laboratory environment. This could be taken as evidence for the transfer of F0 cues from intonationally cued lexical stress contrasts to segmental contrasts, as F0 plays some role in the perception of lexical stress in English. However, since F0 also covaries with VOT in English stops, it remains unclear whether English listeners' ability to process F0 cues in L2 segmental contrasts was caused by their use of F0 cues to segmental contrasts or by their use of F0 cues to intonationally cued lexical stress contrasts (or both). The present study will shed further light on this question by investigating whether F0 cues can transfer from the perception of intonationally cued segmental (i.e., stop) contrasts in the L1 to the perception of intonationally cued lexical stress contrasts in the L2. To do so, two groups of L2 learners—Seoul Korean and French L2 learners of English—will be compared. By addressing this question, the present study will clarify the scope of cue-weighting transfer in L2 prosodic processing.

Korean has a three-way laryngeal stop contrast, which is typologically rare. Prior studies have described Korean as having a short Voice Onset Time (VOT) and high F0 for fortis stops, an intermediate VOT and low F0 for lenis stops, and a long VOT and high F0 for aspirated stops in word-initial position (e.g., Lisker and Abramson, 1964; Cho, 1996). It has also been documented that, in Seoul Korean, the VOT of lenis and aspirated stops has gradually merged over time, with the contrast now depending on the F0 of the following vowel (e.g., Silva, 2006; Kang and Guion, 2008; Kang, 2014). The realization of stops in Seoul Korean is dependent on the prosodic position in which these stops occur, such as the Accentual Phrase (Silva, 2006). More specifically, in trisyllabic Korean words, a low F0 (L) and upward F0 trajectory are observed if the word-initial segment is a lenis stop, and a high F0 (H) and downward F0 trajectory is observed if the initial segment is a non-lenis stop (i.e., fortis and aspirated stops). In other words, the consonant-induced F0 distinction in Korean extends far beyond the initial portion of the immediately following vowel (Jun, 1996; Silva, 2006). Korean listeners have also been found to use F0 cues in the perception of stop contrasts: Lee et al. (2013) and Schertz et al. (2015) demonstrated that Seoul Korean listeners used F0 as a primary cue and VOT as a secondary cue to perceive the lenis-aspirated stop contrast and both F0 and VOT as primary cues to perceive the fortis-lenis stop contrast. Thus, F0 plays an important role in distinguishing stop contrasts for Seoul Korean listeners[1].

What remains unclear is whether Seoul Korean listeners' reliance on F0 for processing segmental distinctions in the L1 can contribute to enhancing their processing of intonationally cued lexical stress contrasts in the L2. Korean listeners have been shown to have more difficulty than Mandarin listeners in the processing of English lexical stress (Lin et al., 2014). It is therefore unlikely that Korean listeners' use of F0 cues to stop contrasts in the L1 would completely overcome any difficulty they may have in the processing of lexical stress in the L2. However, since F0 cues have an extremely high functional weight in Mandarin due to the importance of lexical tones, the Korean-Mandarin group comparison is not one that can determine whether intonational cues to segmental contrasts in the L1 can provide at least some help in the perception of intonationally cued lexical stress in the L2.

To answer this question, the present study compares Seoul Korean L2 learners of English and proficiency-matched Metropolitan French L2 learners of English in the processing of intonationally cued lexical stress. An important intonational unit in Korean is the Accentual Phrase (AP): If the AP has four or more syllables and its initial segment has the feature of [−stiff vocal folds] (e.g., lenis stops or sonorants), the AP has an LHLH tonal pattern; if the phrase-initial segment is [+stiff vocal folds] (e.g., aspirated and fortis stops, coronal fricatives, and /h/), the AP has an HHLH tonal pattern (for details, see Jun, 1998, 2000). The cue to the segmental contrast that is hypothesized to transfer from the L1 to the L2 is associated with a tone that is triggered by the type of segment (i.e., L for lenis stops or H for aspirated and fortis stops). It is predicted that this intonational cue to the segmental contrast may help Seoul Korean listeners when processing intonationally cued English lexical stress.

Despite the typological differences between the two languages, French has a very similar prosodic system to that of Korean, with the AP also being an important intonational unit in French. If the AP has four or more syllables, it has a LHiLH* tonal pattern, where Hi indicates the secondary or initial phrasal prominence

[1] Most perception studies used only one place of articulation (usually bilabial stops) because they assumed that a general cue-weighting pattern would be consistent across places of articulation. In production, however, Broersma (2010) showed the cue-weighting to be generalizable across all three places of articulation.

and H* indicates the primary or final pitch accents (for details, see Jun and Fougeron, 2000, 2002). Crucially, French has only one underlying tonal pattern (i.e., LHiLH*), with the pattern not varying on the basis of the phrase-initial segment. This difference in the underlying tonal patterns of the two languages allows us to investigate whether Seoul Korean L2 learners of English can transfer the use of F0 cues from the processing of segmental contrasts in the L1 to the processing of intonationally cued lexical stress in the L2.

Previous research conducted by Dupoux and colleagues (Dupoux et al., 2001, 2008, 2010) has shown that French monolinguals and French L2 learners of Spanish performed much more poorly than Spanish monolinguals on tasks that required them to process phonetically variable stress under a memory load (sequence recall task), a difficulty that the researchers termed stress "deafness" and attributed to the absence of lexical stress in French. In principle, the stress processing difficulties found in French listeners should be replicated in Seoul Korean listeners, given that Seoul Korean, like French, does not have lexically contrastive stress.

However, and crucially, the cue-based transfer approach would additionally predict that Seoul Korean listeners would outperform French listeners in the processing of intonationally cued English lexical stress because Seoul Korean listeners would transfer the use of F0 cues from the processing of the laryngeal segmental contrasts in the L1 to the processing of lexical stress in the L2. One may ask whether French listeners could, to some degree, transfer the use of F0 as a secondary cue to stop contrasts from the perception of French stops to the perception of intonationally cued English lexical stress. This is unlikely, as Serniclaes (1987) reported that VOT is the dominant cue to the voicing contrast in French; other cues (e.g., F1 onset frequency, duration of formant transition, initial F0, F0 contour, rise time, and burst energy) come into play only when VOT is ambiguous (in perception as well as in production; see Kirby and Ladd, 2015). In other words, VOT provides the major perceptual cue to stop contrasts in French, all the other cues being secondary (e.g., Serniclaes, 1987, cited in Saerens et al., 1989). Since F0 is not a primary cue to stop contrasts in French, but it is for the lenis-aspirated and the fortis-lenis contrast in Seoul Korean, the cue-based transfer approach would predict French listeners to have more difficulty processing intonationally cued English lexical stress compared to Seoul Korean listeners[2].

This hypothesis was tested using a sequence-recall task similar to those used in previous research on the perception of lexical stress contrasts (e.g., Dupoux et al., 2001, 2008, 2010; Lin et al., 2014; Qin et al., 2017; Kim and Tremblay, 2021). In the association phase of a sequence-recall task, participants are trained to associate words that differ in stress with different keys of a computer keyboard. Then, in the testing phase, participants hear auditory word sequences and attempt to recall them by using the same keys in the corresponding order. The auditory words in each sequence are produced by different talkers and thus are acoustically variable. Because of the short-term memory load that this task imposes, listeners must be able to process lexical stress in a phonologically abstract way in order to recall the sequences accurately; processing lexical stress in an acoustic way would impose too high of a demand on short-term memory for the listener to be able to recall the sequence accurately. Because listeners may vary in their short-term memory capacity, a control condition in which listeners hear a phonological contrast that exists in the L1 is also used as the baseline. Hence, this type of task provides a robust method for investigating the phonological processing of lexical stress, and it discourages response strategies given the memory load it imposes on listeners.

The experiment used in the present study manipulated auditory stimuli in which the lexical stress contrast was conveyed only by F0 cues (with duration and intensity being neutralized), as Seoul Korean and French listeners are expected to differ only in the use of F0 cues. Since this study focuses on the processing of intonationally cued lexical stress, the stimuli did not involve vowel reduction cues, which play an important role in native English listeners' perception of lexical stress in English words (e.g., Cooper et al., 2002; Cutler et al., 2007; Zhang and Francis, 2010; Chrabaszcz et al., 2014). Under a cue-weighting transfer view, it is only in the use of F0 cues to lexical stress that Korean and French listeners are expected to show disparities.

Additionally, the current study controlled for Seoul Korean participants' knowledge of other tonal dialects and languages based on a quantitative assessment of their language experience, unlike previous studies that investigated Korean listeners' perception of English lexical stress (e.g., Lin et al., 2014)[3]. In doing so, we can assure that any potential advantage from Seoul Korean listeners in the perception of English lexical stress does not stem from their experience with tonal dialects of Korean (e.g., Gyeongsang Korean) or other tonal languages (e.g., Mandarin, Japanese).

Unlike previous studies, the present experiments used real English words to ensure that participants processed the words in the language in which they were intended (i.e., English). If participants hear non-words, we cannot determine with certainty whether they processed the non-words in English

[2] The approach adopted here is that listeners attend to acoustic cues that have a high functional weight (e.g., that serve to distinguish words) in the L1. This approach assumes that whether or not the L1 has lexical stress does not have much bearing on whether listeners can perceive an intonationally cued lexical stress contrast in the L2, as long as the cue that signals stress is important for distinguishing words in the L1. For example, Taiwan Mandarin listeners, who do not have lexical stress in their L1, have no difficulty perceiving English lexical stress when it is cued with F0, a cue that is very important to the perception of Mandarin lexical tones (e.g., Qin et al., 2017). Some researchers have made phonologically driven predictions for the processing of lexical stress—that is, predictions that are contingent on the higher-level patterning of stress in the L1 (e.g., Peperkamp and Dupoux, 2002; Peperkamp et al., 2010). While the predictions of the two different approaches may coincide in some cases, we believe it is not necessary to make reference to the phonological patterning of stress in the L1 to predict what listeners do when processing lexical stress in the L2.

[3] Altmann (2006)'s study on the perception of English stress tested a variety of L1 groups, including "Seoul" Korean listeners. However, due to the lack of detailed information about the participants' language background, it remains unclear whether the Korean participants she tested had any knowledge of tonal dialects of Korean (e.g., Gyeongsang Korean). Controlling for Korean listeners' Korean dialect is critical because experience with other Korean dialects may change how listeners weight F0 cue when processing the Korean stop contrast (for more detail, see Lee and Jongman, 2019; Kim and Jongman, 2021).

mode, thus processing suprasegmental cues as if they belonged to intonationally cued English lexical stress. Using real English words solves this issue.

METHOD

Participants

The experiment targeted 50 Seoul Korean L2 learners of English, 50 French L2 learners of English, and 50 native English listeners as a control group; Korean participants who did not speak Gyeongsang Korean and were not regularly exposed to it were recruited from universities in Seoul; French participants were recruited from universities in Aix-en-Provence, Grenoble, and Paris; and English participants were recruited from a Midwestern American university. The participants did not have speech or hearing impairments or learning disabilities. The participants were tested *via* a web-based survey design software, Qualtrics (Qualtrics LLC, 2020). Each participant completed three tasks: (1) a language background questionnaire; (2) a sequence-recall task with English stimuli; and (3) a lexical decision task to assess their lexical proficiency in English (LexTALE; Lemhöfer and Broersma, 2012). The complete session took approximately 45 min. Korean and French participants received financial compensation for participating in the experiment. English participants received extra credits for one of the introductory courses in Linguistics.

The English proficiency of Seoul Korean and French L2 learners of English was controlled based on the information obtained in the language background questionnaire and their LexTALE scores. After specific exclusion criteria were applied (see Section Data Analysis for detail), the present study included 42 Seoul Korean L2 learners of English (25 female), 35 French L2 learners of English (15 female), and 32 native English listeners (19 female). **Table 1** summarizes the relevant language background information for all three groups and the English proficiency data for the Seoul Korean and French L2 learners of English. Statistical analyses revealed that the Seoul Korean and French listeners did not show a significant difference in any of the variables reported in **Table 1** except for their self-reported percentage of daily English usage [$t_{(75)} = -6$, $p < 0.001$]. Since the significant difference is in a direction that is not confounded with the predictions, it is not problematic for the interpretation of the results.

TABLE 1 | Participants' language background information.

	Korean	French	English
Age (years)	26.2 (5.1)	25.8 (5)	20.4 (3.5)
LexTALE (/100)	70.6 (12)	74.7 (17)	92 (5.6)
AOE (years)	9.9 (2.3)	12.3 (4.4)	—
LOR (months)	1.5 (3.4)	8.9 (18)	—
LOE (years)	13.8 (4.7)	10.4 (4.3)	—
Daily English usage (%)	10 (9)	34.4 (24.1)	—
Self-rated English proficiency score (1–5)	2.6 (0.8)	3.1 (0.8)	—
Self-rated English accent score (1–10)	5.3 (2.4)	6.1 (1.8)	—

Values are means (standard deviations). AOE, Age of first exposure to English; LOR, Length of residence in an English-speaking country; LOE, Length of English education.

Materials

The lexical items used in this study were identical to those of Kim and Tremblay (2021). The lexical stress contrast stimuli for the experimental condition were a minimal pair of English words that differed in their stress pattern (*TRUSty* vs. *trusTEE* for the practice phase; *OFFset* vs. *offSET* for the test phase)[4]. These stimuli did not involve vowel reduction cues to lexical stress, the contrast being signaled only by suprasegmental cues. The segmental contrast stimuli for the control condition were minimal pair of English words that differed only in the place of articulation of the word-initial segment (*taller* vs. *caller* for the practice phase; *table* vs. *cable* for the test phase). Since English, Korean, and French all have a contrast between the aspirated alveolar stop and the aspirated velar stop, all listeners should be able to perceive this segmental difference. This control condition thus also serves as a test to determine whether participants attended to the task.

The lexical items were recorded by one female and one male native speaker of American English to increase phonetic variability, as was done in previous studies (e.g., Dupoux et al., 2001, 2008, 2010; Lin et al., 2014; Kim and Tremblay, 2021). The speakers recorded each lexical item five times in the carrier sentence, *Say* ____ *again*, using a microphone (Electro Voice N/D 767a) and a digital recorder (Marantz PMD 671) at a sampling rate of 44.1 kHz. From the five repetitions of each stress pattern from each speaker, three best tokens were selected, yielding a total of 24 tokens: 12 experimental tokens (3 tokens × 2 words × 2 speakers) and 12 control tokens (3 tokens × 2 words × 2 speakers). Since it is only in the use of F0 cues to lexical stress that Seoul Korean and French listeners were expected to differ, duration and intensity cues to lexical stress were neutralized. The intensity of all words was first normalized to 70 dB based on the root-mean-square (RMS) amplitude. Each syllable was then manipulated such that its duration and intensity would be that of the average across the two stress patterns. All manipulation procedures were implemented in Praat (Boersma and Weenink, 2019). We used the Pitch Synchronous Overlap and Add (PSOLA) function for duration manipulation, and the Multiply function for intensity. Acoustic measurements of the manipulated stimuli are summarized in **Table 2**.

Procedure

The task was built using web-based survey design software, Qualtrics (Qualtrics LLC, 2020). The sequence-recall task

[4] Since most of the previous studies that conducted sequence-recall experiments used nonword pairs as auditory stimuli, we did not test listeners' knowledge of the English word pairs we used; listeners do not need to know the words in the sequence to be able to classify them as stressed on the first or second syllable based on the auditory information they hear. Note also that we used a single minimal pair in each condition to follow the method developed by Dupoux and colleagues for the sequence recall task. Although this may limit the generalizability of the results, we believe the task would be more difficult (possibly too difficult) if listeners had to categorize different words in different trials.

TABLE 2 | Mean **F0, duration, and intensity** of the English critical stimuli (standard deviation).

	Word-initial stress			Word-final stress		
F0 (Hz)	**σ1**	**σ2**	**Ratio**	**σ1**	**σ2**	**Ratio**
Offset (M)	139 (9.4)	89 (5.3)	1.6	115 (2.4)	160 (12.1)	0.7
Offset (F)	226 (8.7)	181 (3.1)	1.3	173 (3.8)	229 (7.5)	0.8
Duration (ms)	**σ1**	**σ2**	**Ratio**	**σ1**	**σ2**	**Ratio**
Offset (M)	269 (25.2)	476 (11.8)	0.6	262 (2.9)	467 (12.7)	0.6
Offset (F)	314 (1.7)	484 (8.7)	0.6	291 (1.4)	513 (2.9)	0.6
Intensity (dB)	**σ1**	**σ2**	**Ratio**	**σ1**	**σ2**	**Ratio**
Offset (M)	73 (0.4)	72 (0.2)	1.0	73 (0.2)	72 (0.0)	1.0
Offset (F)	70 (0.2)	69 (0.1)	1.0	70 (0.1)	69 (0.0)	1.0

σ1 and σ2 indicate the first and second syllable, respectively; the letters "M" and "F" in parentheses stand for male and female speakers, respectively.

consisted of two tiers. The first tier tested listeners' processing of lexical stress contrasts, and the second tier tested listeners' processing of phonemic contrasts. Each tier consisted of four blocks. The first two blocks of the tier formed the practice phase, and the last two blocks of the tier formed the testing phase. **Figure 1** illustrates the structure of the experiment.

The practice phase involved an association block and a practice block, each with feedback. In the association block of the practice phase, the participants were trained to associate 1 and 2 on a computer keyboard with the two English words that differed in their stress (*TRUSty* vs. *trusTEE*) or in their initial segment (*taller* vs. *caller*). For each contrast type, there were a total of ten association trials (2 stimuli × 5 repetitions). On each trial, immediate feedback was provided to the participants as to whether they associated the stimulus with the correct button.

In the practice block of the practice phase, participants were asked to recall sequences of the stimuli they learned to associate with 1 and 2 in the association block. For example, if a participant heard the sequence of *TRUSty—trusTEE—trustee—TRUSty,* they would need to enter [1221] as a response. The segmental contrast condition had the same logic. Six different orders of four-item sequences (i.e., [1122], [2211], [1212], [2121], [2112], [1221]) were used for the practice block of each contrast type. There were thus six trials (1 repetition × 6 orders) for each contrast type, and the participants received immediate feedback on the accuracy of their responses (i.e., correct or incorrect). Within a four-item sequence, each item was separated by an inter-stimulus interval of 50 ms, as in previous studies (e.g., Dupoux et al., 2001, 2008; Qin et al., 2017; Kim and Tremblay, 2021). The last item in the sequence was followed by a pure tone to prevent participants from using echoic memory to recall the sequences. On each trial, participants had 5 s to respond after they heard the sequence. After 5 s, the next trial automatically began with an inter-trial interval of 1,500 ms. **Figure 2** illustrates the composition of a trial for the four-item sequence-recall task.

The testing phase contained an association block with feedback and a test block without feedback. In the association block, participants were trained to associate 1 and 2 on a keyboard with two other English words that differed in their lexical stress or in their initial segment, this time with the stimuli *OFFset* vs. *offSET* for the stress contrast and *table* vs. *cable* for the segmental contrast. There was a total of ten association trials (2 stimuli × 5 repetitions) for each contrast type. In the association block, the participants received immediate feedback on their accuracy in each trial. In the test block, participants were asked to recall the four-item sequences of English words that differed in stress or segment. Ten different token orders (i.e., [1121], [1122], [1211], [1212], [1221], [2112], [2121], [2122], [2211], [2212]) were used in each of the test blocks. Thus, for each contrast type, the test block included 30 trials (3 tokens × 10 orders). Participants did not receive feedback on the accuracy of their responses in this block. The trials within a block were randomized across participants.

Data Analysis

The data of participants who did not reach a 75% (22/30) accuracy rate on the segmental block (5 Korean listeners, 15 French listeners, 18 English listeners) were excluded from the analyses, under the assumption that they likely did not focus on the task (which is more likely to happen in web-based experiments). Among the remaining Korean participants, 3 participants who self-reported being able to speak Gyeongsang Korean fluently (despite our attempt not to recruit such participants) were also excluded from the analyses. These filtering processes left 42 Seoul Korean, 35 French, and 32 English listeners in the data analyses.

Mixed-effects logistic regression models were conducted on the participants' sequence-recall accuracy. The data were fitted into the model using the *glmer* function of the *lme4* package (Bates et al., 2015) of the statistical software R and R studio (R Development Core Team, 2019). The model focused on the participants' accuracy in the segmental and lexical stress contrast conditions by participants' L1. The dependent variable of the model was ACCURACY, which is a binary response of correct or incorrect. The participants' response on each trial was coded as correct if they correctly recalled the complete sequence of four items and as incorrect if the sequence was incorrectly recalled. The fixed effects in the model were L1 (English vs. Seoul Korean vs. French), CONTRAST TYPE (segmental contrast vs. stress contrast, baseline: stress contrast), and their interactions.

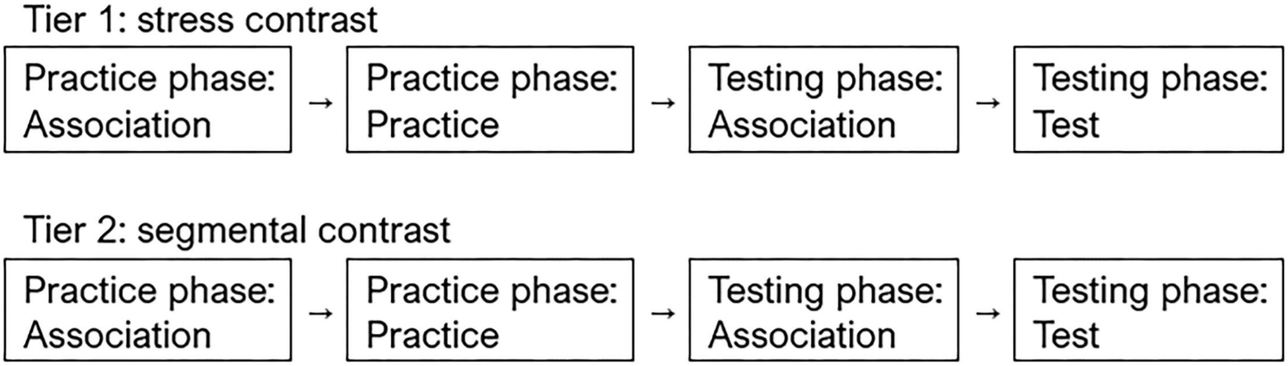

Tier 1: stress contrast
Practice phase: Association
Practice phase: Practice
Testing phase: Association
Testing phase: Test
Tier 2: segmental contrast
Practice phase: Association
Practice phase: Practice
Testing phase: Association
Testing phase: Test

FIGURE 1 | Structure of the experiment.

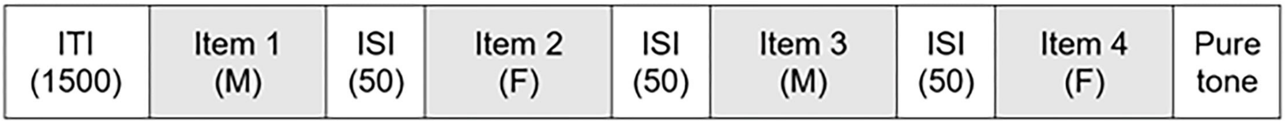


FIGURE 2 | Composition of each trial of the four-item sequence-recall task. The numbers in parentheses are the duration of each interval in milliseconds; the letter M and F in the parentheses stand for male and female speakers, respectively.

Since the effect of L1 has three levels, the model was run once with Seoul Korean listeners as a baseline and once with English listeners as a baseline. Random intercepts included participants, test items, and sequence orders. The best model was automatically selected using the backward fitting function of the *LMERConvenienceFunctions* package (Tremblay and Ransijn, 2015).

If Seoul Korean listeners transfer the use of F0 cues from the processing of intonational cues to segmental contrasts in Seoul Korean to the processing of intonationally cued lexical stress contrasts in English, they should be more accurate than French listeners at processing the lexical stress contrasts in English. If this prediction is correct, we should find a significant interaction between L1 (French) and CONTRAST TYPE in the model with Korean listeners' accuracy on the stress contrasts as a baseline.

RESULTS

Listeners' accuracy on the sequence-recall task is provided in **Figure 3**, and **Table 3** summarizes the fixed-effect coefficients in the mixed-effects logistic regression model.

The model with Korean listeners' accuracy in the stress contrast condition as a baseline (**Table 3A**) revealed that Seoul Korean listeners outperformed French listeners but not English listeners in the stress contrast condition, as evidenced by the significant simple effect of L1 for French listeners but not for English listeners. Seoul Korean listeners' accuracy in the segmental contrast condition was significantly higher than that in the stress contrast condition, as evidenced by the simple effect of CONTRAST TYPE. Additionally, there was a significant interaction between L1 and CONTRAST TYPE for the French group but not for the English group, indicating that Seoul Korean listeners' accuracy showed a smaller difference between the segmental contrast and the stress contrast compared to French listeners. The simple effect of L1 and the interaction effect confirm that French listeners showed greater difficulty processing English lexical stress contrasts than Seoul Korean listeners. Because the segmental condition served as control condition and because the results yielded a significant interaction between L1 and CONTRAST TYPE, the effect of L1 on the stress contrast condition cannot be attributed to short-term memory capacity differences between the two L1 groups.

The model with English listeners' accuracy in the stress contrast condition as a baseline (**Table 3B**) showed that English listeners outperformed French listeners in the stress contrast condition, as evidenced by the significant simple effect of L1. The simple effect of CONTRAST TYPE indicates that English listeners' accuracy in the segmental contrast condition was significantly higher than that in the stress contrast condition. There was a significant interaction effect between L1 and CONTRAST TYPE for the French group, meaning that French listeners' accuracy showed a greater difference between the segmental contrast and the stress contrast compared to English listeners. The simple effect of L1 and the interaction effect indicate that French listeners showed greater difficulty processing lexical stress contrasts than English listeners, a result that again cannot be attributed to short-term memory capacity differences between the two L1 groups.

Additional *post-hoc* analyses showed that Seoul Korean and French listeners' accuracy in the stress contrast condition was

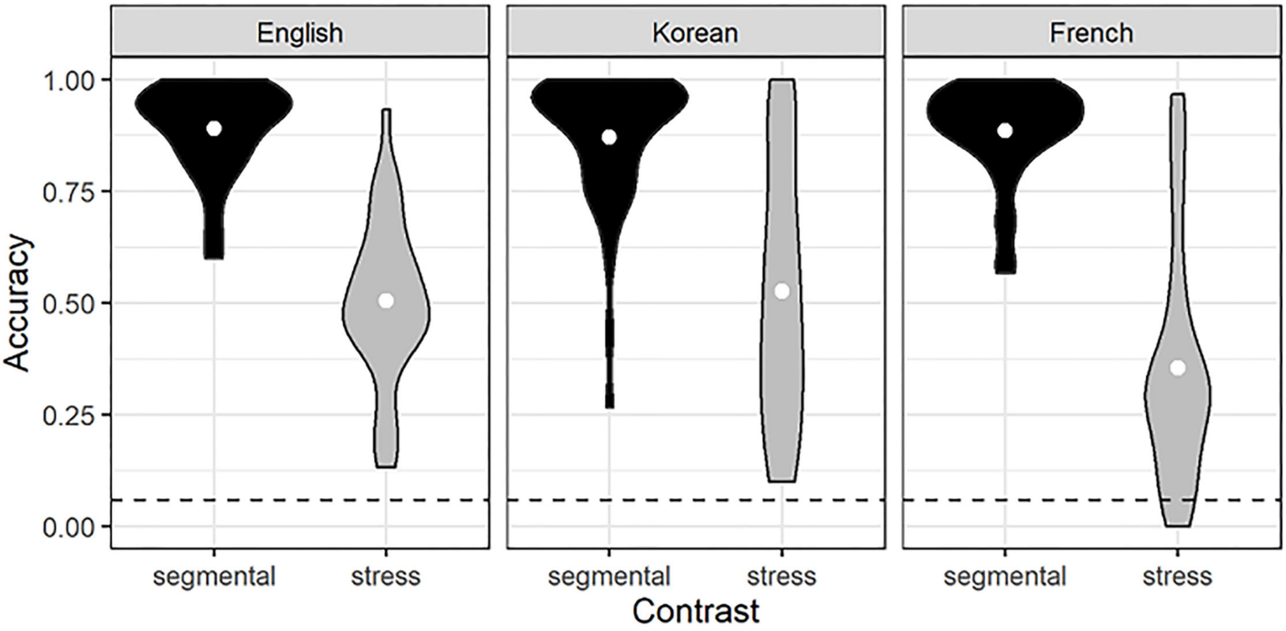


FIGURE 3 | Listeners' accuracy on the sequence-recall task. The length of the violins represents the range of values; the width of the violins at a given *y* value represents the point density at that value; the white dots represent the mean; the dashed line represents chance-level performance (1 hit/16 possible sequence orders = 0.06).

TABLE 3 | Summary of fixed-effect coefficients in the mixed-effects logistic regression model on listeners' accuracy on the sequence-recall task.

Fixed effects	Est.	SE	*z*-value	*Pr*(>|z|)
(A) Model with Korean listeners' accuracy in the stress contrast condition as baseline				
(Intercept)	0.203	0.224	<|1|	0.365
L1 (English)	−0.183	0.279	<|1|	0.512
L1 (French)	−0.934	0.276	−3.388	<0.001
Contrast type (segmental)	2.229	0.126	17.69	<0.001
L1 (English) × Contrast type (segmental)	0.113	0.175	<|1|	0.519
L1 (French) × Contrast type (segmental)	1.007	0.179	5.631	<0.001
(B) Model with English listeners' accuracy in the stress contrast condition as baseline				
(Intercept)	0.019	0.244	<|1|	0.937
L1 (Korean)	0.183	0.279	<|1|	0.512
L1 (French)	−0.750	0.292	−2.568	<0.05
Contrast type (segmental)	2.342	0.140	16.68	<0.001
L1 (Korean) × Contrast type (segmental)	−0.113	0.175	<|1|	0.519
L1 (French) × Contrast type (segmental)	0.895	0.189	4.712	<0.001

not correlated with demographical factors such as L2 learners' self-rated English proficiency score (Korean: $r = 0.24$, $p = 0.13$; French: $r = 0.09$, $p = 0.58$), self-rated English accent score (Korean: $r = 0.25$, $p = 0.11$; French: $r = 0.27$, $p = 0.12$), LexTALE score (Korean: $r = 0.21$, p =0.18; French: $r = 0.12$, $p = 0.49$), length of residence in an English-speaking country (Korean: $r = -0.17$, $p = 0.3$; French: $r = -0.11$, $p = 0.54$), length of English education (Korean: $r = 0.16$, $p = 0.32$; French: $r = 0.29$, $p = 0.09$), or age of first exposure to English (Korean: $r = 0.11$, $p = 0.5$; French: $r = 0.2$, $p = 0.24$). Thus, L2 learners' performance in the stress contrast condition could not be attributed to their proficiency in or familiarity with English. Additionally, Seoul Korean listeners did not show a significant correlation between their accuracy on the task and their degree of exposure to Gyeongsang Korean ($r = 0.001$, $p = 0.99$) or between their accuracy and their self-rated Gyeongsang Korean speaking score ($r = 0.14$, $p = 0.37$), suggesting that Seoul Korean listeners' performance in the stress contrast condition is not related to their knowledge of the tonal dialect of Korean.

DISCUSSION

This study investigated whether listeners transfer the use of intonational cues from the perception of segmental contrasts in the L1 to the perception of intonationally cued lexical stress in the L2. The results showed that Seoul Korean L2 learners of English had an advantage over proficiency-matched French L2 learners of English when processing intonationally cued lexical stress in English words. These results provide support for the hypothesis that L2 learners whose L1 uses a suprasegmental cue (F0) to distinguish segmental features can transfer the use of that cue from one contrast type (i.e., segmental) in the L1 to another (i.e., suprasegmental contrasts) in the L2.

The results provide important evidence on how the use of F0 cues in the L1 can modulate the processing of lexical stress in the L2. Seoul Korean listeners' accuracy on the stress contrast condition was significantly higher than that of French listeners. This suggests that Seoul Korean L2 learners of English, who do not have lexical stress contrasts in their L1, can transfer the use of F0 cues from the processing of intonational cues to the laryngeal stop distinction in Seoul Korean to the processing of intonationally cued lexical stress in English. In other words, the processing of an intonationally cued lexical contrast in the L2 is facilitated when intonational cues signal a segmental contrast in the L1 compared to when they do not.

Interestingly, English listeners' accuracy in the stress contrast condition was on par with that of Seoul Korean listeners. This may be due to the absence of vowel quality cues to lexical stress in the stimuli. English listeners have been shown to use vowel quality as the most important cue when processing lexical stress contrasts, followed by pitch, duration, and intensity cues (e.g., Zhang and Francis, 2010; Chrabaszcz et al., 2014; Tremblay et al., 2021). The unavailability of vowel reduction cues in the present experiment is likely an important factor in explaining English listeners' difficulty in the processing of lexical stress (see also Experiment 2 of Kim and Tremblay, 2021).

For French listeners, the results of this study are consistent with those of previous studies on the processing of lexical stress by French listeners (Dupoux et al., 2001, 2008, 2010). For instance, in Dupoux et al. (2008), the results of French L2 learners of Spanish, who completed two- and four-item sequence-recall tasks with Spanish-like non-words (e.g., *MIpa* vs. *miPA*), are comparable to those in the present study, with a mean accuracy of 28.3% on the Spanish stress contrast condition. Thus, our results provide additional evidence that French listeners have difficulty processing lexical stress contrasts regardless of the L2 that they process.

The current findings clarify the scope of the cue-based transfer approach to L2 lexical stress processing by showing that the use of intonational cues can transfer *across* contrast types. We attribute Seoul Korean listeners' ability to process English lexical stress to their ability to use F0 cues when processing the laryngeal stop contrasts in their L1. One cannot preclude the possibility that French listeners transfer the use of F0 as a secondary cue to stop contrasts from the perception of French stops to the perception of English lexical stress. However, since F0 has a marginal effect on the perception of stop contrasts in French, the amount of transfer taking place is likely limited, whereas F0 is a primary cue to the lenis-aspirated and the fortis-lenis stop contrast in Seoul Korean, resulting in Seoul Korean listeners' superior performance compared to French listeners.

The present results are interesting to compare to those of Kim and Tremblay (2021). Using a similar sequence-recall task, Kim and Tremblay (2021) found that Gyeongsang Korean listeners outperformed Seoul Korean listeners in the processing of English lexical stress, a finding that was attributed to the transfer of F0 cues from lexical pitch accents in Gyeongsang Korean to lexical stress in English. One important implication from their findings and from ours is that cue transfer from the L1 to the L2 is *relative* and depends on the functional weight of the cue in the L1—specifically, how important the cue is for distinguishing lexical candidates (for discussion, see Tremblay et al., 2018). Taken together, the findings of these two studies suggest that F0 has a greater functional weight in Gyeongsang Korean than in Seoul Korean, and it has a greater functional weight in Seoul Korean than in French.

As mentioned in the introduction, speech perception is multidimensional, and acoustic cues do not equally contribute to signaling a sound contrast. This is also true of lexical stress in English. The present study neutralized duration and intensity cues to lexical stress, as Seoul Korean and French listeners were not necessarily predicted to differ in their use of these two cues. It would be interesting to investigate how Seoul Korean and French listeners weight suprasegmental cues to lexical stress in English when all three cues can potentially signal stress. The results of Kim and Tremblay (2021) suggest that Seoul Korean listeners do not benefit from the addition of duration and intensity cues to auditory stimuli that contrast in intonationally cued English lexical stress (unlike English listeners). Further research should compare Seoul Korean and French listeners on the weighting of all three suprasegmental cues to English lexical stress to determine if French listeners show greater reliance on duration and intensity cues to English stress than Seoul Korean listeners as a compensation strategy for their difficulty in using F0 cues.

From a theoretical perspective, the present findings have important implications. The cue-weighting theory of speech perception proposes that the underlying mechanism for learning speech categories or contrasts in both the L1 and the L2 is listeners' selective attention to specific acoustic dimensions, assuming that a phonetic category consists of a multidimensional structure where each dimension corresponds to a feature of the phonetic category (e.g., Iverson and Kuhl, 1995; Kuhl and Iverson, 1995; Francis and Nusbaum, 2002; Francis et al., 2008). Accordingly, the cue-based transfer approach stipulates that L2 learners' ability to attend to a particular cue in the L2 and associate it with a contrast or function that differs from that in the L1 depends on how much weight the cue has in the L1. The findings of this study indicate that intonational cues that have a similar function (e.g., to signal lexical information) can transfer from one type of contrast in the L1 (e.g., segmental contrast) to another type of contrast in the L2 (e.g., suprasegmental contrast). Thus, the results of the present study extend the scope of the cue-based transfer approach to the processing of L2 lexical stress in showing that L1-based cue transfer is not limited by the type of contrast signaled in the L1 and L2.

One important question that arises from the current findings is whether there are limits or constraints on L1-based cue transfer. A cue-based approach conjectures that phonetic learning involves cross-talker, cross-context, and cross-language generalization. Hence, there is no a priori reason to expect cues not to transfer across prosodic categories, contrast types, or functions. However, since what is important for this approach is the relative weight of cues, and because listeners focus their attention on the cues that have been deemed to have the greatest weight (i.e., primary cues; e.g., Francis and Nusbaum, 2002; Kondaurova and Francis, 2010), it is possible that listeners show transfer effects only for primary cues, and not for cues that have a weaker weight (i.e., secondary cues). It may thus be that the limits or constraints of L1-based cue transfer depend not on the prosodic category, contrast type, or function that the cues serve to signal or perform in the L1 and L2, but on the relative importance of specific cues across languages. In other words, cues may be more likely to transfer or have a noticeable effect on L2 speech perception if they are primary cues insofar as listeners are more likely to attend to these cues, and not as much if they are secondary cues, regardless of types of contrasts or functions.

From this perspective, the results of the present study may be interpreted as French listeners having more difficulty increasing the weight given to F0 in the perception of intonationally cued English lexical stress compared to Seoul Korean listeners because F0 is a comparatively less important lexical cue in French than in Seoul Korean. In other words, Seoul Korean listeners' ability to attend to F0 in the L2 may be explained by their relatively more extensive experience attending to this acoustic cue in the L1.

In a similar vein, it would be interesting to investigate whether Seoul Korean and French listeners differ in the use of vowel quality cues to English stress. French does not have lexical stress, but it has a reduced vowel, the schwa, which is never accented intonationally: If a phrase ends with a schwa, the previous syllable receives the phrase-final pitch accent, and the schwa can be pronounced or deleted depending on the context and/or the French dialect (e.g., Jun and Fougeron, 2002; Welby, 2006; Meunier and Espesser, 2011). Even within a phrase, the schwa can be deleted depending on the phonetic context in which it occurs (e.g., Jun and Fougeron, 2002). By contrast, Korean does not have a reduced vowel, but it has vowels that can assimilate to reduced vowels in English, with these vowels not having any relationship with accenting. The cue-based transfer approach would predict that Seoul Korean and French listeners would not necessarily differ in the processing of stress when the unstressed syllable is reduced, unlike what was predicted for the present study: French listeners would be able to transfer their use of vowel quality cues in the L1 to the processing of English reduced vowels in unstressed syllables, and Korean listeners would also be expected to process vowel quality cues to English stress but for a different reason—although Korean does not have vowel reduction, Korean listeners would be able to process vowel quality cues to English stress by assimilating full and reduced vowels to different Korean vowels (for such a proposal, see Connell et al., 2018).

CONCLUSION

The present study investigated whether the use of intonational cues can transfer from the processing of segmental contrasts in the L1 to the processing of intonationally cued lexical stress in the L2. A comparison of Seoul Korean and proficiency-matched French L2 learners of English showed that Seoul Korean listeners transferred their use of F0 cues from the processing of the laryngeal stop contrasts in Korean to the processing of lexical stress in English, as evidenced by their greater ability to process English stress compared to French listeners. From a theoretical perspective, this study further specified the scope of the cue-based transfer hypothesis, suggesting that listeners can transfer the use of intonational cues from the processing of segmental contrasts to the processing of lexical stress. Further research and more empirical data are needed to better understand the nature of the limits or constraints on cue-weighting transfer.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors upon request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Human Research Protection Program, The University of Kansas (IRB ID: STUDY00145019). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

HK created and recorded the stimuli, and analyzed the data. All authors contributed to conception, design of the study, wrote the first draft of the manuscript, contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by a Linguistics Research Scholarship awarded to HK from the Department of Linguistics at the University of Kansas. The article processing charges related to the publication of this article were supported in part by The University of Kansas (KU) One University Open Access Author Fund sponsored jointly by the KU Provost, KU Vice Chancellor for Research & Graduate Studies, and KUMC Vice Chancellor for Research and managed jointly by the Libraries at the Medical Center and KU - Lawrence.

ACKNOWLEDGMENTS

We thank the reviewers for their invaluable comments and suggestions. We also thank Dr. Allard Jongman, Dr. Joan Sereno, Dr. Jie Zhang, and the student members of LING 850 at the

University of Kansas for their insightful feedback on this work. We would also like to thank Dr. Elsa Spinelli and Dr. Pauline Welby for their help with advertising the study in France. Any misunderstandings are, of course, our own.

REFERENCES

Altmann, H. (2006). *The perception and production of second language stress: a cross-linguistic experimental study* (dissertation). The University of Delaware, Newark, DE, United States.

Bates, D., Maechler, B., 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

Boersma, P., and Weenink, D. (2019). *Doing Phonetics by Computer (Version 6.0.46)*. Retrieved from: http://www.praat.org (accessed January 28, 2019).

Braun, B., Galts, T., and Kabak, B. (2014). Lexical encoding of L2 tones: The role of L1 stress, pitch accent and intonation. *Second Lang. Res.* 30, 323–350. doi: 10.1177/0267658313510926

Broersma, M. (2010). "Korean lenis, fortis, and aspirated stops: effect of place of articulation on acoustic realization," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association* (Makuhari), 941–944. doi: 10.21437/Interspeech.2010-317

Chan, I. L., and Chang, C. B. (2019). Perception of nonnative tonal contrasts by Mandarin-English and English-Mandarin sequential bilinguals. *J. Acoust. Soc. Am.* 146, 956–972. doi: 10.1121/1.5120522

Cho, T. (1996). *Vowel correlates to consonant phonation: an acoustic-perceptual study of Korean obstruents* (master's thesis). The University of Texas at Arlington, Arlington, TX, United States.

Choi, W. (2022). Theorizing positive transfer in cross-linguistic speech perception: the Acoustic-Attentional-Contextual hypothesis. *J. Phonet.* 91, 101135. doi: 10.1016/j.wocn.2022.101135

Choi, W., Tong, X., and Samuel, A. G. (2019). Better than native: tone language experience enhances English lexical stress discrimination in Cantonese-English bilingual listeners. *Cognition* 189, 188–192. doi: 10.1016/j.cognition.2019.04.004

Chrabaszcz, A., Winn, M., Lin, C. Y., and Idsardi, W. J. (2014). Acoustic cues to perception of word stress by English, Mandarin, and Russian speakers. *J. Speech Lang. Hear. Res.* 57, 1468–1479. doi: 10.1044/2014_JSLHR-L-13-0279

Connell, K., Hüls, S., Martínez-García, M. T., Qin, Z., Shin, S., Yan, H., et al. (2018). English learners' use of segmental and suprasegmental cues to stress in lexical access: an eye-tracking study. *Lang. Learn.* 68, 635–668. doi: 10.1111/lang.12288

Cooper, N., Cutler, A., and Wales, R. (2002). Constraints of lexical stress on lexical access in English: evidence from native and nonnative listeners. *Lang. Speech* 45, 207–228. doi: 10.1177/00238309020450030101

Cutler, A., Wales, R., Cooper, N., and Janssen, J. (2007). "Dutch listeners' use of suprasegmental cues to English stress," in *Proceedings of the 16th International Congress for Phonetic Sciences,* eds J. Trouvain and W. J. Barry (Dudweiler: Pirrot) 1913–1916.

Dupoux, E., Peperkamp, S., and Sebastián-Gallés, N. (2001). A robust method to study stress 'deafness'. *J. Acoust. Soc. Am.* 110, 1606–1618. doi: 10.1121/1.1380437

Dupoux, E., Peperkamp, S., and Sebastián-Gallés, N. (2010). Limits on bilingualism revisited: stress 'deafness' in simultaneous French-Spanish bilinguals. *Cognition* 114, 266–275. doi: 10.1016/j.cognition.2009.10.001

Dupoux, E., Sebastián-Gallés, N., Navarrete, E., and Peperkamp, S. (2008). Persistent stress 'deafness': the case of French learners of Spanish. *Cognition* 106, 682–706. doi: 10.1016/j.cognition.2007.04.001

Francis, A. L., Baldwin, K., and Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Percept. Psychophys.* 62, 1668–1680. doi: 10.3758/BF03212164

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

Francis, A. L., and Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *J. Exp. Psychol. Hum. Percept. Perform.* 28, 349–366. doi: 10.1037/0096-1523.28.2.349

Holt, L. L., and Lotto, A. J. (2006). Cue weighting in auditory categorization: implication for first and second language acquisition. *J. Acoust. Soc. Am.* 119, 3059–3071. doi: 10.1121/1.2188377

Iverson, P., and Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *J. Acoust.Soc. Am.* 97, 553–562. doi: 10.1121/1.412280

Jun, S.-A. (1996). The influence of the microprosody on the macroprosody: a case of phrase initial strengthening. *UCLA Work. Pap. Phonet.* 92, 97–116.

Jun, S.-A. (1998). The accentual phrase in the Korean prosodic hierarchy. *Phonology* 15, 189–226. doi: 10.1017/S0952675798003571

Jun, S.-A. (2000). K-ToBI (Korean ToBI) labeling conventions. *UCLA Work. Pap. Phonet.* 99, 149–173.

Jun, S.-A., and Fougeron, C. (2000). "A phonological model of French intonation," in *Intonation: Analysis, Modeling and Technology,* ed A. Botinis (Dordrecht: Kluwer Academic Publishers), 209–242. doi: 10.1007/978-94-011-4317-2_10

Jun, S.-A., and Fougeron, C. (2002). Realizations of accentual phrase in French intonation. *Probus* 14, 147–172. doi: 10.1515/prbs.2002.002

Kang, K. H., and Guion, S. G. (2008). Clear speech production of Korean stops: changing phonetic targets and enhancement strategies. *J. Acoust. Soc. Am.* 124, 3909–3917. doi: 10.1121/1.2988292

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

Kim, H., and Jongman, A. (2021). The influence of inter-dialect contact on the Korean three-way laryngeal distinction: an acoustic comparison among Seoul Korean speakers and Gyeongsang speakers with limited and extended residence in Seoul. *Lang. Speech*. doi: 10.1177/00238309211037720

Kim, H., and Tremblay, A. (2021). Korean listeners' processing of suprasegmental lexical contrasts in Korean and English: a cue-based transfer approach. *J. Phonet.* 87, 101059. doi: 10.1016/j.wocn.2021.101059

Kirby, J. P., and Ladd, D. R. (2015). "Stop voicing and F0 perturbations: evidence from French and Italian," in *Proceedings of the 18th International Congress of Phonetic Sciences* (Glasgow: The University of Glasgow).

Kondaurova, M., and Francis, A. L. (2010). The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: comparison of three training methods. *J. Phonet.* 38, 569–587. doi: 10.1016/j.wocn.2010.08.003

Kuhl, P., and Iverson, P. (1995). "Linguistic experience and the "perceptual magnet effect"," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, ed W. Strange (Baltimore, MD: York Press), 121–154.

Lee, H., and Jongman, A. (2019). Effects of sound change on the weighting of acoustic cues to the three-way laryngeal stop contrast in Korean: diachronic and dialectal comparisons. *Lang. Speech* 62, 509–530. doi: 10.1177/0023830918786305

Lee, H., Politzer-Ahles, S., and Jongman, A. (2013). Speakers of tonal and non-tonal Korean dialects use different cue weightings in the perception of the three-way laryngeal stop contrast. *J. Phonet.* 41, 117–132. doi: 10.1016/j.wocn.2012.12.002

Lemhöfer, K., and Broersma, M. (2012). Introducing LexTALE: a quick and valid Lexical Test for Advanced Learners of English. *Behav. Res. Methods* 44, 325–343. doi: 10.3758/s13428-011-0146-0

Lin, C. Y., Wang, M., Idsardi, W. J., and Xu, Y. (2014). Stress processing in Mandarin and Korean second language learners of English. *Bilingual. Lang. Cogn.* 17, 316–346. doi: 10.1017/S1366728913000333

Lisker, L., and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20, 384–422. doi: 10.1080/00437956.1964.11659830

Meunier, C., and Espesser, R. (2011). Vowel reduction in conversational speech in French: the role of lexical factors. *J. Phonet.* 39, 271–278. doi: 10.1016/j.wocn.2010.11.008

Peperkamp, S., and Dupoux, E. (2002). A typological study of stress 'deafness'. *Labo. Phonol.* 7, 203–240. doi: 10.1515/9783110197105.203

Peperkamp, S., Vendelin, I., and Dupoux, E. (2010). Perception of predictable stress: a cross-linguistic investigation. *J. Phonet.* 38, 422–430. doi: 10.1016/j.wocn.2010.04.001

Qin, Z., Chien, Y. F., and Tremblay, A. (2017). Processing of word-level stress by Mandarin-speaking second language learners of English. *Appl. Psycholinguist.* 38, 541–570. doi: 10.1017/S0142716416000321

Qin, Z., Tremblay, A., and Zhang, J. (2019). Influence of within-category tonal information in the recognition of Mandarin-Chinese words by native and non-native listeners: an eye-tracking study. *J. Phonet.* 73, 144–157. doi: 10.1016/j.wocn.2019.01.002

Qualtrics, LLC (2020). *Qualtrics [Computer Program]*. Qualtrics. Retrieved from: https://www.qualtrics.com (accessed January 1, 2020).

R Development Core Team (2019). *R: A Language and Environment for Statistical Computing (Version Version 1.2.1335)*. Vienna: R Foundation for Statistical Computing. Retrieved from: http://www.rproject.org

Saerens, M., Serniclaes, W., and Beeckmans, R. (1989). Acoustic versus contextual factors in stop voicing perception in spontaneous French. *Lang. Speech* 32, 291–314. doi: 10.1177/002383098903200401

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

Serniclaes, W. (1987). *Etude expérimentale de la perception du trait de voisement des occlusives du français* (dissertation). Université Libre de Bruxelles, Brussels, Belgium.

Shport, I. A. (2015). Perception of acoustic cues to Tokyo Japanese pitch-accent contrasts in native Japanese and naive English listeners. *J. Acoust. Soc. Am.* 138, 307–318. doi: 10.1121/1.4922468

Silva, D. J. (2006). Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology* 23, 287–308. doi: 10.1017/S0952675706000911

Tremblay, A., Broersma, M., and Coughlin, C. E. (2018). The functional weight of a prosodic cue in the native language predicts the learning of speech segmentation in a second language. *Bilingual. Lang. Cogn.* 21, 640–652. doi: 10.1017/S136672891700030X

Tremblay, A., Broersma, M., Zeng, Y., Kim, H., Lee, J., and Shin, S. (2021). Dutch listeners' perception of English lexical stress: a cue-weighting approach. *J. Acoust. Soc. Am.* 149, 3703–3714. doi: 10.1121/10.0005086

Tremblay, A., and Ransijn, J. (2015). *Package 'LMERConvenienceFunctions'*.

Welby, P. (2006). French intonational structure: Evidence from tonal alignment. *J. Phonetics*. 34, 343–371.

Wiener, S., and Goss, S. (2019). Second and third language learners' sensitivity to Japanese pitch accent is additive: an information-based model of pitch perception. *Stud. Second Lang. Acquisit.* 41, 897–910. doi: 10.1017/S0272263119000068

Zhang, Y., and Francis, A. (2010). The weighting of vowel quality in native and non-native listeners' perception of English lexical stress. *J. Phonet.* 38, 260–271. 6

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

REVIEW
published: 03 June 2022
doi: 10.3389/fpsyg.2022.906848

How Tone, Intonation and Emotion Shape the Development of Infants' Fundamental Frequency Perception

Liquan Liu[1,2,3], Antonia Götz[1,4], Pernelle Lorette[5] and Michael D. Tyler[1,3]*

[1]MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Penrith, NSW, Australia, [2]Center for Multilingualism in Society Across the Lifespan, University of Oslo, Oslo, Norway, [3]Australian Research Council Centre of Excellence for the Dynamics of Language, Canberra, ACT, Australia, [4]Department of Linguistics, University of Potsdam, Potsdam, Germany, [5]Department of English Linguistics, University of Mannheim, Mannheim, Germany

Fundamental frequency (f_0), perceived as pitch, is the first and arguably most salient auditory component humans are exposed to since the beginning of life. It carries multiple linguistic (e.g., word meaning) and paralinguistic (e.g., speakers' emotion) functions in speech and communication. The mappings between these functions and f_0 features vary within a language and differ cross-linguistically. For instance, a rising pitch can be perceived as a question in English but a lexical tone in Mandarin. Such variations mean that infants must learn the specific mappings based on their respective linguistic and social environments. To date, canonical theoretical frameworks and most empirical studies do not view or consider the multi-functionality of f_0, but typically focus on individual functions. More importantly, despite the eventual mastery of f_0 in communication, it is unclear how infants learn to decompose and recognize these overlapping functions carried by f_0. In this paper, we review the symbioses and synergies of the lexical, intonational, and emotional functions that can be carried by f_0 and are being acquired throughout infancy. On the basis of our review, we put forward the Learnability Hypothesis that infants decompose and acquire multiple f_0 functions through native/environmental experiences. Under this hypothesis, we propose representative cases such as the synergy scenario, where infants use visual cues to disambiguate and decompose the different f_0 functions. Further, viable ways to test the scenarios derived from this hypothesis are suggested across auditory and visual modalities. Discovering how infants learn to master the diverse functions carried by f_0 can increase our understanding of linguistic systems, auditory processing and communication functions.

Keywords: lexical tone, intonation, Prosody, phonological theory, sensory processing, cognitive processing, cross-linguistic transfer, emotional tone

OPEN ACCESS

Edited by:
Hatice Zora,
Max Planck Institute for Psycholinguistics, Netherlands

Reviewed by:
Aijun Li,
Institute of Linguistics (CASS), China
Linda Polka,
McGill University, Canada

****Correspondence:***
Liquan Liu
l.liu@westernsydney.edu.au

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

Received: *29 March 2022*
Accepted: *10 May 2022*
Published: *03 June 2022*

Citation:
Liu L, Götz A, Lorette P and Tyler MD (2022) How Tone, Intonation and Emotion Shape the Development of Infants' Fundamental Frequency Perception. Front. Psychol. 13:906848. doi: 10.3389/fpsyg.2022.906848

INTRODUCTION

From the beginning of life, humans are exposed to the fundamental frequency (f_0; Titze et al., 2015). The f_0 carries a wide range of information. This includes linguistic (e.g., lexical tone), paralinguistic (e.g., speaker intent, emotion, Crystal and Quirk, 1964; Gussenhoven, 2002), and extralinguistic information (e.g., melody, Johnson, 1990; He et al., 2007). While some crucial communicative functions carried by f_0 appear to be universal, such as intonation (Best, 2019), others can vary across the world's languages (e.g., signalling grammatical information; Hyman, 2011, 2016; Remijsen, 2016). For example, a syllable /ja/ with a rising f_0 can be recognised as an attention getter for a Dutch speaker, but as the word "tooth" for a speaker of Mandarin. Thus, to acquire the language of their environment, infants are faced with a complex task. They must learn to disambiguate, decompose, recognise, and learn the patterns of f_0 variability that apply to different linguistic, paralinguistic, and non-linguistic domains.

It is impressive that infants process different sources of speech information and eventually learn to disentangle functions of f_0 during speech perception, yet *how* they achieve this has received little attention in the empirical or theoretical literature. Research on infants' perception, production, and learning of the functions carried on f_0 has focused mainly on a single specific domain of interest, for example, music, lexical tone, or intonation. To explain how infants learn to perceive the multifaceted and cross-domain f_0 signal, it will be necessary to integrate findings across those different domains of interest. The purpose of this paper is to sketch out an approach to doing that across three f_0 functions: tone, intonation and emotion. We first review empirical studies on infants' acquisition of the three functions of interest along with their interactions. After that, theoretical considerations are discussed, followed by the proposal of a novel hypothesis.

INFANTS' ACQUISITION OF TONE, INTONATION, AND EMOTION CARRIED ON f_0

Tone

Around 60–70% of world languages are tonal (Yip, 2002), predominantly using contrastive f_0 variations to differentiate lexical and grammatical changes. Spreading across Asia, Africa, (indigenous) America, Europe and South Pacific regions (Maddieson, 2013), tone languages are spoken by more than half of the world's population (Fromkin, 2014). Among tones, the predominant f_0 changes lie in pitch height (level, register) and pitch direction (contour, slope; Chao, 1947; Gandour, 1983; Gussenhoven, 2004). Of particular interest are tone languages that rely on lexical tones to distinguish word meanings. For instance, the syllable [ji] in Cantonese means "cure" when bearing a high level tone, but "son" with a low falling tone (Francis et al., 2008). In a tone language such as Mandarin, f_0 carries the primary cues for perception (Gandour, 1983; Massaro et al., 1985; Lee and Lee, 2010), in addition to secondary cues such as intensity and duration (Jongman et al., 2006). The stark difference in f_0 functions in lexical-tone versus non-tone languages raises important questions about how these typological differences influence the development of speech perception, speech production, and word learning.

Speech perception research has shown clear differences in the way that speech is perceived by tone and non-tone language learning infants (Fikkert et al., 2020) as well as by adults (Burnham and Singh, 2018; Liu et al., 2022). Such studies have demonstrated increased tonal sensitivity over the first year after birth for tone language learners and decreased sensitivity for non-tone language learners (Mattock and Burnham, 2006; Mattock et al., 2008; Yeung et al., 2013). However, empirical evidence in the last decade appears to challenge these canonical patterns. For instance, there appears to be an age-based increase in sensitivity to certain tonal contrasts for both tone and non-tone language learning infants (Chen and Kager, 2016; Chen et al., 2017; Tsao, 2017; Ramachers et al., 2018; Singh et al., 2018), and behavioural and neural studies report that bilingual infants tend to be more resilient in perceiving and learning tones even when they do not exist in these infants' linguistic repertoires (Graf Estes and Hay, 2015; Liu and Kager, 2017a; Liu et al., 2019). Further, a U-shaped sensitivity has been reported in non-tone language learning infants, such that the decline in sensitivity observed over the first year of life is reversed in their second year (Liu and Kager, 2014, 2017a; Götz et al., 2018). Thus, while initial investigations into infant speech perception showed expected declines in sensitivity for tonal contrasts for infants learning a non-tone language, more recent studies suggest that the developmental trajectory requires a more nuanced theoretical interpretation (for similar observations on the development of consonant perception, see Tyler et al., 2014; Liu and Kager, 2015).

Tone production studies typically involve tone language-learning infants, who start producing f_0 contours around 7 months (Chen and Kent, 2009). It is unclear whether the f_0 produced is on a lexical or utterance level (or both), however, because adults cannot identify the ambient language when listening to the babbling of 8–12-month-old English and Mandarin-learning infants extracted from recordings (Lee et al., 2017). Mature production can be observed shortly after 2 years of age (Li and Thompson, 1977; So and Dodd, 1995; Hua and Dodd, 2000; Hua, 2002; To et al., 2013, for a review, see Peng and Chen, 2020). Recent acoustic analyses challenged this conclusion, however, as they have revealed substantial differences between children and adults' tone production. Mandarin-learning children have been found not to reach an adult level of tonetic realisation until the age of 5 (Wong et al., 2005; Wong, 2012a,b, 2013), possibly due to complex tone articulation (Wong, 2012a) or tonal rules (Chen et al., 2015; Wewalaarachchi and Singh, 2016).

The conflicting findings also extend to word learning. To learn a tone language, children need to associate lexical items that differ minimally in tonal contrasts with different word meanings. Making such associations does not appear to be easy for children at 2–3 years (Shi et al., 2017) and the lexical encoding does not stabilise until around 4–5 years (Singh et al., 2015). Sensitivity to tonal contrast is not required for non-tone language learning infants, yet they are sensitive to f_0 variations on words at 7.5 and 18 months (Singh et al., 2008, 2014). While 14-month-olds are able to associate non-native tones with different objects, that ability decreases at 18 months (Hay et al., 2015; Liu and Kager, 2018). By

2.5 years, they no longer consider f_0 change to be lexically relevant (Quam and Swingley, 2010).

Mixed findings in perception, production, and learning trajectories among tone and non-tone language-learning infants require further investigation. In this paper, we raise the hypothesis that these discrepancies can be attributed to other functional uses of f_0, which are linguistically and paralinguistically relevant in all spoken languages as they can also manifest on the utterance level, such as intonation.

Intonation

All spoken languages employ intonation (Best, 2019), where f_0 acts at a phrasal level (distinct from the word-level tone, and in addition to other cues such as voice quality; Ladd et al., 1985). When learning a language like English, children need to know that different f_0 contours applied to the same utterance can signal different (e.g., narrative, interrogative) connotations. Intonation can convey linguistic information, facilitate the acquisition of other linguistic components (e.g., words; Thiessen et al., 2005), raise attention (Sullivan and Horowitz, 1983), and carry speakers' intentions (Gussenhoven, 2002; Esteve-Gibert et al., 2017). Adult listeners can encode both focus and interrogative meaning in intonation (Liu and Xu, 2005). Arguably, this makes intonation a unique component, as it spreads across linguistic and paralinguistic fields and serves grammatical, pragmatic and affective functions (Snow and Balog, 2002). Furthermore, intonation plays a crucial role in caretaker-infant interactions and communications (Stern et al., 1982; Fernald and Simon, 1984; Fernald, 1989).

Infants' perception and production of intonation develop concurrently with tone throughout infancy and early childhood. Newborns are sensitive to intonation in speech (Nazzi et al., 1998; Sambeth et al., 2008) and 6-month-olds can use pitch contours to parse utterances into clauses (e.g., Seidl, 2007). By 6 and 9 months, European Portuguese-learning infants can discriminate single prosodic-word utterances differing in statement (falling) or yes–no question (falling-rising) intonation (Frota et al., 2014; Frota and Butler, 2018). Despite their sensitivity to the f_0 differences that characterise intonation, children do not appear to rely strongly on intonation to signal conversational turn taking until 3 years and onwards (Keitel et al., 2013). Why they are reluctant to do so at earlier ages needs to be understood.

Arguably, intonation production starts from birth with crying (Mampe et al., 2009) and vocalisation shortly after birth (Kent and Murray, 1982). Newborn infants' crying patterns already reflect the intonation patterns of their native language (Mampe et al., 2009; Wermke et al., 2016, 2017; Manfredi et al., 2019; Prochnow et al., 2019). Infants begin with a predominant falling pitch contour then progress to other f_0 patterns, with accent range increasing with age (Snow, 2001). The production of pitch register stabilises in the single-word period, and core features are controlled in the two-word stage (Snow and Balog, 2002). However, the development of intonation production in the first 2 years of life is not linear. At the end of the first year after birth, rising and falling contours are produced with a smaller accent range in comparison to the 6–9 and above 18-month-olds. This U-shaped pattern needs further investigation and explanation.

With respect to the interaction between tone and intonation, researchers are prone to argue for a linguistic status of tone and an ambiguous status of intonation: from a categorical perspective, studies favour evidence for discrete tone but not intonation categories, as one "intoneme" may consist of various intonational elements (Tonkova-Yampol'skaya, 1969; but see So and Best, 2014 on "i-category"). Tone-language speakers show distinct tone and intonation processing differences on single-syllable units, not only in the neural organisation of subcortical and cortical structures but also hemispheric lateralisations (Chien et al., 2020), although to date, no consensus has been reached on whether intonation is dominantly processed in the left or right hemisphere. An utterance-final rising f_0 tends to be a universal cue for the perception of interrogation (Gussenhoven and Chen, 2000; Liang and Heuven, 2007), but perception of intonation appears to be tone-dependent. In Mandarin, a yes/no question is more easily identified when the utterance ends with a falling than a rising tone (Yuan, 2011), and a declarative versus interrogative contrast elicits strong mismatch negativity responses on syllables with falling but not rising tones (Ren et al., 2013). Research connecting intonation with word learning is relatively scarce. Although English speakers demonstrate the presence of long-term memory traces for prosodic information in the brain (Zora et al., 2015), English-learning 2-year-olds do not interpret salient pitch contour differences (rising-falling vs. falling-rising) as inherent to novel words (Quam and Swingley, 2010).

Such tone-intonation interaction in perception is not restricted to speakers of a tone language. Among non-tone language speakers, the component that stabilises the earliest, pitch register (Snow and Balog, 2002), can facilitate the perception of non-native tone contrasts (Liu et al., 2022). Non-tone language speakers' knowledge of intonation also appears to influence tone perception. For instance, the rising versus falling tones in Mandarin Chinese are similar to the declarative versus interrogative f_0 patterns in languages such as English (Braun and Johnson, 2011; So and Best, 2011, 2014). Indeed, when examining American English-learning infants' Mandarin tone-object association at 14 months, infants were more successful for words with a rising tone than for words with the other three Mandarin tones (Hay et al., 2015, 2019). This suggests that they may have been able to capitalise on their developing sensitivity to English rising pitch intonation for perception of non-native words differing by lexical tone. Adopting intonation patterns from a non-tonal native language for perception of non-native tones is consistent with theories of perceptual assimilation (Best, 1994, 2019; Best et al., 2009; So and Best, 2010, 2014), which may provide a potential theoretical explanation for the U-shaped developmental pattern reported in infant perception of non-native tones (Liu and Kager, 2014, 2017a; Götz et al., 2018). Children learning non-tone languages may become less sensitive to certain f_0 patterns as they recognise that tonal variations do not signal lexical distinctions in their native language, while also learning the complementary functions that *are* carried on f_0.

The nonlinear developmental trajectory for intonation from infancy to toddlerhood (Snow and Balog, 2002), the restricted use of f_0 as a cue in intonation in early childhood (Keitel et al., 2013), and the overlap between tone and intonation in adulthood across the world's languages (e.g., Gussenhoven and Chen, 2000) all highlight the need to comprehensively understand f_0 functions along the developmental trajectory. Additionally, research on infants' acquisition of intonation may benefit from considering the prosodic and information structures of intonation, but few studies have taken this approach (Frota and Butler, 2018). For example, according to Autosegmental-Metrical accounts (Pierrehumbert, 1980; Grice et al., 2006; Ladd, 2008; Arvaniti and Fletcher, 2020), intonation is composed of a series of tonal events. To reveal the trajectory and mechanisms infants use to recognise word- and phrase-level prosody from continuous speech, it may be necessary to take the componential structure of intonation into consideration. Further, the visual aspect of intonation, often discussed in sign languages (e.g., Dachkovsky and Sandler, 2009), it still poorly understood in spoken languages. While expressing uncertainty, speakers not only use prosodic cues such as rising intonation, but also facial cues involving eyebrow raising, head tilting, furrowing, etc. (Dijkstra et al., 2006; Roseano et al., 2016).

In the next section, we attempt to explore the f_0 function in the domain of emotion, as well as the entanglement between the intonational and emotional functions in speech directed to infants.

Emotion

At first glance, there are differences in how theories consider f_0 between linguistic and emotional domains. This is not surprising since emotion theories typically focus on visual emotional signals (e.g., facial expressions) rather than how emotion is coded in speech. Theoretical debates centre on whether humans possess innate basic emotion categories, in both facial expressions (Chong et al., 2003; Gendron et al., 2018) and emotions in vocalisations (Sauter et al., 2010, 2015; Gendron et al., 2014, 2015). Empirical evidence suggests distinct processing of f_0 functions in intonation and in emotion. Emotional voice cues are processed predominantly in the auditory cortical areas in the right hemisphere, whereas phonemic cues are processed mainly in the left (Kotz et al., 2006; Scott and McGettigan, 2013). Limited studies have discussed the interaction between linguistic and emotional f_0 functions (Kotz and Paulmann, 2007; Pell and Kotz, 2011). It is unclear whether certain regions are responsible for f_0 variations in both emotional and linguistic states (Frühholz et al., 2012; Liebenthal et al., 2016).

For preverbal infants, perception of emotion is critical for survival in a social world, as it constitutes one of the critical social cognition skills. While emotion signals in the visual domain are most representative in a speaker's face and body language, they are carried primarily by f_0 in speech (Remez et al., 1981; Ladd et al., 1985; Scherer, 1986, 2003; Goldbeck et al., 1988). There are also secondary cues for emotion in speech (Murray and Arnott, 1993; Banse and Scherer, 1996; Bänziger et al., 2015; Pell et al., 2015), including intensity and speech rate (Scherer, 1986), pausing structure (Cahn, 1990) and duration (Mozziconacci, 1998), and timbre/voice quality (Gobl et al., 2002; Gobl and Chasaide, 2003; Yanushevskaya et al., 2018). In particular, f_0 modulates and strengthens the affective and motivational contexts in both infants (Stern et al., 1982) and adults (Frick, 1985). It also has an advantage over other cues, such as timbre, that it is simple to measure and quantify.

With respect to emotion perception, infants' ability to experience and perceive emotion has been hypothesised to develop as a function of neural development, increasing the capacity of processing emotional concepts with the aim of assigning meaning to sensory inputs and guiding behaviour (Hoemann et al., 2019). In their first year of life, infants are sensitive to emotions expressed from different cultures (Liu et al., 2021), and employ different attentional strategies based on their native culture (Geangu et al., 2016). Although emotional f_0 is highly salient in the environment from the beginning of life (ManyBabies Consortium, 2020), and its development is likely linked with the neuro-cognitive development of socio-emotions, the detailed trajectory of emotional f_0 remains unclear.

There appear to be f_0 patterns with distinct acoustic characteristics for different emotions (Liu and Pell, 2012; Wang and Lee, 2015), although findings are mixed on whether emotional f_0 patterns are universal or culturally-specific (Murray and Arnott, 1993; Pell et al., 2009; Li, 2015). Some perception studies have suggested a universal association between high f_0 and positive emotion (e.g., happiness, Ortony et al., 1990; Ilie and Thompson, 2006; Belyk and Brown, 2014), but the same trend has not been observed in other corpus studies (Laukka et al., 2005; Goudbeek and Scherer, 2010). The f_0 acoustics of the same emotional tone can vary across studies in height and range (Pell et al., 2009), along with other cues such as intensity and duration (Wang and Lee, 2015; Wang and Qian, 2018). Furthermore, cross-linguistic and cultural differences have been reported in both the acoustic manifestation (Douglas-Cowie et al., 2003; Anolli et al., 2008; Wang et al., 2018) and the interpretation (Koeda et al., 2013) of f_0. Despite this substantial variation, infants appear to identify regularities to build their knowledge.

There has been a debate in the literature on the processing of emotions in (visual) facial expressions about whether universal categories of basic emotional categories (e.g., happiness, anger) exist (Gendron et al., 2018). Infants can disambiguate between some emotional categories (Caron et al., 1985; Haviland and Lelwica, 1987; Soken and Pick, 1999; for a review, see Widen, 2013), yet it is unclear whether they conceptualise and abstract emotional features such as valence or arousal (Ruba et al., 2020). In comparison, research on processing of (auditory) vocal expressions of emotion is relatively scarce. Unlike 3-month-olds, infants at 5 months can discriminate between vocal expressions of positive and negative valence, but they do so reliably only in the presence of a face (Walker-Andrews and Grolnick, 1983; Walker-Andrews and Lennon, 1991). Infants aged 7 months process emotions of positive and negative valence differently, not only in facial expressions (Nelson and De Haan, 1996) but also in emotional prosody (Grossmann et al., 2005). With respect to the production of emotional f_0, a parental rating study has shown that vocalisations of 2-month-olds can be judged to fit along a comfort-discomfort dimension (Papoušek, 1989). Infants often use prosody, including (high) f_0, to signal what is perceived by their caretakers as emotional cues, be it wailing of fear or crying for attention (for a review, see Bryant, 2021). Little is known about the two-way relationship of f_0 functions in tone and emotion, although language background (tone vs. non-tone languages) has been shown to play a role. Larger f_0 variations of emotional tones are produced by non-tone than tone language speakers (Ross et al., 1986; Anolli et al., 2008; Wang et al., 2018), suggesting that the lexical function of f_0 constrains its use for emotional function.

Some studies have shown that emotional f_0 can facilitate word learning (for a review, see Doan, 2010). For example, words with emotional variations are better recognised in fluent speech by English-learning 7-8-month-olds than words without such variability (Singh, 2008). Infants aged 10.5 months showed significant positive recognition scores for words familiarised in happy but not in neutral emotion text passages (Singh et al., 2004). Words produced with an emotional f_0 assist infants in establishing representations and facilitate their word learning. While this does not automatically imply that they have decoded the emotional function carried on f_0, they are clearly sensitive to the f_0 differences between words produced with a neutral versus emotional f_0. Infants in their first year of life appear to have the capacity to separate linguistic and emotional functions of f_0, but no direct evidence of that has been reported.

Discussion on the interaction between intonational and emotional f_0 functions can be found in the area of infant-directed speech (IDS), a distinctive speech style that caretakers use to communicate with infants (Fernald, 1985, 1992). IDS is more exaggerated, with higher f_0 and wider f_0 ranges than adult-directed speech (ADS). Infants prefer IDS over ADS across the world's languages (ManyBabies Consortium, 2020). Some identify intonation as the key reason for this preference (Katz et al., 1996), whereas others attribute it to its attention-grabbing qualities (Burnham et al., 2002) and the positive emotion embedded in IDS (Singh et al., 2002). Infants appear to be sensitive to f_0 variations as early as 4 months of age, when they prefer f_0 but not amplitude or duration variations in IDS (Fernald and Kuhl, 1987). The fact that pragmatic functions encompassing both intonation and emotion, such as approval or prohibition, are more clearly expressed in IDS than in ADS, suggests that infants are capable of identifying those f_0 functions (Fernald, 1989; Moore et al., 1997). Indeed, as early as 5 months, infants are able to associate positive emotion in IDS with approval vocalisations, and negative emotion with prohibition vocalisations (Fernald, 1993). The functions of IDS appear to change over the first year of life, with ratings of mothers' IDS showing general decrease in comforting and soothing functions, and an increase in attentional and directive functions (Kitamura and Burnham, 2003). Infants' preferences for those functions appear to follow the same developmental trend (Kitamura and Lam, 2009). Despite infants' clear sensitivity to these f_0 patterns, another study suggests that children do not consider f_0 in speech as a reliable cue to indicate emotions until around 4–5 years of age (Quam and Swingley, 2012).

To our knowledge, no study has attempted to tease apart the three-way interaction between tone, intonation, and emotional functions in f_0. Trends may be observed in emotional f_0 from its immense variations, but not "rules" in the same sense as tone (e.g., "a tone language has a set of fixed pitch variations") or intonation (e.g., "a question usually has a rising pitch"). Thus, while there are broad indicators about the association between f_0 and emotion, this relationship, as well as its consistency across languages and cultures, is still under investigation. The interactions in between tone, intonation and emotion remain unclear, and research on IDS cannot efficiently disentangle its impact from intonational or emotional perspectives.

Summary

The fluctuating f_0 signal contains overlapping information from different sources that infants need to decompose and recognise. We have focused on three distinct functions carried by f_0; tone, intonation, and emotion. It is not yet clear whether languages differ from each other in the way that emotion is expressed using f_0, but there are clear differences in the ways that languages use f_0 for tone and intonation. Infants do not know innately whether the information in f_0 refers to tone, intonation, or emotion. They must learn which aspects of the fluctuating f_0 signal correspond to different functions.

Studies on the developmental trajectories of infants' sensitivity to the tonal, intonational, and emotion aspects carried on f_0 have yielded mixed findings. Unstable and fluctuating developmental trajectories have been reported for tone, not only for infants learning a tone language but also for those learning a non-tone language in the first 2 years of life. Similarly, infants' intonation development does not appear to be linear before Year 2, and children do not use f_0 for intonation reliably until after Year 3. Although the contribution of f_0 on emotion is widely acknowledged, incongruent findings have been reported across the world's languages. Reliable use of f_0 as a cue to indicate emotion has only been found after Year 4 (Quam and Swingley, 2012).

Research on infant speech perception has only recently begun to focus on f_0 and there is certainly more work that needs to be done to establish clear developmental patterns. Nevertheless, it is clear that infants are sensitive to f_0 across domains, in tone (Liu and Kager, 2014), intonation (Frota et al., 2014) and emotion (Singh et al., 2004), and it appears that robust knowledge about tone is learned ahead of intonation and emotion. This observation is consistent with the idea that discrete categories for tone seem to be established earlier and more easily than they are for intonation (Tonkova-Yampol'skaya, 1969; Snow, 2006; Yeung et al., 2013). Indeed, it could be argued that the variability in the way that the three functions are represented in f_0 increases from tone, to intonation, then emotion. Such variability would make an infant's job of learning the f_0 patterns even more challenging, which may explain the developmental progression and fluctuation across domains.

Although traces of overlap in between these domains appear in literature, there is insufficient empirical data to disentangle the interactions between tone, intonation and emotion in the development of f_0 perception. To arrive at a clear explanation of how infants learn to use f_0 cues in linguistic and paralinguistic functions, it is necessary to formulate a theoretical framework that incorporates f_0 functions across multiple domains.

THEORETICAL CONSIDERATIONS

Investigating how infants solve the puzzle of decomposing f_0 into different functions is a rare opportunity to observe language development across different communicative domains. One interesting aspect of f_0, from a developmental perspective, is that an f_0 pattern that signals a tonal function in one language could be perceived as intonation in another. Proposing a perspective that can conceptually integrate across all three lines of inquiry – tone, intonation and emotion – may seem ambitious, but it is necessary to consider all of these aspects to understand how infants learn to decode f_0. Given the developmental patterns that have been observed for the three domains, a purely bottom-up statistical learning solution seems unlikely. Rather, infants may require multimodal experiences from their environment to develop functional speech communication skills. Our current understanding of how tone or intonation is coded in the visual modality, and how emotion is coded in the auditory speech signal is rudimentary. Nevertheless, addressing the multifunctionality of the speech signal using a global approach, conveying linguistic, paralinguistic, and affective information simultaneously, is critical for a comprehensive model of speech development. Any theory addressing f_0 perception and development will need to be able to explain how children acquire their native f_0 functions and account for the mixed findings observed in previous literature. On these bases, we argue for four critical aspects that must be properly addressed by any theories concerning f_0 perception and development.

- *Disambiguation*: how infants disentangle and recognise multiple overlapping f_0 patterns
- *Categorisation*: how infants learn that those patterns correspond to a given (native) linguistic or paralinguistic function
- *Accomodation*: how infants tackle f_0 functions that deviate from their native functional use
- *Interaction*: why recognition, learning and cue weighting of f_0 fluctuate along the development

Below, we consider how developmental theories of speech perception, cognition, and statistical learning may contribute to a broad theoretical approach to explaining the eventual successful acquisition of f_0 functions.

Speech Perception

From a developmental perspective, *Perceptual Attunement* accounts (Werker and Hensch, 2015; Reh et al., 2020) propose that an infant's perception gradually shifts from universal into native or environmentally-attenuated perception patterns. Such changes occur across domains and modalities, fitting well in the aspect of *categorisation*. Such accounts associate well with, and arguably, lay the foundation of speech processing theories. For linguistic functions such as tone and intonation, infants typically exhibit initial biases or universal sensitivity, and quickly tune into the f_0 patterns of their native language (Burnham, 1986). Meanwhile, assimilations or perceptual difficulties surface since non-native or unfamiliar f_0 patterns are tuned out. Having said that, discrepancies from the attunement process have been reported for native and non-native f_0 patterns (Fikkert et al., 2020). Though overlapping f_0 patterns have been used as a possible explanation for these findings, theories of perceptual attunement will need to demonstrate *disambiguation*: how infants overcome overlaps in (e.g., f_0) functions along the developmental trajectory.

Further, models and theories of infants' acquisition of their L1 phonological system have been devised to explain how infants tune in to the phonetic features that signal phonological similarities and differences in the language of their environment (e.g., Best, 1994; Escudero, 2005; Kuhl et al., 2008; Polka and Bohn, 2011). The focus of these models has been on the acquisition of consonants and vowels (henceforth, *phones*). Here, we use the framework of the *Perceptual Assimilation Model* (PAM; Best, 1994; Best et al., 2009, Tyler et al., 2014) to consider how such models might account for the acquisition of f_0 functions.

A key empirical observation that led to the development of PAM was that English infants and adults had high discrimination accuracy for non-native Zulu click consonants despite never having encountered them before (Best et al., 1988). When asked to write down what they heard, all participants reported relying on non-speech characteristics of the consonants (e.g., water dripping, fingers snapping, or tongue popping). To account for this, PAM proposes that non-native phones may be perceived as speech (i.e., assimilated to the native phonological system) or as non-speech. When perceived as speech, a non-native phone may be assimilated as categorised (as a good, medium, or poor exemplar of a native phonological category) or uncategorised (not a clear exemplar of any single L1 category). Discrimination of non-native phonemes that are perceived as speech is crucially dependent on how it is assimilated to the native phonological system. Sometimes natively tuned perception will support discrimination (e.g., when each non-native phone is assimilated to a different L1 phonological category) and sometimes it will make it difficult to perceive any differences between them (e.g., when the non-native phones are perceived as equally good or poor exemplars of the same L1 category). Contrasting non-native phones that are perceived as non-speech (e.g., click consonants) are discriminated well by adults because they learned that the phonetic features of these categories are not used for linguistic purposes in their native language. Consistent with this account, native speakers of the click languages Zulu and Sesothu predominantly perceived non-native!Xóõ click consonants as speech (Best et al., 2003). Both click consonants in one of the !Xóõ contrasts were perceived as the same L1 click consonant category by both Zulu and Sesothu listeners. Importantly, English listeners perceived the same click consonants predominantly as non-speech and their discrimination of the contrast was more accurate than both groups of click language speakers. It appears that the English speaking adults had learned, as infants, that the phonetic characteristics that correspond to click consonants were not part of the L1 phonological space.

According to PAM, infants transition from language-independent phonetic sensitivity to natively tuned perception by recognising higher order invariant information in articulatory patterns through processes of perceptual learning (Gibson and Pick, 2000). Phonetic variability is crucial for phonological development because infants need to learn not only those phonetic differences that signal a difference in meaning (the principle of phonological distinctiveness), but also those variable phonetic characteristics that define a category (the principle of phonological constancy; Best et al., 2009; Best, 2015). The region of phonetic space that is dedicated to speech is known as the phonological space. Click consonants would fall outside of the phonological space for English speakers but they would fall inside the phonological space for click language speakers. The development of phonological categories is beneficial for L1 perception because it supports accurate and rapid detection of the critical phonetic differences that signal a potential difference in word meaning. However, once infants have begun to tune into the L1 phonology, non-native speech is also perceived in terms of its similarities and differences to their developing L1 phonological categories. If they happen to perceive each phoneme in a non-native contrast as different L1 phonological category (e.g., one phoneme as /b/and the other as/d/, a PAM two-category assimilation) then their natively tuned perception will still support rapid and accurate discrimination. However, if both non-native phonemes are perceived as the same L1 category (e.g., the Hindi dental vs. retroflex plosive contrast for English native speakers, Werker and Logan, 1985; a PAM single-category assimilation), then discrimination is poor.

If fluctuating f_0 patterns were considered in a similar way as the varying articulatory-acoustic patterns that demarcate consonants and vowels, then it is conceivable that infants might use similar learning mechanisms to separate the linguistic, paralinguistic, or extralinguistic functions carried on f_0. For example, the f_0 patterns that are used in a tone language for lexical distinctions may be similar to those used for other functions in a non-tone language, such as intonation (for a discussion, see, Best, 2019). The developmental changes in infants' responses to f_0 fluctuation might then be explained by infants' learning and recognition of the various functions at different ages. For infants who experience phonological characteristics of a non-tone language, f_0 is irrelevant for lexical distinctions. This may explain why discrimination of tonal contrasts initially declines. The subsequent improvement would then be due to the development of sensitivity to other types of f_0 information. Thus, from the perspective of the Perceptual Assimilation Model, *disambiguation* and *categorisation* occur through processes of perceptual learning. *Accommodation* may be observed if infants perceive a non-native f_0 pattern as consistent with a different type of function in their L1, and *interaction* may be explained by the different timescales for perceptual development of linguistic, paralinguistic, and extralinguistic information.

Cognition

Another potential joinder of the three areas of f_0 functions resides in cognitive competition. Theories such as the *Functional Load Hypothesis* (FLH, Berinstein, 1979) postulate that our prosodic space of a given language is finite, and therefore, assume competition in phonological processing. Under FLH, it would be more cognitively demanding to process f_0 contours that simultaneously carry more than one type of function.

The FLH predictions provide indirect explanations for *disambiguation,* as presumably, competition across diverse f_0 functions may facilitate their recognition, disentanglement and establishment of f_0 categories. These predictions also offer viable ways of empirically examining FLH as a hypothesis. Having said that, existing findings are mixed (van Heuven, 2018). FLH is supported by studies investigating parameters competing within the prosodic domain. Supported by phonological and acoustic analyses, Remijsen (2002) has shown that it is highly unlikely for a tone language to feature lexical stress because that would create competition (and thus ambiguity) between the pragmatic and the lexical functions of f_0. Using phonological and acoustic analyses, Remijsen (2002) concluded that it is implausible for lexical tone and lexically contrastive stress accent to co-exist in the word-prosodic system of a language. Nevertheless, challenges appear to lie in *interaction*: FLH would need to explain how parameters from different domains within phonology (e.g., prosodic vs. segmental domains) and beyond (e.g., linguistic vs. paralinguistic domains) compete against one another. In other words, it is unclear whether and to what extent information across domains and modalities fights for cognitive resources during processing. FLH concentrates on the linguistic domain and the emotional aspect has not been directly considered (although it was alluded to in Chen, 2005). Nevertheless, the FLH postulation seems to imply that languages encoding f_0 in both tone and intonation would have less functional space left to encode f_0 in emotions. Note that caregivers may assist, consciously or unconsciously, in the reduction of functional loads in the course of infants' learning. For instance, they may package messages in IDS to reduce processing challenges for certain f_0 functions.

The FLH faces challenges incorporating cross-domain or cross-modal facilitation effects. That is, information perceived in one domain (e.g., vision) may support perception and learning of information in another domain (e.g., speech). These are often referred to as bootstrapping or anchoring effects. For instance, the prosodic bootstrapping hypothesis suggests that infants may use prosodic information to discover utterance and word boundaries (Seidl and Johnson, 2006; Johnson et al., 2014), and knowledge of word semantics may further cue syntactic categories (Höhle, 2009). Along the same lines, various sources of information from the ambient environment provide anchors to facilitate children's f_0 *disambiguation* and *categorisation* along the developmental trajectory. The command of one f_0 function may facilitate another even when they are simultaneously presented. FLH, or any cognitive model, will need to clearly explain the degree of interaction between competition and facilitation in co-occurring functions.

What has not been discussed, but links closely with the FLH mechanisms, is how infants cope with cognitive demands and how increased neurocognitive ability affects children's perception and learning. It takes children years to master linguistic and pragmatic functions. Taking *Theory of Mind* (ToM) as an example, ToM refers to the understanding of distinctions between individuals' mental states, mental constructs, physical entities and their overt actions (Gopnik and Wellman, 1992; Wellman, 1992). ToM is crucial for children's socio-emotional development. What needs to be explored is how children's gross and specific (e.g., socio-emotions) cognitive development attributes to the learning of emotional f_0.

Statistical Learning

Statistical learning refers to the ability to acquire information solely based on relevant statistical distributions in the ambient environment, and *Statistical Learning* accounts argue that infants utilise their innate statistical (Saffran and Kirkham, 2018) and relational (Ferry et al., 2015) learning ability to acquire new information. For instance, 8-month-olds are able to segment words from fluent speech based on one and only one cue: the statistical relationships between neighbouring syllables (Saffran et al., 1996; but see Johnson and Tyler, 2010).

While statistical learning accounts have been used to describe acquisition of a single f_0 function (e.g., lexical tone, Liu and Kager, 2017b), its explanatory power faces evident challenges in *disambiguation* and *categorisation*. A purely bottom-up learning of a statistical distribution does not appear sufficient to explain *disambiguation* if f_0 is the only statistical distribution available. By comparison, vowels may be disambiguated on the basis of multiple information sources (e.g., the first, second, and third formants, and duration). Even though f_0 serves as the primary acoustic correlate of emotional tones (Scherer, 2003), its usage differs between tone and non-tone language speakers, with greater f_0 variations in the productions of the latter group. It seems likely that statistical learning of f_0 patterns would require correlated statistical distributions from other information sources. This may include phonation type (e.g., creaky voice) or tone-vowel interactions (Shaw and Tyler, 2020) for tone, and voice quality for both intonation (Ladd et al., 1985) and emotion (Yanushevskaya et al., 2018). Cue-weighting, or differences in listeners' weighting of acoustic cues (e.g., between f_0 and secondary cues such as amplitude and duration, Ross et al., 1986), likely further modulates statistical learning.

With respect to *categorisation*, statistical learning ability does not appear to be constant across ages. Its efficacy changes dynamically over a child's development. However, the direction of such change, or the statistical learning efficacy across ages, is currently a matter of debate. On the one hand, a meta-analysis has reported increased effect sizes with age in the first year of life (Cristia, 2018), suggesting that older infants are increasingly sensitive to this learning mechanism. On the other hand, behavioural (Yoshida et al., 2010) and neural (Wanrooij et al., 2014) evidence has shown that this learning mechanism may be maturationally delimited, along the perceptual attunement trajectory during which phonetic perception is refined (Liu and Kager, 2017b; Reh et al., 2021). The latter evidence suggests that the learning of sound frequency distributions become increasingly resistant as children grow. Discrepancies in literature have been explained by the different perceptual attunement time windows of speech sounds differing in phonetic representations, space and perceptual/acoustic salience (Werker and Hensch, 2015; Reh et al., 2021). Hence, statistical learning of speech sounds may be at its peak of efficacy during perceptual attunement, when infants' perception exhibits enhanced sensitivity to input from the environment.

Although the learning mechanism is considered domain- (and even species-) general, individual studies and models typically investigate statistical learning in a domain-specific fashion. Despite the challenge in *disambiguation* and the debate in *categorisation*, in order to achieve learning of diverse f_0 functions, models of statistical learning would require additional focus on the *interaction* mechanisms, with modelling of certain (e.g., f_0) statistical distributions across domains.

Summary

Similar to the lack of empirical research in studying the interaction of distinct linguistic and paralinguistic functions carried on f_0, none of the existing models and hypotheses seems sufficient in addressing how different f_0 functions disassociate in sensory and cognitive processes, or the extent to which they are processed simultaneously or separately. A theoretical account is required for how infants manage to decompose these overlapping f_0 functions while taking into consideration the differences between these functions across languages/cultures, as well as information integration across modalities.

As summarised in the beginning of this section, to achieve successful learning, infants must rely on fundamental aspects (disambiguation, categorisation, accommodation and interaction). These aspects point out directions where the exploration of diverse f_0 functions may converge. These directions are crucial for us to understand how infants resolve puzzles identified in the literature:

- *Neuro-cognitive Development*, which reflects age-related developmental and maturational changes
- *Environmental Information*, where learning of language and social-emotions from the ambient resources occur
- *Competition and Facilitation*, within and across perceptual and/or cognitive spaces and modalities (e.g., auditory, visual) where information gathers and integrates

Infants eventually sort out their native linguistic and socio-emotional functions carried on f_0. Thus, developmental and environmental aspects such as age and experience will need to be considered when exploring f_0 functions, in line with the first two directions. With respect to category learning, future research should focus on the establishment of f_0 categories for tone, intonation, and emotion. Further that, the degree of flexibility and assimilation when facing a novel/non-native category will need to be explored. Regarding bootstrapping, a theoretical basis will require that infants effectively integrate environmental sources of information and existing knowledge to recognise and disambiguate f_0 functions.[1] A multimodal view into the issue is also consistent with an ecological approach to perceptual learning and development (e.g., Gibson and Pick, 2000).

To summarise across the four directions, future research should concentrate on how infants decompose and acquire linguistic and paralinguistic functions carried on f_0; to what extent reinforcement or interference may occur with infants' perception and learning of f_0 functions; and how infants employ environmental resources to disambiguate these functions.

Our Hypothesis

Considering the gap in discussion of f_0 functions across linguistic and socio-emotional domains, the four aspects concerning f_0 perception and development, and the four directions essential to achieve its functional learning, we propose a *Learnability Hypothesis* that infants require multimodal environmental experiences to decompose and acquire overlapping linguistic, paralinguistic, and extralinguistic f_0 functions. Its predictions are as follows: When faced with f_0 contours carrying multiple functions, perception and learning of a certain function should be enhanced if other functions are not ambiguous, and should be affected if other functions have not been properly learned or cannot be properly identified. Moreover, infants use acquired, environmental and multi-modal cues to anchor and facilitate learning whenever possible.

A representative and measurable case of the learnability hypothesis can be viewed as the "synergy scenario." For example, infants can use visual cues to disambiguate and decompose different auditory f_0 functions. Congruent audiovisual cues of the same function will lead to corresponding enhancements as well as reduced sensitivity to others. In contrast, incongruent cues may capture infants' attention, as is the case for deviants against standards in an oddball paradigm in electroencephalogram, or regained attention to new information in a behavioural habituation paradigm. These predictions provide us with viable ways of testing the hypothesis.

One way to examine this scenario would be to use an experimental paradigm that reflects the real world lives and interactions that infants experience, such as using stimuli that mirror real communications that occur in infant-caregiver interactions. Following an associative learning paradigm (Hay et al., 2015; Liu and Kager, 2018), infants' ability to associate novel objects with an instructor's f_0s that represent tones could be measured with or without the instructor's visual intonational and emotional information. Here, the f_0s could be ambiguous, not only reflecting tonal but also intonational or emotional f_0 that are relevant in infants' native environment. In this case, when the presented visual information matched intonational or emotional f_0, infants should show a reduction in associative learning.

CONCLUSION

A diverse array of linguistic and paralinguistic functions are carried simultaneously on f_0. Patterns of f_0 variability differ across languages, such that an f_0 pattern that serves a particular function in one language may serve a different function in another. Adults use native f_0 functions effortlessly, but how infants acquire them remains a mystery. Infants' unstable learning trajectories raise important questions. For instance, when they no longer treat f_0 differences as potential signals to a change in a certain function, is it due to an insensitivity to f_0 features or due to those features being used for a different communicative purpose? Do infants adopt top-down or bottom-up processing when disambiguating different functions carried on the same f_0? These questions surface from the mixed findings in the literature, across tone, intonation, and emotional domains.

It is important to seek answers to these questions and solutions to the discrepancies observed in the literature. The body of literature needs to be expanded to include infants from a broader range of language environments so that we can understand the course of acquisition. Obtaining the answers through a theoretical and empirical approach, such as the research ideas spawned by our *Learnability Hypothesis*, will improve and integrate theories across research fields, especially when existing models do not appear sufficiently inclusive to address the learning process.

The early years of life lay solid foundations for child learning, assisting our young learners to navigate through the complexities of our modern world. The understanding of how children command the multiple f_0 functions using an ecological approach will function as a benchmark guiding pitch learning in the natural environment; help with the identification of speech or cognitive impairments; better support typical child development; and contribute to multilingual/vulnerable language learning, second/foreign language learning, as well as learning across the lifespan.

AUTHOR CONTRIBUTIONS

LL and MT drafted and revised the manuscript. AG and PL revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

LL's writing was partially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 798658 hosted by Center for Multilingualism across the Lifespan at the University of Oslo, financed by Research Council of Norway through its Centers of Excellence funding scheme grant agreement No. 223265. The open access publication fee was supported by Western Sydney University.

[1] Recent evidence suggests that there may be information about f_0 in the face and in head movements that can be used to discriminate lexical tone contrasts (Burnham et al., 2022), but it is not clear from these findings whether such auditory visual speech information would be useful for disambiguating different f_0 functions. Here we consider the role of non-speech environmental information on the acquisition of f_0 functions.

REFERENCES

Anolli, L., Wang, L., Mantovani, F., and De Toni, A. (2008). The voice of emotion in Chinese and Italian young adults. *J. Cross-Cult. Psychol.* 39, 565–598. doi: 10.1177/0022022108321178

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.

Banse, R., and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *J. Pers. Soc. Psychol.* 70, 614–636. doi: 10.1037/0022-3514.70.3.614

Bänziger, T., Hosoya, G., and Scherer, K. R. (2015). Path models of vocal emotion communication. *PLoS One* 10:e0136675. doi: 10.1371/journal.pone.0136675

Belyk, M., and Brown, S. (2014). Perception of affective and linguistic prosody: An ALE meta-analysis of neuroimaging studies. *Soc. Cogn. Affect. Neurosci.* 9, 1395–1403. doi: 10.1093/scan/nst124

Berinstein, A. E. (1979). A cross-linguistic Study on the Perception and Production of Stress Unpublished master's dissertation. University of California Los Angeles.

Best, C. T. (1994). "The emergence of native-language phonological influences in infants: A perceptual assimilation model," in *The Development of speech Perception: The transition from speech Sounds to Spoken Words*. eds. J. C. Goodman and H. C. Nusbaum (United States: MIT Press), 167–244.

Best, C. T. (2015). "Devil or angel in the details? Perceiving phonetic variation as information about phonological structure," in *Phonetics-phonology interface: Representations and methodologies*. eds. J. Romero and M. Riera (Amsterdam: John Benjamins), 3–31.

Best, C. T. (2019). The diversity of tone languages and the roles of pitch variation in non-tone languages: considerations for tone perception research. *Front. Psychol.* 10:364. doi: 10.3389/fpsyg.2019.00364

Best, C. T., McRoberts, G. W., and Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *J. Exp. Psychol. Hum. Percept. Perform.* 14, 345–360. doi: 10.1037/0096-1523.14.3.345

Best, C. T., Traill, A., Carter, A., Harrison, K. D., and Faber, A. (2003). "!Xóõ Click Perception by English, Isizulu, and Sesotho Listeners." in *Proceedings of the 15th International Congress of Phonetic Sciences.* eds. M. J. Solé, D. Recasens and J. Romero; August 3–9, 2003; Causal Productions; 853–856.

Best, C. T., Tyler, M. D., Gooding, T. N., Orlando, C. B., and Quann, C. A. (2009). Development of phonological constancy: Toddlers' perception of native-and Jamaican-accented words. *Psychol. Sci.* 20, 539–542. doi: 10.1111/j.1467-9280.2009.02327.x

Braun, B., and Johnson, E. K. (2011). Question or tone 2? How language experience and linguistic function guide pitch processing. *J. Phon.* 39, 585–594. doi: 10.1016/j.wocn.2011.06.002

Bryant, G. A. (2021). The evolution of human vocal emotion. *Emot. Rev.* 13, 25–33. doi: 10.1177/1754073920930791

Burnham, D. K. (1986). Developmental loss of speech perception: exposure to and experience with a first language1. *Appl. Psycholinguist.* 7, 207–239. doi: 10.1017/S0142716400007542

Burnham, D., Kitamura, C., and Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science* 296:1435. doi: 10.1126/science.1069587

Burnham, D. K., and Singh, L. (2018). Coupling tonetics and perceptual attunement: The psychophysics of lexical tone contrast salience. *J. Acoust. Soc. Am.* 144:1716. doi: 10.1121/1.5067611

Burnham, D., Vatikiotis-Bateson, E., Barbosa, A. V., Menezes, J. V., Yehia, H. C., Morris, R. H., et al. (2022). Seeing lexical tone: head and face motion in production and perception of Cantonese lexical tones. *Speech Comm.* 141, 40–55. doi: 10.1016/j.specom.2022.03.011

Cahn, J. E. (1990). The generation of affect in synthesized speech. *J. Am. Voice I/O Soc.* 8:1

Caron, R. F., Caron, A. J., and Myers, R. S. (1985). Do infants see emotional expressions in static faces? *Child Dev.* 56, 1552–1560. doi: 10.2307/1130474

Chao, Y. R. (1947). *Cantonese Primer*. United States: Harvard University Press.

Chen, A. (2005). *Universal and Language-specific Perception of Paralinguistic Intonational Meaning*. Utrecht: LOT.

Chen, A., and Kager, R. (2016). Discrimination of lexical tones in the first year of life. *Infant Child Dev.* 25, 426–439. doi: 10.1002/icd.1944

Chen, L. M., and Kent, R. D. (2009). Development of prosodic patterns in mandarin-learning infants. *J. Child Lang.* 36, 73–84. doi: 10.1017/S0305000908008878

Chen, A., Liu, L., and Kager, R. (2015). Cross-linguistic perception of mandarin tone sandhi. *Lang. Sci.* 48, 62–69. doi: 10.1016/j.langsci.2014.12.002

Chen, A., Stevens, C. J., and Kager, R. (2017). Pitch perception in the first year of life, a comparison of lexical tones and musical pitch. *Front. Psychol.* 8:297. doi: 10.3389/fpsyg.2017.00297

Chien, P. J., Friederici, A. D., Hartwigsen, G., and Sammler, D. (2020). Neural correlates of intonation and lexical tone in tonal and non-tonal language speakers. *Hum. Brain Mapp.* 41, 1842–1858. doi: 10.1002/hbm.24916

Chong, S. C. F., Werker, J., Russell, J. A., and Carroll, J. M. (2003). Three facial expressions mothers direct to their infants. *Infant Child Dev.* 12, 211–232. doi: 10.1002/icd.286

Cristia, A. (2018). Can infants learn phonology in the lab? A meta-analytic answer. *Cognition* 170, 312–327. doi: 10.1016/j.cognition.2017.09.016

Crystal, D., and Quirk, R. (1964). *Systems of Prosodic and Paralinguistic Features in English*. Berlin: De Gruyter Mouton.

Dachkovsky, S., and Sandler, W. (2009). Visual intonation in the prosody of a sign language. *Lang. Speech* 52, 287–314. doi: 10.1177/0023830909103175

Dijkstra, C., Krahmer, E., and Swerts, M. (2006). "Manipulating uncertainty: The contribution of different audiovisual prosodic cues to the perception of confidence." in *Proceedings of the Speech Prosody Conference, Dresden*. May 2–5, 2006.

Doan, S. N. (2010). The role of emotion in word learning. *Early Child Dev. Care* 180, 1065–1078. doi: 10.1080/03004430902726479

Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech: towards a new generation of databases. *Speech Comm.* 40, 33–60. doi: 10.1016/S0167-6393(02)00070-5

Escudero, P. (2005). *Linguistic Perception and second Language Acquisition: Explaining the Attainment of Optimal Phonological Categorization*. Netherlands: Netherlands Graduate School of Linguistics.

Esteve-Gibert, N., Prieto, P., and Liszkowski, U. (2017). Twelve-month-olds understand social intentions based on prosody and gesture shape. *Infancy* 22, 108–129. doi: 10.1111/infa.12146

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behav. Dev.* 8, 181–195. doi: 10.1016/S0163-6383(85)80005-9

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: is the melody the message? *Child Dev.* 60, 1497–1510. doi: 10.2307/1130938

Fernald, A. (1992). "Human maternal vocalizations to infants as biologically relevant signals," in *The Adapted mind: Evolutionary Psychology and the Generation of Culture*. eds. J. Barkow, L. Cosmides and J. Tooby (Oxford, England: Oxford University Press), 391–428.

Fernald, A. (1993). Approval and disapproval: infant responsiveness to vocal affect in familiar and unfamiliar languages. *Child Dev.* 64, 657–674. doi: 10.2307/1131209

Fernald, A., and Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behav. Dev.* 10, 279–293. doi: 10.1016/0163-6383(87)90017-8

Fernald, A., and Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Dev. Psychol.* 20, 104–113. doi: 10.1037/0012-1649.20.1.104

Ferry, A. L., Hespos, S. J., and Gentner, D. (2015). Prelinguistic relational concepts: investigating analogical processing in infants. *Child Dev.* 86, 1386–1405. doi: 10.1111/cdev.12381

Fikkert, P., Liu, L., and Ota, M. (2020). "The acquisition of word prosody," in *The Oxford Handbook of Language Prosody* (London: Oxford University Press), 541–552.

Francis, A. L., Ciocca, V., Ma, L., and Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *J. Phon.* 36, 268–294. doi: 10.1016/j.wocn.2007.06.005

Frick, R. W. (1985). Communicating emotion: The role of prosodic features. *Psychol. Bull.* 97, 412–429. doi: 10.1037/0033-2909.97.3.412

Fromkin, V. A. (Ed.). (2014). *Tone: A linguistic Survey*. United States: Academic Press.

Frota, S., and Butler, J. (2018). "Early development of intonation," in *The Development of Prosody in first Language Acquisition*. eds. P. Prieto and N. Esteve-Gibert (Netherland: John Benjamins), 145–164.

Frota, S., Butler, J., and Vigário, M. (2014). Infants' perception of intonation: is it a statement or a question? *Infancy* 19, 194–213. doi: 10.1111/infa.12037

Frühholz, S., Ceravolo, L., and Grandjean, D. (2012). Specific brain networks during explicit and implicit decoding of emotional prosody. *Cereb. Cortex* 22, 1107–1117. doi: 10.1093/cercor/bhr184

Gandour, J. (1983). Tone perception in far eastern languages. *J. Phon.* 11, 149–175. doi: 10.1016/S0095-4470(19)30813-7

Geangu, E., Ichikawa, H., Lao, J., Kanazawa, S., Yamaguchi, M. K., Caldara, R., et al. (2016). Culture shapes 7-month-olds' perceptual strategies in discriminating facial expressions of emotion. *Curr. Biol.* 26, R663–R664. doi: 10.1016/j.cub.2016.05.072

Gendron, M., Crivelli, C., and Barrett, L. F. (2018). Universality reconsidered: diversity in making meaning of facial expressions. *Curr. Dir. Psychol. Sci.* 27, 211–219. doi: 10.1177/0963721417746794

Gendron, M., Roberson, D., and Barrett, L. F. (2015). Cultural variation in emotion perception is real: A response to Sauter, Eisner, Ekman, and Scott (2015). *Psychol. Sci.* 26, 357–359. doi: 10.1177/0956797614566659

Gendron, M., Roberson, D., van der Vyver, J. M., and Barrett, L. F. (2014). Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion* 14, 251–262. doi: 10.1037/a0036052

Gibson, E. J., and Pick, A. D. (2000). *An Ecological Approach to Perceptual Learning and Development*. England: Oxford University Press.

Gobl, C., Bennett, E., and Chasaide, A. N. (2002). Expressive synthesis: how crucial is voice quality?. In *Proceedings of 2002 IEEE Workshop on Speech Synthesis*. September 11–13, 2002; IEEE; 91–94.

Gobl, C., and Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Comm.* 40, 189–212. doi: 10.1016/S0167-6393(02)00082-1

Goldbeck, T., Tolkmitt, F., and Scherer, K. R. (1988). "Experimental studies on vocal affect communication," in *Facets of Emotion: Recent Research*. ed. K. R. Scherer (Mahwah: Lawrence Erlbaum), 119–137.

Gopnik, A., and Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind Lang.* 7, 145–171. doi: 10.1111/j.1468-0017.1992.tb00202.x

Götz, A., Yeung, H. H., Krasotkina, A., Schwarzer, G., and Höhle, B. (2018). Perceptual reorganization of lexical tones: effects of age and experimental procedure. *Front. Psychol.* 9:477. doi: 10.3389/fpsyg.2018.00477

Goudbeek, M., and Scherer, K. (2010). Beyond arousal: valence and potency/control cues in the vocal expression of emotion. *J. Acoust. Soc. Am.* 128, 1322–1336. doi: 10.1121/1.3466853

Graf Estes, K., and Hay, J. F. (2015). Flexibility in bilingual infants' word learning. *Child Dev.* 86, 1371–1385. doi: 10.1111/cdev.12392

Grice, M., Baumann, S., and Benzmüller, R. (2006). *Autosegmental-Metrical Phonology*. Prosodic typology: The phonology of intonation and phrasing, 55–83.

Grossmann, T., Striano, T., and Friederici, A. D. (2005). Infants' electric brain responses to emotional prosody. *Neuroreport* 16, 1825–1828. doi: 10.1097/01.wnr.0000185964.34336.b1

Gussenhoven, C. (2002). "Intonation and interpretation: phonetics and phonology." in *Proceedings of the 1st International Conference on Speech Prosody*. April 11–13, 2002; 47–57.

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

Gussenhoven, C., and Chen, A. (2000). Universal and language-specific effects in the perception of question intonation. In *6th International Conference on Spoken Language Processing (ICSLP 2000)* 91–94.

Haviland, J. M., and Lelwica, M. (1987). The induced affect response: 10-week-old infants' responses to three emotion expressions. *Dev. Psychol.* 23, 97–104. doi: 10.1037/0012-1649.23.1.97

Hay, J. F., Cannistraci, R. A., and Zhao, Q. (2019). Mapping non-native pitch contours to meaning: perceptual and experiential factors. *J. Mem. Lang.* 105, 131–140. doi: 10.1016/j.jml.2018.12.004

Hay, J. F., Graf Estes, K., Wang, T., and Saffran, J. R. (2015). From flexibility to constraint: The contrastive use of lexical tone in early word learning. *Child Dev.* 86, 10–22. doi: 10.1111/cdev.12269

He, C., Hotson, L., and Trainor, L. J. (2007). Mismatch responses to pitch changes in early infancy. *J. Cogn. Neurosci.* 19, 878–892. doi: 10.1162/jocn.2007.19.5.878

Hoemann, K., Xu, F., and Barrett, L. F. (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Dev. Psychol.* 55, 1830–1849. doi: 10.1037/dev0000686

Höhle, B. (2009). Bootstrapping mechanisms in first language acquisition. *Linguistics* 47, 359–382. doi: 10.1515/LING.2009.013

Hua, Z. (2002). *Phonological Development in specific Contexts: Studies of Chinese-Speaking children. Vol. 3* United Kingdom: Multilingual Matters.

Hua, Z., and Dodd, B. (2000). The phonological acquisition of Putonghua (modern standard Chinese). *J. Child Lang.* 27, 3–42. doi: 10.1017/S030500099900402X

Hyman, L. M. (2011). "Tone: is it different?" in *The Handbook of Phonological Theory. Vol. 75*. eds. J. A. Goldsmith, J. Riggle and A. C. L. Yu (United States: John Wiley and Sons), 50–80.

Hyman, L. M. (2016). Lexical vs. grammatical tone: sorting out the differences. In *proceedings of the 5th international symposium on tonal aspects of languages (TAL 2016)*. May 24–27, 2016; 6–11.

Ilie, G., and Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music. Percept.* 23, 319–330. doi: 10.1525/mp.2006.23.4.319

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *J. Acoust. Soc. Am.* 88, 642–654. doi: 10.1121/1.399767

Johnson, E. K., Seidl, A., and Tyler, M. D. (2014). The edge factor in early word segmentation: utterance-level prosody enables word form extraction by 6-month-olds. *PLoS One* 9:e83546. doi: 10.1371/journal.pone.0083546

Johnson, E. K., and Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Dev. Sci.* 13, 339–345. doi: 10.1111/j.1467-7687.2009.00886.x

Jongman, A., Wang, Y., Moore, C., and Sereno, J. (2006). "Perception and production of mandarin tone," in *Handbook of East Asian Psycholinguistics. Vol. 1*. eds. P. Li, L. H. Tan, E. Bates and O. J. L. Tzeng (England: Cambridge University Press), 209–217.

Katz, G. S., Cohn, J. F., and Moore, C. A. (1996). A combination of vocal f0 dynamic and summary features discriminates between three pragmatic categories of infant-directed speech. *Child Dev.* 67, 205–217. doi: 10.1111/j.1467-8624.1996.tb01729.x

Keitel, A., Prinz, W., Friederici, A. D., Von Hofsten, C., and Daum, M. M. (2013). Perception of conversations: The importance of semantics and intonation in children's development. *J. Exp. Child Psychol.* 116, 264–277. doi: 10.1016/j.jecp.2013.06.005

Kent, R. D., and Murray, A. D. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *J. Acoust. Soc. Am.* 72, 353–365. doi: 10.1121/1.388089

Kitamura, C., and Burnham, D. (2003). Pitch and communicative intent in mother's speech: adjustments for age and sex in the first year. *Infancy* 4, 85–110. doi: 10.1207/S15327078IN0401_5

Kitamura, C., and Lam, C. (2009). Age-specific preferences for infant-directed affective intent. *Infancy* 14, 77–100. doi: 10.1080/15250000802569777

Koeda, M., Belin, P., Hama, T., Masuda, T., Matsuura, M., and Okubo, Y. (2013). Cross-cultural differences in the processing of non-verbal affective vocalizations by Japanese and Canadian listeners. *Front. Psychol.* 4:105. doi: 10.3389/fpsyg.2013.00105

Kotz, S. A., Meyer, M., and Paulmann, S. (2006). Lateralization of emotional prosody in the brain: An overview and synopsis on the impact of study design. *Prog. Brain Res.* 156, 285–294. doi: 10.1016/S0079-6123(06)56015-7

Kotz, S. A., and Paulmann, S. (2007). When emotional prosody and semantics dance cheek to cheek: ERP evidence. *Brain Res.* 1151, 107–118. doi: 10.1016/j.brainres.2007.03.015

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., and Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Phil. Trans. R. Soc. B: Biol. Sci.* 363, 979–1000. doi: 10.1098/rstb.2007.2154

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

Ladd, D. R., Silverman, K. E., Tolkmitt, F., Bergmann, G., and Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *J. Acoust. Soc. Am.* 78, 435–444. doi: 10.1121/1.392466

Laukka, P., Juslin, P., and Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognit. Emot.* 19, 633–653. doi: 10.1080/02699930441000445

Lee, C. C., Jhang, Y., Chen, L. M., Relyea, G., and Oller, D. K. (2017). Subtlety of ambient-language effects in babbling: a study of English-and Chinese-learning infants at 8, 10, and 12 months. *Lang. Learn. Dev.* 13, 100–126. doi: 10.1080/15475441.2016.1180983

Lee, C. Y., and Lee, Y. F. (2010). Perception of musical pitch and lexical tones by mandarin-speaking musicians. *J. Acoust. Soc. Am.* 127, 481–490. doi: 10.1121/1.3266683

Li, A. (2015). *Encoding and Decoding of Emotional speech: A cross-Cultural and Multimodal Study between Chinese and Japanese*. United States: Springer.

Li, C. N., and Thompson, S. A. (1977). The acquisition of tone in mandarin-speaking children. *J. Child Lang.* 4, 185–199. doi: 10.1017/S0305000900001598

Liang, J., and Heuven, V. J. (2007). Chinese tone and intonation perceived by L1 and L2 listeners. In C. Gussenhoven and T. Riad (Eds.), *Tones and Tunes, Vol. 2: Experimental Studies in Word and Sentence Prosody*. Berlin: De Gruyter Mouton. (27–62)

Liebenthal, E., Silbersweig, D. A., and Stern, E. (2016). The language, tone and prosody of emotions: neural substrates and dynamics of spoken-word emotion perception. *Front. Neurosci.* 10:506. doi: 10.3389/fnins.2016.00506

Liu, L., du Toit, M., and Weidemann, G. (2021). Infants are sensitive to cultural differences in emotions at 11 months. *PLoS One* 16:e0257655. doi: 10.1371/journal.pone.0257655

Liu, L., and Kager, R. (2014). Perception of tones by infants learning a non-tone language. *Cognition* 133, 385–394. doi: 10.1016/j.cognition.2014.06.004

Liu, L., and Kager, R. (2015). Bilingual exposure influences infant VOT perception. *Infant Behavior and Development* 38, 27–36.

Liu, L., and Kager, R. (2017a). Perception of tones by bilingual infants learning non-tone languages. *Biling. Lang. Congn.* 20, 561–575. doi: 10.1017/S1366728916000183

Liu, L., and Kager, R. (2017b). Statistical learning of speech sounds is most robust during the period of perceptual attunement. *J. Exp. Child Psychol.* 164, 192–208. doi: 10.1016/j.jecp.2017.05.013

Liu, L., and Kager, R. (2018). Monolingual and bilingual infants' ability to use non-native tone for word learning deteriorates by the second year after birth. *Front. Psychol.* 9:117. doi: 10.3389/fpsyg.2018.00117

Liu, L., Lai, R., Singh, L., Kalashnikova, M., Wong, P. C. M., Kasisopa, B., et al. (2022). The tone atlas of perceptual discriminability and perceptual distance: Four tone languages and five language groups. *Brain Lang.* 229:105106. doi: 10.1016/j.bandl.2022.105106

Liu, P., and Pell, M. D. (2012). Recognizing vocal emotions in mandarin Chinese: a validated database of Chinese vocal emotional stimuli. *Behav. Res. Methods* 44, 1042–1051. doi: 10.3758/s13428-012-0203-3

Liu, L., Varghese, P., and Weidemann, G. (2019). "A bilingual advantage in infant pitch processing." in *Proceedings of the 19th International Congress of Phonetic Sciences*. August 5–9, 2019; International Phonetic Association; 1397–1401.

Liu, L., and Xu, R. (2005). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica* 62, 70–87.

Maddieson, I. (2013). "Tone" in *The World Atlas of Language Structures Online*. eds. M. S. Dryer, S. Matthew and M. Haspelmath (Germany: Max Planck Institute for Evolutionary Anthropology).

Mampe, B., Friederici, A. D., Christophe, A., and Wermke, K. (2009). Newborns' cry melody is shaped by their native language. *Curr. Biol.* 19, 1994–1997. doi: 10.1016/j.cub.2009.09.064

Manfredi, C., Viellevoye, R., Orlandi, S., Torres-García, A., Pieraccini, G., and Reyes-García, C. A. (2019). Automated analysis of newborn cry: relationships between melodic shapes and native language. *Biomed. Sig. Proces. Cont.* 53:101561. doi: 10.1016/j.bspc.2019.101561

ManyBabies Consortium (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Adv. Methods Pract. Psychol. Sci.* 3, 24–52. doi: 10.1177/2515245919900809

Massaro, D. W., Cohen, M. M., and Tseng, C. Y. (1985). The evaluation and integration of pitch height and pitch contour in lexical tone perception in mandarin Chinese. *J. Chinese Ling.* 267–289.

Mattock, K., and Burnham, D. (2006). Chinese and English infants' tone perception: evidence for perceptual reorganization. *Infancy* 10, 241–265. doi: 10.1207/s15327078in1003_3

Mattock, K., Molnar, M., Polka, L., and Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition* 106, 1367–1381. doi: 10.1016/j.cognition.2007.07.002

Moore, D. S., Spence, M. J., and Katz, G. S. (1997). Six-month-olds' categorization of natural infant-directed utterances. *Dev. Psychol.* 33, 980–989. doi: 10.1037/0012-1649.33.6.980

Mozziconacci, S. J. L. (1998). *Speech Variability and Emotion: Production and Perception*. Netherlands: Technical University Eindhoven.

Murray, I. R., and Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *J. Acoust. Soc. Am.* 93, 1097–1108. doi: 10.1121/1.405558

Nazzi, T., Floccia, C., and Bertoncini, J. (1998). Discrimination of pitch contours by neonates. *Infant Behav. Dev.* 21, 779–784. doi: 10.1016/S0163-6383(98)90044-3

Nelson, C. A., and De Haan, M. (1996). Neural correlates of infants' visual responsiveness to facial expressions of emotion. *Dev. Psychobiol.* 29, 577–595. doi: 10.1002/(SICI)1098-2302(199611)29:7<577::AID-DEV3>3.0.CO;2-R

Ortony, A., Clore, G. L., and Collins, A. (1990). *The Cognitive Structure of Emotions*. England: Cambridge University Press.

Papoušek, M. (1989). Determinants of responsiveness to infant vocal expression of emotional state. *Infant Behav. Dev.* 12, 507–524. doi: 10.1016/0163-6383(89)90030-1

Pell, M. D., and Kotz, S. A. (2011). On the time course of vocal emotion recognition. *PLoS One* 6:e27256. doi: 10.1371/journal.pone.0027256

Pell, M. D., Monetta, L., Paulmann, S., and Kotz, S. A. (2009). Recognizing emotions in a foreign language. *J. Nonverbal Behav.* 33, 107–120. doi: 10.1007/s10919-008-0065-7

Pell, M. D., Rothermich, K., Liu, P., Paulmann, S., Sethi, S., and Rigoulot, S. (2015). Preferential decoding of emotion from human non-linguistic vocalizations versus speech prosody. *Biol. Psychol.* 111, 14–25. doi: 10.1016/j.biopsycho.2015.08.008

Peng, G., and Chen, F. (2020). "Speech development in mandarin-speaking children," in *Speech Perception, Production and Acquisition: Multidisciplinary Approaches in Chinese Languages*. eds. H. Liu, F. Tsao and P. Li (United States: Springer)

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

Polka, L., and Bohn, O. S. (2011). Natural referent vowel (NRV) framework: an emerging view of early phonetic development. *J. Phon.* 39, 467–478. doi: 10.1016/j.wocn.2010.08.007

Prochnow, A., Erlandsson, S., Hesse, V., and Wermke, K. (2019). Does a 'musical' mother tongue influence cry melodies? A comparative study of Swedish and German newborns. *Music. Sci.* 23, 143–156. doi: 10.1177/1029864917733035

Quam, C., and Swingley, D. (2010). Phonological knowledge guides 2-year-olds' and adults' interpretation of salient pitch contours in word learning. *J. Mem. Lang.* 62, 135–150. doi: 10.1016/j.jml.2009.09.003

Quam, C., and Swingley, D. (2012). Development in children's interpretation of pitch cues to emotions. *Child Dev.* 83, 236–250. doi: 10.1111/j.1467-8624.2011.01700.x

Ramachers, S., Brouwer, S., and Fikkert, P. (2018). No perceptual reorganization for Limburgian tones? A cross-linguistic investigation with 6-to 12-month-old infants. *J. Child Lang.* 45, 290–318. doi: 10.1017/S0305000917000228

Reh, R. K., Dias, B. G., Nelson, C. A., Kaufer, D., Werker, J. F., Kolb, B., et al. (2020). Critical period regulation across multiple timescales. *Proc. Natl. Acad. Sci.* 117, 23242–23251. doi: 10.1073/pnas.1820836117

Reh, R. K., Hensch, T. K., and Werker, J. F. (2021). Distributional learning of speech sound categories is gated by sensitive periods. *Cognition* 213:104653. doi: 10.1016/j.cognition.2021.104653

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science* 212, 947–950. doi: 10.1126/science.7233191

Remijsen, B. (2002). "Lexically contrastive stress accent and lexical tone in Ma'ya," in *Laboratory Phonology 7* (Berlin: De Gruyter Mouton), 585–614.

Remijsen, B. (2016). "Tone," in *The Oxford Research Encyclopedia of Linguistics Online* (Oxford: Oxford University Press).

Ren, G. Q., Tang, Y. Y., Li, X. Q., and Sui, X. (2013). "Pre-attentive processing of mandarin tone and intonation: evidence from event-related potentials," in *Functional brain Mapping and the Endeavor to Understand the Working brain*. eds. F. Signorelli and D. Chirchiglia (Austria: Intech), 95–108.

Roseano, P., González, M., Borràs-Comes, J., and Prieto, P. (2016). Communicating epistemic stance: how speech and gesture patterns reflect epistemicity and evidentiality. *Discourse Process.* 53, 135–174. doi: 10.1080/0163853X.2014.969137

Ross, E. D., Edmondson, J. A., and Seibert, G. B. (1986). The effect of affect on various acoustic measures of prosody in tone and non-tone languages: a comparison based on computer analysis of voice. *J. Phon.* 14, 283–302. doi: 10.1016/S0095-4470(19)30669-2

Ruba, A. L., Meltzoff, A. N., and Repacholi, B. M. (2020). Superordinate categorization of negative facial expressions in infancy: The influence of labels. *Dev. Psychol.* 56, 671–685. doi: 10.1037/dev0000892

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* 274, 1926–1928. doi: 10.1126/science.274.5294.1926

Saffran, J. R., and Kirkham, N. Z. (2018). Infant statistical learning. *Annu. Rev. Psychol.* 69, 181–203. doi: 10.1146/annurev-psych-122216-011805

Sambeth, A., Ruohio, K., Alku, P., Fellman, V., and Huotilainen, M. (2008). Sleeping newborns extract prosody from continuous speech. *Clin. Neurophysiol.* 119, 332–341. doi: 10.1016/j.clinph.2007.09.144

Sauter, D. A., Eisner, F., Ekman, P., and Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proc. Natl. Acad. Sci.* 107, 2408–2412. doi: 10.1073/pnas.0908239106

Sauter, D. A., Eisner, F., Ekman, P., and Scott, S. K. (2015). Emotional vocalizations are recognized across cultures regardless of the valence of distractors. *Psychol. Sci.* 26, 354–356. doi: 10.1177/0956797614560771

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychol. Bull.* 99, 143–165. doi: 10.1037/0033-2909.99.2.143

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Comm.* 40, 227–256. doi: 10.1016/S0167-6393(02)00084-5

Scott, S. K., and McGettigan, C. (2013). Do temporal processes underlie left hemisphere dominance in speech perception? *Brain Lang.* 127, 36–45. doi: 10.1016/j.bandl.2013.07.006

Seidl, A. (2007). Infants' use and weighting of prosodic cues in clause segmentation. *J. Mem. Lang.* 57, 24–48. doi: 10.1016/j.jml.2006.10.004

Seidl, A., and Johnson, E. K. (2006). Infant word segmentation revisited: edge alignment facilitates target extraction. *Dev. Sci.* 9, 565–573. doi: 10.1111/j.1467-7687.2006.00534.x

Shaw, J. A., and Tyler, M. D. (2020). Effects of vowel coproduction on the timecourse of tone recognition. *J. Acoust. Soc. Am.* 147, 2511–2524. doi: 10.1121/10.0001103

Shi, R., Gao, J., Achim, A., and Li, A. (2017). Perception and representation of lexical tones in native mandarin-learning infants and toddlers. *Front. Psychol.* 8:1117. doi: 10.3389/fpsyg.2017.01117

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition* 106, 833–870. doi: 10.1016/j.cognition.2007.05.002

Singh, L., Fu, C. S., Seet, X. H., Tong, A. P., Wang, J. L., and Best, C. T. (2018). Developmental change in tone perception in mandarin monolingual, English monolingual, and mandarin–English bilingual infants: divergences between monolingual and bilingual learners. *J. Exp. Child Psychol.* 173, 59–77. doi: 10.1016/j.jecp.2018.03.012

Singh, L., Goh, H. H., and Wewalaarachchi, T. D. (2015). Spoken word recognition in early childhood: comparative effects of vowel, consonant and lexical tone variation. *Cognition* 142, 1–11. doi: 10.1016/j.cognition.2015.05.010

Singh, L., Hui, T. J., Chan, C., and Golinkoff, R. M. (2014). Influences of vowel and tone variation on emergent word knowledge: a cross-linguistic investigation. *Dev. Sci.* 17, 94–109. doi: 10.1111/desc.12097

Singh, L., Morgan, J. L., and Best, C. T. (2002). Infants' listening preferences: baby talk or happy talk? *Infancy* 3, 365–394. doi: 10.1207/S15327078IN0303_5

Singh, L., Morgan, J. L., and White, K. S. (2004). Preference and processing: The role of speech affect in early spoken word recognition. *J. Mem. Lang.* 51, 173–189. doi: 10.1016/j.jml.2004.04.004

Singh, L., White, K. S., and Morgan, J. L. (2008). Building a word-form lexicon in the face of variable input: influences of pitch and amplitude on early spoken word recognition. *Lang. Learn. Dev.* 4, 157–178. doi: 10.1080/15475440801922131

Snow, D. (2001). Intonation in the monosyllabic utterances of 1-year-olds. *Infant Behav. Dev.* 24, 393–407. doi: 10.1016/S0163-6383(02)00084-X

Snow, D. (2006). Regression and reorganization of intonation between 6 and 23 months. *Child Dev.* 77, 281–296. doi: 10.1111/j.1467-8624.2006.00870.x

Snow, D., and Balog, H. L. (2002). Do children produce the melody before the words? A review of developmental intonation research. *Lingua* 112, 1025–1058. doi: 10.1016/S0024-3841(02)00060-8

So, C. K., and Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: effects of native phonological and phonetic influences. *Lang. Speech* 53, 273–293. doi: 10.1177/0023830909357156

So, C. K., and Best, C. T. (2011). Categorizing mandarin tones into listeners' native prosodic categories: The role of phonetic properties. *Poznań Stud. Contemp. Ling.* 47:133. doi: 10.2478/psicl-2011-0011

So, C. K., and Best, C. T. (2014). Phonetic influences on English and French listeners' assimilation of mandarin tones to native prosodic categories. *Stud. Second. Lang. Acquis.* 36, 195–221. doi: 10.1017/S0272263114000047

So, L. K., and Dodd, B. J. (1995). The acquisition of phonology by Cantonese-speaking children. *J. Child Lang.* 22, 473–495. doi: 10.1017/S0305000900009922

Soken, N. H., and Pick, A. D. (1999). Infants' perception of dynamic affective expressions: do infants distinguish specific expressions? *Child Dev.* 70, 1275–1282. doi: 10.1111/1467-8624.00093

Stern, D. N., Spieker, S., and MacKain, K. (1982). Intonation contours as signals in maternal speech to prelinguistic infants. *Dev. Psychol.* 18, 727–735. doi: 10.1037/0012-1649.18.5.727

Sullivan, J. W., and Horowitz, F. D. (1983). The effects of intonation on infant attention: The role of the rising intonation contour. *J. Child Lang.* 10, 521–534. doi: 10.1017/S0305000900005341

Thiessen, E. D., Hill, E. A., and Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy* 7, 53–71. doi: 10.1207/s15327078in0701_5

Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., et al. (2015). Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. *J. Acoust. Soc. Am.* 137, 3005–3007. doi: 10.1121/1.4919349

To, C. K., Cheung, P. S., and McLeod, S. (2013). A population study of children's acquisition of Hong Kong Cantonese consonants, vowels, and tones. *J. Speech Lang. Hear. Res.* 56, 103–122. doi: 10.1044/1092-4388(2012/11-0080)

Tonkova-Yampol'skaya, R. V. (1969). Development of speech intonation in infants during the first two years of life. *Sov. Psychol.* 7, 48–54. doi: 10.2753/RPO1061-0405070348

Tsao, F. M. (2017). Perceptual improvement of lexical tones in infants: effects of tone language experience. *Front. Psychol.* 8:558. doi: 10.3389/fpsyg.2017.00558

Tyler, M. D., Best, C. T., Goldstein, L. M., and Antoniou, M. (2014). Investigating the role of articulatory organs and perceptual assimilation in infants' discrimination of native and non-native fricative place contrasts. *Dev. Psychobiol.* 56, 210–227. doi: 10.1002/dev.21195

van Heuven, V. J. (2018). Acoustic correlates and perceptual cues of word and sentence stress: towards a cross-linguistic perspective. In R. Goedemans, J. Heinz and HulstH. van der (Eds.), *The Study of word Stress and accent: Theories, Methods and data*. (15–59). England: Cambridge University Press.

Walker-Andrews, A. S., and Grolnick, W. (1983). Discrimination of vocal expressions by young infants. *Infant Behav. Dev.* 6, 491–498. doi: 10.1016/S0163-6383(83)90331-4

Walker-Andrews, A. S., and Lennon, E. (1991). Infants' discrimination of vocal expressions: contributions of auditory and visual information. *Infant Behav. Dev.* 14, 131–142. doi: 10.1016/0163-6383(91)90001-9

Wang, T., and Lee, Y. C. (2015). Does restriction of pitch variation affect the perception of vocal emotions in mandarin Chinese? *J. Acoust. Soc. Am.* 137, EL117–EL123. doi: 10.1121/1.4904916

Wang, T., Lee, Y. C., and Ma, Q. (2018). Within and across-language comparison of vocal emotions in mandarin and English. *Appl. Sci.* 8:2629. doi: 10.3390/app8122629

Wang, T., and Qian, Y. (2018). Are pitch variation cues indispensable to distinguish vocal emotions. In *Proceedings of the 9th International Conference on Speech Prosody* 324–328.

Wanrooij, K., Boersma, P., and van Zuijen, T. L. (2014). Distributional vowel training is less effective for adults than for infants. A study using the mismatch response. *PLoS One* 9:e109806. doi: 10.1371/journal.pone.0109806

Wellman, H. M. (1992). *The child's Theory of mind.*United States: MIT Press.

Werker, J. F., and Hensch, T. K. (2015). Critical periods in speech perception: new directions. *Annu. Rev. Psychol.* 66, 173–196. doi: 10.1146/annurev-psych-010814-015104

Werker, J. F., and Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Percept. Psychophys.* 37, 35–44. doi: 10.3758/BF03207136

Wermke, K., Ruan, Y., Feng, Y., Dobnig, D., Stephan, S., Wermke, P., et al. (2017). Fundamental frequency variation in crying of mandarin and German neonates. *J. Voice* 31:255.e30. doi: 10.1016/j.jvoice.2016.06.009

Wermke, K., Teiser, J., Yovsi, E., Kohlenberg, P. J., Wermke, P., Robb, M., et al. (2016). Fundamental frequency variation within neonatal crying: does ambient language matter? *Speech Lang. Hear.* 19, 211–217. doi: 10.1080/2050571X.2016.1187903

Wewalaarachchi, T. D., and Singh, L. (2016). Effects of suprasegmental phonological alternations on early word recognition: evidence from tone sandhi. *Front. Psychol.* 7:627. doi: 10.3389/fpsyg.2016.00627

Widen, S. C. (2013). Children's interpretation of facial expressions: The long path from valence-based to specific discrete categories. *Emot. Rev.* 5, 72–77. doi: 10.1177/1754073912451492

Wong, P. (2012a). Acoustic characteristics of three-year-olds' correct and incorrect monosyllabic mandarin lexical tone productions. *J. Phon.* 40, 141–151. doi: 10.1016/j.wocn.2011.10.005

Wong, P. (2012b). Monosyllabic mandarin tone productions by 3-year-olds growing up in Taiwan and in the United States: Interjudge reliability and perceptual results. *J. Speech Lang. Hear. Res.* 55, 1423–1437. doi: 10.1044/1092-4388(2012/11-0273)

Wong, P. (2013). Perceptual evidence for protracted development in monosyllabic mandarin lexical tone production in preschool children in Taiwan. *J. Acoust. Soc. Am.* 133, 434–443. doi: 10.1121/1.4768883

Wong, P., Schwartz, R. G., and Jenkins, J. J. (2005). Perception and production of lexical tones by 3-year-old, mandarin-speaking children. *J. Speech Lang. Hear. Res.* 48, 1065–1079. doi: 10.1044/1092-4388(2005/074)

Yanushevskaya, I., Gobl, C., and Ní Chasaide, A. (2018). Cross-language differences in how voice quality and f0 contours map to affect. *J. Acoust. Soc. Am.* 144, 2730–2750. doi: 10.1121/1.5066448

Yeung, H. H., Chen, K. H., and Werker, J. F. (2013). When does native language input affect phonetic perception? The precocious case of lexical tone. *J. Mem. Lang.* 68, 123–139. doi: 10.1016/j.jml.2012.09.004

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Yoshida, K. A., Pons, F., Maye, J., and Werker, J. F. (2010). Distributional phonetic learning at 10 months of age. *Infancy* 15, 420–433. doi: 10.1111/j.1532-7078.2009.00024.x

Yuan, J. (2011). Perception of intonation in mandarin Chinese. *J. Acoust. Soc. Am.* 130, 4063–4069. doi: 10.1121/1.3651818

Zora, H., Schwarz, I. C., and Heldner, M. (2015). Neural correlates of lexical stress: mismatch negativity reflects fundamental frequency and intensity. *Neuroreport* 26, 791–796. doi: 10.1097/WNR.0000000000000426

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: 27 June 2022
doi: 10.3389/fcomm.2022.849132

Perception of Intonation on Neutral Tone in Mandarin

Yixin Zhang[1]*, *Elaine Schmidt*[1,2] and *Brechtje Post*[1]

[1] *Phonetics Laboratory, Faculty of Modern and Medieval Languages and Linguistics (MMLL), University of Cambridge, Cambridge, United Kingdom,* [2] *Cambridge Assessment, University of Cambridge, Cambridge, United Kingdom*

In Mandarin, lexical tone has been found to interact with intonational tone to influence intonation perception, with the falling T4 facilitating the perception of the statement/question contrast the most, and the rising T2 the least. However, in addition to the four citation tones T1-T4, Mandarin has "neutral tone" which marks weak, non-initial syllables that do not carry a citation tone. The prevailing view is that neutral tone is, in fact, phonologically toneless. It is unknown whether neutral tone can also affect intonation perception. However, it is reasonable to hypothesize that if neutral tone is indeed toneless, it cannot interact with intonational tone in the same way as citation tones do. We investigated this novel hypothesis with a perception experiment in which 22 Mandarin speakers had to determine whether disyllabic citation tone and neutral tone words were a question or statement. Results show that the identification of intonation contours is more accurate for neutral tone than for T2, and similarly accurate for neutral tone and T4, regardless of whether the neutral tone is intrinsic or derived. Furthermore, both T4 and neutral tone are realized with a reduced pitch range at a higher pitch level in questions, unlike T2, which is characterized by a slightly expanded pitch range and a higher pitch level. It is possible that intonation perception in Mandarin is facilitated by changes in the phonetic shapes of lexical tones brought by intonation rather than the phonological interaction between lexical tones and intonation. The importance of pitch changes to the intonation perception in Mandarin was further tested in a second perception experiment with the same 22 participants and disyllabic stimuli with manipulated pitch level and range. Results indicate that the use of pitch cues in intonation perception shows tone-specific differences, namely, pitch range is more important in signaling the question/statement contrast in utterances ending with T4 or neutral tone, while pitch level is the only perceptual cue to interrogativity for utterances ending in T2.

Keywords: neutral tone, Tone and intonation, lexical tone, intonation perception, Mandarin Chinese

OPEN ACCESS

Edited by:
Carlos Gussenhoven,
Radboud University
Nijmegen, Netherlands

Reviewed by:
Liquan Liu,
Western Sydney University, Australia
Bettina Braun,
University of Konstanz, Germany

****Correspondence:***
Yixin Zhang
yz510@cantab.net

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

Received: *05 January 2022*
Accepted: *18 May 2022*
Published: *27 June 2022*

Citation:
Zhang Y, Schmidt E and Post B (2022)
Perception of Intonation on Neutral
Tone in Mandarin.
Front. Commun. 7:849132.
doi: 10.3389/fcomm.2022.849132

INTRODUCTION

In tone languages, f_0 is used as the primary acoustic parameter for two important prosodic features, tone and intonation. At a lexical level, f_0 is employed to distinguish word meanings, and at an utterance level, it conveys intonational information such as discourse function (e.g., signaling questions or statements). Therefore, the realization of intonation in tone languages is more restricted than in non-tone languages. In tone languages, intonation is often realized through a change in pitch register throughout the whole utterance, the insertion of boundary tones, register re-set, and/or through the suspension of downdrift (e.g., Shen, 1989, 1992b; Yuan et al., 2002). In other words, it seems that intonation interacts with rather than overrides lexical tones (Yip, 2002, p. 261). However, there are also tone-less elements in tone languages, like neutral tone in Mandarin. This raises the question addressed in this paper: How is intonation signaled when the syllables involved are phonologically toneless, and there are therefore no lexical tones to interact with it?

Neutral Tone in Mandarin

Mandarin Chinese is a tone language in which lexical tones are part of the phonological specification of morphemes, in addition to vowels and consonants. The majority of Mandarin morphemes are monosyllabic, and each either bears one of the four citation tones (CTs: T1, high-level tone, T2, mid-rising tone, T3, low-convex tone and T4, high-falling tone) or a neutral tone (NT). Syllables with NT are prosodically weak and cannot appear in word-initial positions or on their own, but must be attached to a syllable that carries a CT, whereby more than one NT-bearing syllable can be attached to the same preceding CT-bearing syllable. In this study, we focused on disyllabic words with a single NT. Henceforth, we will refer to words that contain a CT followed by an NT as "NT words". The f_0 realization of NT depends on the preceding CT: NT has a high-falling f_0 contour when following a high-level T1 or mid-rising T2, a high-level contour when following a low-dipping T3, and a mid or low falling contour when following a high-falling T4, and in addition, any following CT may also influence the realization of NT (Lin and Yan, 1980; Lin, 1983; Cao, 1986; Wang, 1996; Lee and Zee, 2014). However, some recent phonetic studies suggest that NT has a static mid target which is implemented with weak articulatory strength (Li, 2003; Chen and Xu, 2006).

From a morpho-phonological perspective, NT is not a homogeneous phenomenon either. Shen (1992a), for instance, proposed a three-way categorization of NT: toneless, detonic and atonic, based on their morphological status combined with their ability to be realized with CTs. Duanmu (2007, p. 248–250) instead identified NT as a stress phenomenon, categorizing Shen's toneless NT as associated with unstressed syllables, while all other tone-bearing syllables are stressed. Zhang (2018, 2021), by contrast, distinguishes two types of NT with different tonal representations but similar phonetic realizations in neutral utterances without narrow focus, based on a series of production and perception experiments: *Intrinsic NT* is carried by functional morphemes that have lost their etymological tone and is phonologically toneless; *Derived NT* is carried by notional morphemes which lose their CT on the surface in particular words when not in focus. In that account, a Derived NT is phonologically represented as the CT it is derived from in all its occurrences, regardless of its surface realization Thus, Derived NT is not phonologically toneless, unlike Intrinsic NT, and the two may therefore affect the production and perception of intonation in different ways. The only study investigating the realization of intonation on NT, however, focused exclusively on Intrinsic NT. It finds that like statements, questions are realized with a gradual f_0 declination when multiple NTs are pronounced in sequence at the end of an utterance, although the declination is not as steep as in statements. In contrast, the high-level T1s are realized with a slightly rising contour in questions (Liu and Xu, 2007). This suggests that in production at least, question intonation does not manifest itself more straightforwardly on NT than on CT. This raises two hitherto unanswered questions: (i) how are different intonation types realized on different types of NT, and (ii) how is intonation perceived on different types of NTs? This study focuses on the second question.

Intonation on Mandarin CTs

In Mandarin, two mechanisms for signaling question intonation have been identified for the final syllables of an utterance, an overall higher f_0 compared to statement intonation and a terminal rise. The implementation of these mechanisms is tone-dependent (Cao, 1986; Shen, 1989, 1992b; Yuan et al., 2002; Liu and Xu, 2005; Peng et al., 2005; Xu, 2005; Yuan, 2006, 2011). To be specific, in addition to raising their overall f_0, the high-level T1 becomes slightly rising, the mid-rising T2 and low-convex T3 have an expanded range, while the high-falling T4 is flattened as its final tonal target is raised (Yuan, 2004; Liu and Xu, 2005; Peng et al., 2005).

The perception of intonation in Mandarin has also been shown to be tone-dependent (Yuan, 2006, 2011; Ren et al., 2013). According to Yuan (2006, 2011), yes/no questions were easiest to identify in utterances ending with a falling T4, and hardest in utterances ending with a rising T2. In other words, the more saliently rising T2 was not necessarily interpreted as a question but led to greater bias toward statements. This finding, according to Yuan (2011), indicates that the phonological identity of tone "intervenes in the mapping of f_0 contours to intonational categories" (p. 19), and that hence tone and intonation interact at a phonological and linguistic level. Furthermore, in an electroencephalographic (EEG) oddball paradigm study using naturally produced monosyllabic stimuli which were controlled for differences in duration, Ren et al. (2013) found that the question-statement contrast elicits a clear mismatch negativity for T4-bearing syllables, but not for T2-bearing ones, indicating that the question-statement contrast is more salient on T4 than on T2. These findings suggest that the phonological identity of the utterance-final tone in a sentence determines the relative "ease" with which they are identified in perception.

However, in a more recent study, Liu et al. (2016) found that questions in utterances ending with T2 and T4 were equally difficult to identify while the identification of statements was difficult in sentences ending with T2 but not in those ending with T4. In other words, while the results of Liu et al. confirmed that there was tone-specific asymmetry in Mandarin intonation perception, they found it in statement perception. This is different from Yuan (2006, 2011) in which the asymmetry was found in question perception, because more questions on utterances ending with the rising T2 were misinterpreted as statements compared to utterances ending with the falling T4. An explanation suggested by Liu et al. (2016) to account for these potentially contradictory findings is that the realization of the question-final T4 used in Liu et al. (2016) differed from the f_0

contour used for T4 in Yuan's and Ren et al.'s studies in that the f_0 curve of the T4 in Liu et al. (2016) was not flattened as much as in the other two studies. This raises an alternative possibility that the ease of intonational perception in a tone language like Mandarin Chinese depends on the size of the difference between statement and question intonations of any given tone.

To summarize, if the phonological representations of lexical tones interfere with intonation perception as Yuan (2006, 2011) suggested, the perception of questions carried by Intrinsic NT syllables should differ from questions carried by CTs. Furthermore, the acoustic realization and interpretation of question intonation on phonologically toneless Intrinsic NT may also differ from phonologically specified Derived NT. We examine these possibilities in Experiment 1 by testing whether intonation is easier to perceive on intrinsic neutral tone because it is phonologically toneless than on CT and derived NT, because these are phonologically specified, at least in their underlying their forms. Following on from the findings of Experiment 1, Experiment 2 then investigates the relative contribution of different pitch cues to the perception of intonation type (in this case, question intonation).

EXPERIMENT 1

In Experiment 1, we investigated intonation perception (question vs. statement) in short utterances ending with Intrinsic NT, Derived NT, and T2 and T4 as the baseline citation form conditions.

H1. Since Intrinsic NT is phonologically toneless, the identification of intonation type for Intrinsic NT stimuli should be more accurate and faster compared to stimuli that are phonologically specified for tone (i.e., Derived NTs and CTs).

Methodology

Participants

Twenty-two Northern Mandarin speakers (6 males, 16 females) aged between 18 and 29 (mean age 23.7) participated in the experiment. All participants were current students at the Shanghai Jiao Tong University, Minhang Campus. None of the participants had lived in Shanghai for more than 3 years, as they all completed their pre-university education in the Huabei region and reported Northern Mandarin as the main language they used in school and at home. Therefore, the influence of the local Wu dialect in Shanghai on these participants is very limited. All of them were right-handed and none of them reported any hearing impairments. Informed consent was obtained prior to the experiment.

Stimuli

To test H1, we chose disyllabic stimuli in which the second syllables carried the target tone (i.e., Intrinsic NT, Derived NT, T2, or T4) and the first syllables carried T1. T1-T2 and T1-T4 words were chosen as the representative CT words, since previous studies found that intonation type was the hardest to identify in the case of utterances ending with T2 and the easiest in utterances ending with T4. The high-level T1 was chosen as the first syllable tone because it allowed for the most natural range of f_0 manipulations, and to keep the overall duration of the experiment to a reasonable time, no other CTs were used as the first syllable tone. For each of the four tone conditions, 32 stimulus words were used (in the Derived NT condition, 8 words were phonologically specified as T1+T1, 8 words as T1+T2, 8 words as T1+T3 and 8 words T1+T4), resulting in 128 items in total. Due to the limitation of the natural language, the second syllable of the items in the different tone conditions (Intrinsic NT, Derived NTs, T2 or T4) did not have the same segments, but the segmental complexity of the items (calculated by dividing the number of segments in the second syllable by the number of segments in the first syllable) was matched across the tone conditions. Furthermore, the NT items chosen here do not have minimal pairs carrying T2 or T4, and the T2 and T4 items did not exist in minimal pairs carrying NT (**Table 1**; **Appendix A**). Thirty-six disyllabic Mandarin words with T1 as the tone on the first syllable were added as fillers. Half of those had T1 as the second syllable tone and the other half had T3 as the second syllable tone.

The experimental items and the fillers were recorded by the first author, a native northern Mandarin speaker aged 27, and another female speaker of very similar age, education and language background. Both speakers hold the Level 1 (the top level) certificate of the National Mandarin Test for native speakers. Two speakers rather than one were recorded to ensure that participants could not just focus on the acoustic differences occurring within a single speaker's productions.

The recordings were made separately by each speaker in a quiet room with a Zoom H1 handy recorder at 96.000 Hz/26Bit. Stimuli that were not clear enough to allow for f_0 manipulation straightforwardly (e.g., due to co-articulation) were re-recorded. The naturalness of the stimuli was examined by the two speakers as well as a naive male northern Mandarin speaker by asking them to pick out the unnatural stimuli.

The recordings were then cross-spliced to neutralize the acoustic parameters of the preceding T1-bearing syllable between the two intonation conditions (declarative and interrogative) using Praat (Boersma and Weenink, 2021). For each recording, the initial syllable and second syllable were separated into two sound files. The pitch height, range and duration of the initial syllables were manipulated into an average pitch height, pitch range and duration of the statement and the question versions of the same stimulus word produced across all words by the same speaker. The second syllables were then spliced onto the manipulated initial syllable with the other intonation type, that is, the second syllables in statement intonation were attached to the initial syllables in question intonation of the same word. Equally, second syllables in question intonation were attached to the initial syllables in statement intonation of the same word and speaker. The intonation of the stimuli as discussed henceforth was determined by the intonation of the second syllable of the stimuli. The fillers were all manipulated in the same way as the stimulus words.

After cross-splicing, the average intensity of the stimuli was scaled to 75 dB. The digitally edited recordings were judged as natural by two native speakers who did not participate in the study. The stimuli were separated into two equal sets of

TABLE 1 | Examples of stimuli in each condition.

Tone	Word	Pinyin transcription	IPA transcription	Glossary
Intrinsic NT	鸽子	ge1zi0	/kɤ1 tsi0/	Pigeon
Derived NT from T1	孙家	sun1jia0(1)	/sʊn1 tɕia0 (1)/	The Sun's family
Derived NT from T2	敦实	dun1shi0(2)	/tun1 ʂi0 (2)/	Stoky
Derived NT from T3	家里	jia1li0(3)	/tɕia1 li0 (3)/	(At) home
Derived NT from T4	吃过	chi1guo0(4)	/tʊhi1 kuɔ (3)/	Have eaten
T2	清除	qing1chu2	/tɕhiŋ1 tʂhu2/	Delete
T4	捉住	zhuo1zhu1	/tʂuo1 tʂu4/	Get hold of

The numbers in Pinyin Transcription and IPA Transcription indicate tones (0 = NT) and the numbers in bracket indicate the phonological CTs of Derived NTs.

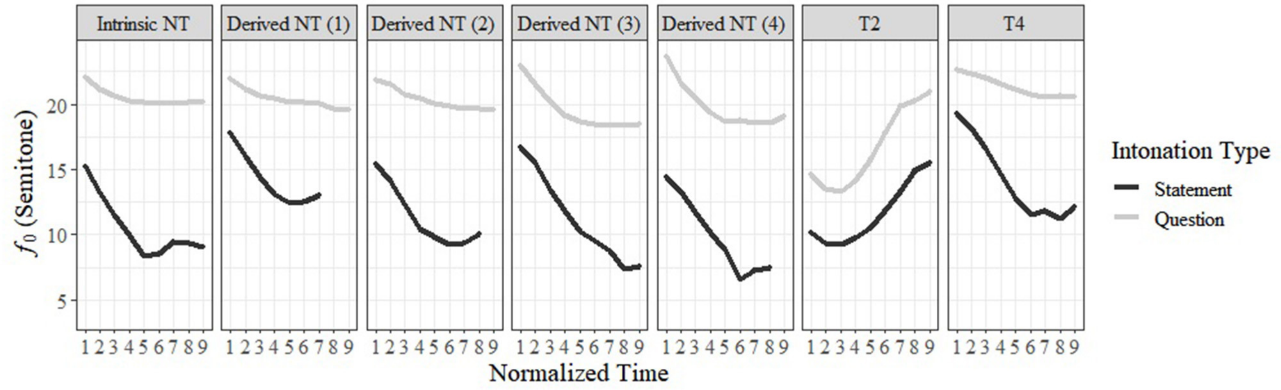


FIGURE 1 | Contours of the 2nd-syllable tone by tone and intonation (Numbers in brackets indicate the phonological tones of Derived NTs).

TABLE 2 | Average f_0 height and range of the second tones.

Tone	Intonation	f_0 Height (semitones)		Intonation	f_0 Range (semitones)	
		Average	SE		Average	SE
Intrinsic NT	Statement	14.96	0.06	Statement	8.65	0.06
	Question	20.5	0.02	Question	2.13	0.02
Derived NT (1)	Statement	13.08	0.38	Statement	11.43	0.72
	Question	20.41	0.23	Question	2.4	0.23
Derived NT (2)	Statement	12.33	0.39	Statement	11.46	0.62
	Question	19.62	0.32	Question	4.74	0.32
Derived NT (3)	Statement	11.81	0.33	Statement	12.72	0.62
	Question	19.86	0.31	Question	6.05	0.42
Derived NT (4)	Statement	11.3	0.32	Statement	7.79	0.45
	Question	20.47	0.17	Question	2.25	0.12
T2	Statement	11.61	0.01	Statement	6.75	0.02
	Question	16.4	0.02	Question	10.7	0.04
T4	Statement	14.43	0.02	Statement	10.45	0.04
	Question	21.35	0.02	Question	2.38	0.02

word-pairs with different intonations. Each set had half of the recordings of the stimulus words from one speaker who did the recording and the other half from the other. The order of the stimuli was pseudorandomized. Half the participants were tested with one set and half with the other.

To be better able to interpret the perception data, we conducted acoustic analyses of the stimuli after manipulation with Praat (Boersma and Weenink, 2021). Firstly, we analyzed the f_0s of the second syllables. A Praat script was applied to extract f_0 values (converted to semitones with 1 Hz as the reference value)

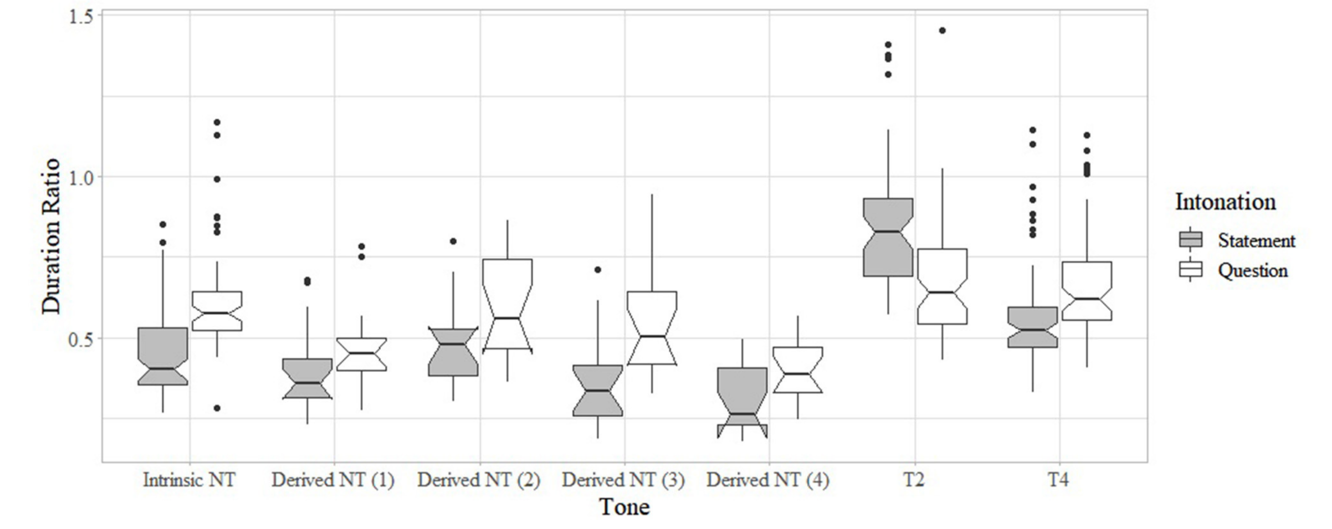


FIGURE 2 | Duration ratio by tone and intonation (Numbers in brackets indicate the phonological tones of Derived NTs).

TABLE 3 | Duration ratio and duration of the 2nd syllable by tone and intonation.

Tone	Intonation	Ratio		Intonation	Duration of the 2nd syllable (ms)	
		Average	SE		Average	SE
Intrinsic NT	Statement	0.46	0.00	Statement	181.52	1.95
	Question	0.61	0.01	Question	238.82	2.57
Derived NT (1)	Statement	0.40	0.03	Statement	242.26	20.19
	Question	0.47	0.04	Question	292.97	24.41
Derived NT (2)	Statement	0.48	0.04	Statement	278.78	23.23
	Question	0.60	0.05	Question	341.98	28.50
Derived NT (3)	Statement	0.36	0.03	Statement	203.86	16.99
	Question	0.54	0.05	Question	309.00	25.75
Derived NT (4)	Statement	0.31	0.03	Statement	200.13	16.68
	Question	0.40	0.03	Question	259.11	21.59
T2	Statement	0.84	0.01	Statement	355.09	3.95
	Question	0.68	0.01	Question	273.97	2.99
T4	Statement	0.57	0.01	Statement	235.17	2.53
	Question	0.67	0.01	Question	268.15	2.88

of the sonorous part in the second syllable of each stimulus. The f_0 contours were time-normalized by dividing the sonorous parts into 10 equal intervals, and f_0 values were extracted at each 10% step. Time-normalized rather than raw f_0-values were used to better illustrate any differences in pitch movement, range and height, as NTs and CTs differ in duration. The last value was excluded to reduce effects of final creakiness on f_0, and tokens with creakiness (i.e., no f_0 value extracted at a measure point) in more than 50% of the measured points were excluded from the f_0 analysis (thirty-eight statement Intrinsic NT and two statement T4 tokens; note that they were included in the perceptual experiment). We also analyzed the f_0 height and range (i.e., the difference between the minimum and maximum f_0 values) of the second syllables, and used linear-mixed effect (LME) models to evaluate the effects of Tone, Intonation, Speaker and their interactions on these two f_0 parameters. The model-building process is presented in detail in Section Data Analysis.

The f_0 contours realized on stimuli with <50% creakiness are illustrated in **Figure 1**. A clear difference between question and statement could be observed for all tones. **Figure 1** shows that question intonation raised the f_0 level in all tone stimuli, and that the range of the falling contour of Intrinsic NT was reduced. All other tones tested here show the same pattern except T2, which is only realized with raised pitch but slightly expanded range.

Analyses of average f_0 height and range confirmed these observations (**Table 2**). LME models showed that Tone, Intonation, Speaker, and the two-way interactions between them all had significant effects on the average f_0 height and range

TABLE 4 | Identification accuracy, hit rate (H, i.e., the identification accuracy of statement intonation), false alarm (FA), discriminability (A') and Bias (B''_D) in each tone condition.

Tone	Identification accuracy	Intonation	Identification accuracy	Hit rate (H)	False alarm (FA)	A''	B''_D
Intrinsic NT	95.48%	Statement	96.72%	96.72%	5.81%	0.98	0.29
		Question	94.19%				
Derived NT (1)	93.75%	Statement	95.45%	95.45%	7.95%	0.97	0.29
		Question	92.05%				
Derived NT (2)	92.90%	Statement	94.03%	94.03%	8.24%	0.96	0.17
		Question	91.76%				
Derived NT (3)	93.89%	Statement	94.89%	94.89%	7.10%	0.97	0.17
		Question	92.90%				
Derived NT (4)	94.03%	Statement	95.17%	95.17%	7.10%	0.97	0.20
		Question	92.90%				
T2	85.85%	Statement	86.77%	86.77%	15.00%	0.91	0.07
		Question	85.00%				
T4	93.17%	Statement	96.88%	96.88%	10.65%	0.96	0.57
		Question	89.35%				

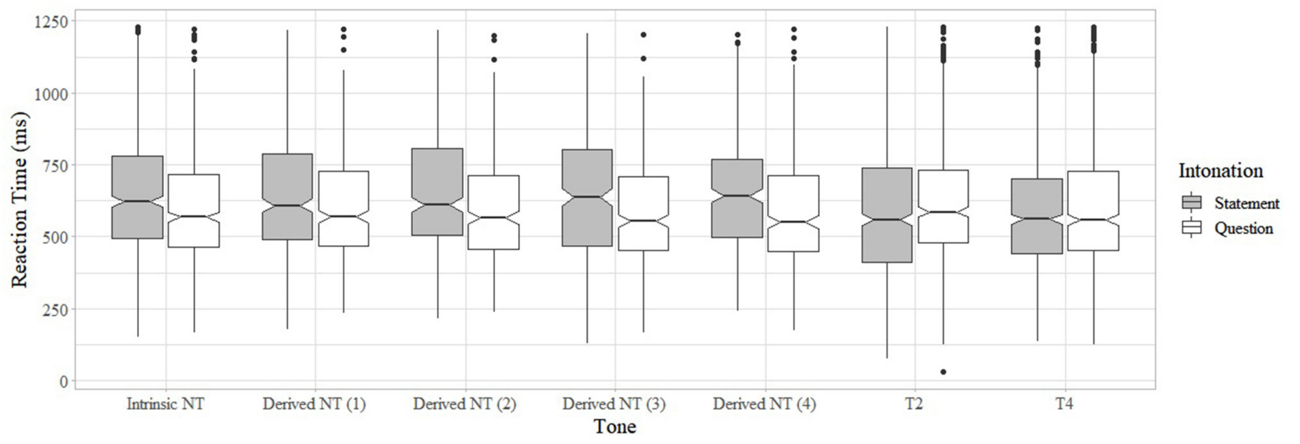


FIGURE 3 | Reaction time by tone and intonation (Numbers in brackets indicate the phonological tones of Derived NTs).

of the second tones, and the three-way interaction between Tone, Intonation and Speaker only affected average f_0 height ($ps < 0.0001$; for the full model, see **Supplementary Table 1** in **Appendix B**). Tukey *post-hoc* comparisons showed that the average f_0 height of questions was significantly higher than that of statements in all tone conditions ($ps < 0.001$). As to f_0 range, questions showed a significantly smaller f_0 range than statements in Intrinsic NT, Derived NT (3), Derived NT (4) and T4 ($ps < 0.001$). However, in T2, the pattern was reversed, namely, the pitch range for statements was significantly smaller than the pitch range for questions ($p < 0.001$). The interaction between Tone, Intonation and Speaker was significant due to the speakers consistently differing in their production of different tones in different intonation types. Since this is not relevant for our study, this will not be further presented.

Furthermore, the duration ratio for all stimuli in each tone and intonation condition was calculated (=duration of 2nd syllable/duration of the 1st syllable) and evaluated using an LME model. In general, we focused on the acoustic differences between the two intonation types within the same tone condition, rather than differences between tones with the same intonation, as different tones are already expected to have different f_0 and durational realizations.

In terms of duration ratio, the LME model showed that Tone, Intonation, Speaker, the interaction between Tone and Intonation, as well as the interaction between Tone and Speaker had significant effects on the duration ratio ($ps < 0.0001$; for the full model, see **Supplementary Table 2** in **Appendix B**). As can be seen in **Figure 2** and **Table 3**, stimuli with T2 also showed a reversed pattern to the other tones, namely, the duration ratio was smaller in T2 when it was a question, while it was larger when it was a statement all other tone conditions, and the same patterns were observed for the absolute duration of the 2nd syllable (all $ps < 0.001$).

TABLE 5 | Reaction time by tone and intonation.

Tone	Reaction time (ms)		Intonation	Reaction time (ms)		***Post-hoc*** **comparisons between statement and question**
	Average	**SE**		**Average**	**SE**	
Intrinsic NT	626.76	6.06	Statement	652.72	12.07	$p < 0.005$
			Question	598.99	10.7	
Derived NT (1)	624.78	8.28	Statement	649.52	9.02	0.18
			Question	599.57	8.29	
Derived NT (2)	625.58	8.33	Statement	657.65	12.05	$p < 0.005$
			Question	592.25	10.9	
Derived NT (3)	617.95	8.32	Statement	648.76	12.65	$p < 0.05$
			Question	586.34	10.53	
Derived NT (4)	617.79	8.33	Statement	644.79	11.88	$p < 0.05$
			Question	589.85	11.84	
T2	602.28	6.63	Statement	583.44	10.1	0.14
			Question	620.05	8.53	
T4	593.83	6.23	Statement	588.46	8.8	0.99
			Question	599.81	8.86	

Procedure

The experiment was programmed in PsychoPy 3.0 (Peirce et al., 2019). Participants heard a manipulated recording of the stimuli (**Table 1**; **Appendix A**) while watching a screen on the experimental laptop which showed two horizontally arranged icons, "?" and "!" to record whether they heard a question or a statement.[1] Participants were asked to indicate their choice by pressing the keys on the keyboard labeled "?" or "!" after which the next trial automatically started. If no button was pressed within 3,000 ms, the next trial automatically started. The "?" was assigned to the left keyboard response button for half the participants and the right for the other half to avoid interference from handedness. The 42 trials with null results were treated as incorrect answers in the data analyses.

The experiment consisted of 256 test trials, 36 fillers and 32 trials of repeated words (324 trials in total) with two participant-controlled breaks available in between. The 32 trials of repeated words had randomly chosen words from the other stimulus set (eight words per tone condition) that appeared only once either in statement or in question intonation to prevent predictability, that is, the within-subject manipulation (i.e., each utterance is presented in both intonation conditions) may lead participants to choose the other response options for strategic reasons. Nine practice trials were given at the beginning to help participants familiarize themselves with the procedure. The whole experiment took about 40 min including instructions and practice trials.

Data Analysis

Identification accuracy was calculated to measure how well an intonational function (i.e., statement vs. question) was recognized by the listeners. The response was considered accurate only if the intonation was identified as the same intonation that the speakers were asked to produce. The effects of Tone, Intonation and individual differences between Speakers on identification accuracy, a binary categorical variable (Accurate vs. Inaccurate), were evaluated by logistic mixed effects models, using glmer in the lmerTest package (Kuznetsova et al., 2017) in R (R Core Team, 2020). We assigned value 0 to Inaccurate and 1 to Accurate, and selected the optimal fixed structure by using stepwise comparisons from the most complex structure to the simplest and the optimal random effect structure according to the smallest Akaike Information Criterion (AIC). The anova() function served to compare different models to determine whether excluding factors from the analysis led to a better fit (Field et al., 2012). The details of the final models are presented in **Appendix B**.

According to Signal Detection Theory (see Macmillan and Creelman, 2004, for an introduction), identification involves not only the ability to discriminate between the two intonation conditions, but also the bias toward one of them in ambiguous situations. More specifically, Signal Detection Theory applies to the situation in which participants are asked to determine which of two categories (i.e., statement and question in our case) a stimulus belongs to. The task generates two measures of behavioral performance: the hit rate and the false alarm rate. In the present study, the response option of the statement was arbitrarily assigned to the signal, the question to the noise. Then, a hit (H) referred to when "the signal (statement) was presented and chosen" (i.e., the correct identification), a miss to when "the signal (statement) was presented but not chosen", a false alarm (F) to when "the noise (question) was presented but not chosen" and a correct rejection to when "the noise (question)

[1] The symbol for a full stop "." was not used for statements to avoid the visual imbalance between "." and "?", and it was made clear in the instructions that "!" stood for "statement" rather than "exclamation". In Mandarin, "!" can be used to express strong emotions ranging from surprise, happiness to sadness and regrets in declarative sentences (The National Bureau of Quality Technical Supervision, 1996). This may have slowed down statement compared to question responses, but this is not relevant here, since such a bias would have applied across all tone conditions.

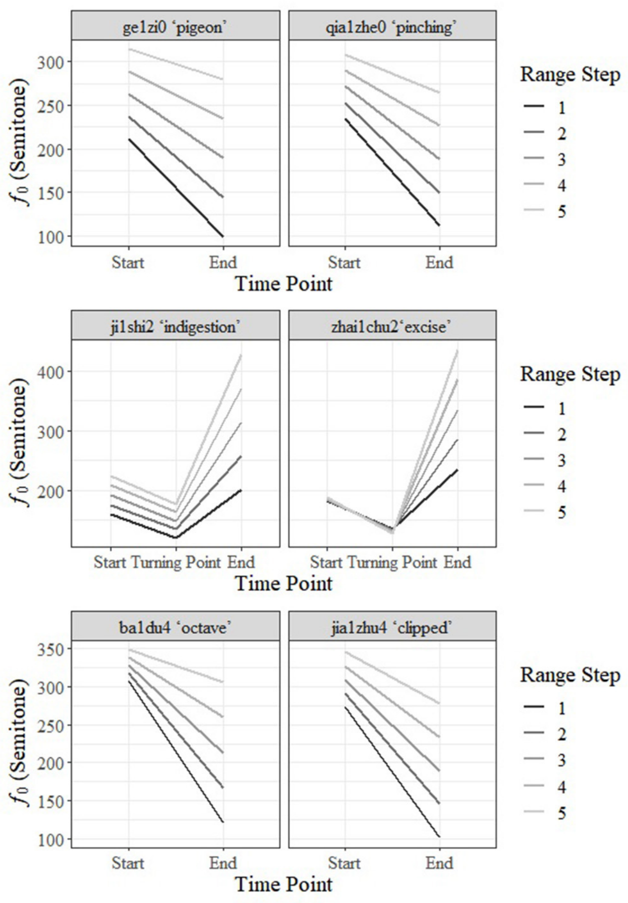


FIGURE 4 | Range steps of each stimulus at Height Step 1.

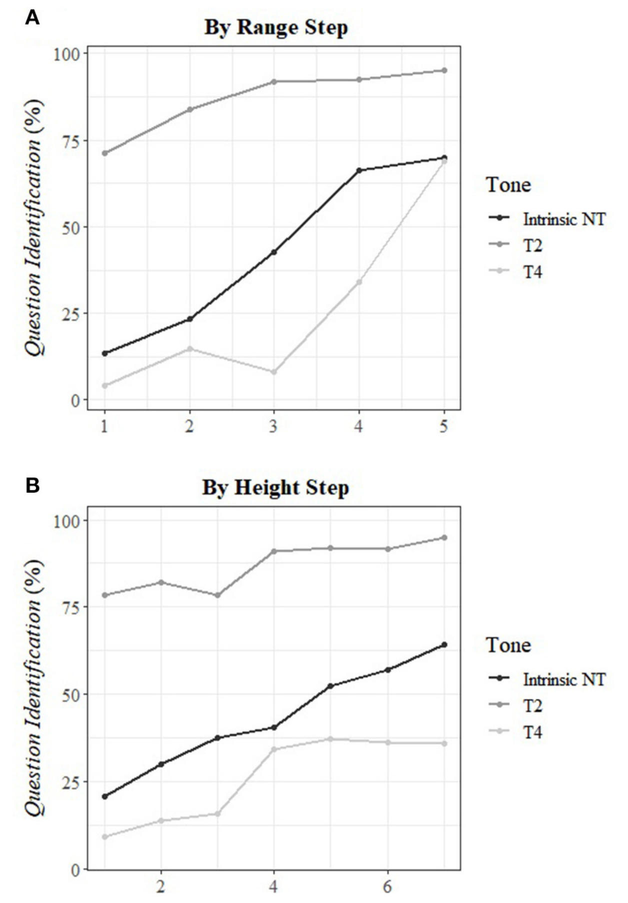


FIGURE 5 | Percentage of stimuli identified as questions (*question identified*) by Range Step **(A)** and Height Step **(B)**.

was presented and chosen". In studies using Signal Detection Theory, H and F are transformed into indices of identification sensitivity like A$'$ based on statistical models, which indicates the discriminablility between the signal and the noise (Pollack and Norman, 1964; Smith, 1995; Zhang and Mueller, 2005). Calculated as (1), A$'$ ranges between 0 and 1, 1 indicating maximum performance and 0.5 indicating chance performance (Zhang and Mueller, 2005, p. 207). The larger A$'$ is, the better the perceptual result is.

$$\mathrm{A}' = \begin{cases} 0.75 + \frac{\mathrm{H-F}}{4} - \mathrm{F(1-H)} & \text{when } \mathrm{F} \leq 0.5 \leq \mathrm{H} \\ 0.75 + \frac{\mathrm{H-F}}{4} - \frac{\mathrm{F}}{4\mathrm{H}} & \text{when } \mathrm{F} \leq \mathrm{H} < 0.5 \\ 0.75 + \frac{\mathrm{H-F}}{4} - \frac{\text{1-H}}{\text{4(1-F)}} & \text{when } 0.5 < \mathrm{F} \leq \mathrm{H} \end{cases} \tag{1}$$

The participants' response bias was indexed by B$''_{\mathrm{D}}$, which correlates to the slope of the receiver operating characteristic function at the point of observation. B$''_{\mathrm{D}}$ was calculated following Pallier (2002) as (2) and ranges from −1 (maximum bias to the question) to 1 (maximum bias to statement). The absolute value of B$''_{\mathrm{D}}$ reflects the perceptual bias. The smaller it is, the better the perceptual result. We calculated A$'$ and B$''_{\mathrm{D}}$ by tone condition.

$$B''_{\mathrm{D}} = \frac{(1-\mathrm{H}) \times (1-\mathrm{F}) - \mathrm{H} \times \mathrm{F}}{(1-\mathrm{H}) \times (1-\mathrm{F}) + \mathrm{H} \times \mathrm{F}} \tag{2}$$

Reaction time (=the time of key-pressing minus the offset time of the auditory stimulus) was collected alongside as a measure of the difficulty of identifying intonation. Null results were excluded from the analyses here. Outliers in the reaction time data were removed following the Interquartile Rule (Tukey, 1977), and the effects of Tone, Intonation and Speakers on reaction time (a continuous numeric variable), were evaluated by linear mixed effect (LME) models. LME models were built through a similar process to the logistic mixed effect model but used lmer in the lmerTest package (Kuznetsova et al., 2017) in R (R Core Team, 2020). The details are presented with the results. To establish the LME models, the skewed data were transformed using square root transformation (Hothorn and Everitt, 2006).

TABLE 6 | Percentage of stimuli identified as question, by contour steps and height steps.

Tone	Contour step	Average (%)	SE (%)	Height step	Average (%)	SE (%)
Intrinsic NT	1 (Statement)	13.31	2.74	1 (The lowest)	20.68	7.29
T2		71.1	3.01		78.41	5.51
T4		4.06	2.16		9.09	4.81
Intrinsic NT	2	23.54	4.77	2	30	8.48
T2		83.93	3.83		82.05	6.75
T4		14.77	4.37		13.86	8.14
Intrinsic NT	3	42.7	7.48	3	37.5	12.28
T2		91.88	3.56		78.41	3.8
T4		8.28	2.93		15.91	9.04
Intrinsic NT	4	66.4	10.48	4	40.46	13.58
T2		92.53	3.28		90.91	3.39
T4		34.09	8.19		34.09	15.88
Intrinsic NT	5 (Question)	69.97	7.25	5	52.5	14.33
T2		94.97	2.22		92.05	5.01
T4		68.83	10.94		37.05	14.88
Intrinsic NT	-	6	57.04	12.47		
T2			91.59	4.97		
T4			36.14	16		
Intrinsic NT	-	7 (The highest)	64.09	15.13		
T2			94.77	4.14		
T4			35.91	17.77		

Results

Identification and Bias

The logistic regression model found that identification accuracy was significantly influenced by Tone ($p < 0.005$), Intonation ($p < 0.0001$), Speaker ($p < 0.0001$), and interactions between Tone and Intonation ($p < 0.01$) and between Speaker and Intonation ($p < 0.0005$) (for the full model, see **Supplementary Table 3** in **Appendix B**).

Tukey *post-hoc* comparisons showed that the identification accuracy of intonation for T2 was significantly lower than for all the other tones, namely, the four Derived NTs (ps < 0.05), Intrinsic NT ($p < 0.001$) and T4 ($p < 0.001$; **Table 4**). None of the accuracy differences between the other tones were significant, and neither were the differences among the four Derived NTs. When examined by intonation type, with regard to the identification of statement, the accuracy for T2 was significantly lower than that for Intrinsic NT ($p < 0.001$) and T4 ($p < 0.001$), but the other accuracy differences between tones were not significant. With regard to the identification of questions, the accuracy on T2 was significantly lower than that on Intrinsic NT ($p < 0.005$), while the differences between the other tones were not significant. To summarize, significant differences were mainly found between T2 and the other tones, especially between T2 and Intrinsic NT and T4, but not between the two types of NTs or between NTs and T4. Since there was no consistent difference in the identification of the same intonational contours produced by different speakers, the interaction between Intonation and Speaker is not relevant here and will thus not be further analyzed.

Discriminability and bias results showed that intonation on NTs and T4 was highly differentiable, more differentiable than intonation on T2 (**Table 4**). B''_D values were positive in all the conditions, suggesting that there was a bias toward statements, in line with previous findings (Yuan, 2006). However, the identification bias was larger in the T4 condition than in the other tone conditions, but smallest in the T2 condition.

Reaction Time

Significant effects of Intonation ($p < 0.001$), Tone ($p < 0.0001$) and the interaction between Tone and Intonation ($p < 0.0001$) on reaction time were found for reaction times (for the full model, see **Supplementary Table 4** in **Appendix B**). On average, the reaction time for question identification was significantly shorter than the reaction time for statement identification ($p < 0.001$). With regards to reaction time differences between tones, Tukey *post-hoc* comparisons showed that reaction time in the Intrinsic NT condition was significantly longer than that in T4 ($p < 0.01$), but the other differences between tone conditions were not significant.

When examined more closely, the reaction time differences between intonation types were only significant for Intrinsic NT ($p < 0.005$), Derived NT phonologically specified as T2 ($p < 0.005$), T3 ($p < 0.05$), and T4 ($p < 0.05$), but not for Derived NT phonologically specified as T1 or the two CTs (**Figure 3**; **Table 5**). When examined by intonation type, no significant differences were observed between tones with regard to question identification, but it took significantly longer to identify statements on Intrinsic NT and Derived NTs compared to T2 and T4 (ps < 0.05).

Discussion of Experiment 1

The present experiment examined the identification of intonation on Intrinsic NT and Derived NT in comparison to that of two CTs, the rising T2 and falling T4. The findings confirmed that intonation perception is easiest on T4 and hardest on T2, as has been found in previous studies (e.g., Yuan, 2011). However, the results did not confirm the hypothesis that the intonation type realized on Intrinsic NT is identified faster and more accurately than intonation on the other tones tested here, on the basis that it does not have phonologically specified tones that interact with intonation (H1). Instead, we found that the identification accuracy for Intrinsic NT was only higher than T2, but it was not significantly different from that for Derived NTs and T4. Moreover, there was similarly high discriminability (A') of intonation types in the Intrinsic NT, Derived NT and T4 conditions, higher than that in T2. In other words, Intrinsic NT patterned with the other falling tones (i.e., Derived NTs and T4) in accuracy, suggesting that the phonetic shape of the tonal contour provides the crucial explanatory information in tone-intonation interaction in Mandarin. There may be a ceiling effect in play as the identification accuracy in Intrinsic NT, Derived NT and T4 conditions was above 90% (Huang and Johnson, 2010).

In terms of identification bias, although all tones showed a bias toward statement interpretation in line with previous studies (e.g., Yuan, 2006, 2011; Liu et al., 2016), the largest bias was found for T4, and the smallest for T2. B''_D values in the NT conditions were all smaller than 0.3, and not comparable to the B''_D value of 0.57 in T4 ($B''_D = 0$, no bias; $B''_D = 1$, maximum bias to statement). It seems that although the identification accuracy for T4 was as high as for the NTs, it contained more bias toward statements. Also, although the intonation identification accuracy on T2 was more problematic, the smallest bias toward statements was found in the T2 condition, which indicates that the identification of question and statement were equally problematic, in line with Liu et al. (2016).

The bias results so far seemed to suggest that Intrinsic NT facilitates intonation perception in a more balanced way in comparison to T4. It is possible that in the absence of a phonological interaction between lexical tone and intonation in phonologically toneless Intrinsic NT syllables, intonation can somehow be better accommodated than in syllables with phonologically specified lexical tones. However, this interpretation fails to explain the low B''_D values found for Derived NTs. Derived NTs seem to have phonological tones which are assumed to interact with intonation just like T2 and T4. Moreover, we found that Derived NT phonologically specified as T4 did not enable a higher identification accuracy nor a less biased identification than Derived NTs with the other phonological tones, which also indicates that intonation perception for Derived NTs is not affected by the phonological identity of the tone. Note that high B''_D value found in T4 may in fact be affected by the fact that it was on a small number of misses and false alarms (Stanislaw and Todorov, 1999; Zhang and Mueller, 2005).

The reaction time results did not support H1 either. Normally, we would expect higher accuracy and shorter reaction times to indicate ease of identification, as was the case for T4 vs. T2, but against the hypothesis, intonation identification took significantly longer for NTs than T4. It is possible that the participants found it harder to identify the NT stimuli due to their weak surface realization and short duration, diminishing the salience of the relevant perceptual cues. It is also possible to attribute this finding to the very short duration of the NT-bearing syllables and the statements in general, because the key-pressing process will always need a certain amount of time.

Taken together, these findings show that any interaction that may take place at a phonological level between intonation and lexical tones cannot account for the intonation identification data analyzed here. Instead, the facilitative effect observed for T4 as opposed to T2, as well as the absence of a significant difference between T4 and both types of NT suggest that it is, in fact, the surface f_0 pattern that is of crucial importance here. More specifically, unlike T2, T4 and both types of NT all have a falling contour which is raised and flattened under question intonation, which makes their surface realizations quite unlike their realization in statement contexts. T2 also shows a slight shift in range and height, but otherwise, the contour is identical in the two intonation conditions.

EXPERIMENT 2

In Experiment 1, despite tone-specific differences, the raising of f_0 to signal question intonation was clearly found across all tone conditions as well as a changed f_0 range due to a further raising of the utterance-final targets (see **Figure 1** above). Liang and Heuven (2009) used a sentence made up of seven syllables carrying high-level T1 to investigate the relative weighting of these two cues in intonation perception. By manipulating the overall f_0 height of the utterance and the terminal f_0 height of the final syllable, they established that the f_0 rise in the utterance-final tone was a more important cue to question intonation than the overall height of the utterance. What Liang and Heuven (2009) could not fully investigate by using T1 syllables only is the potential effects of individual lexical tones on the perpetual cues to intonation type, especially the changed f_0 range of the utterance-final tones. Since question intonation changes the surface contour of question-final tones in a tone-specific manner, how cues to questions are weighted in perception may also vary between different utterance-final tones. More specifically, we hypothesized that:

H2: The change in pitch range is more important to the perception of intonation type on tones with falling contours (i.e., Neutral tone and T4) while changes in both pitch range and height are important cues to intonation perception in the rising T2.

Methodology

Participants

The same 22 participants that participated in Experiment 1 also took part in Experiment 2. Experiment 2 took place about 2 months after Experiment 1.

Stimuli

Two Intrinsic NT words, two T2 words and two T4 words from Experiment 1 were recorded by the first author as representative NT and CTs, and cross-spliced as in Experiment 1. We then manipulated the duration of the first and the second syllables of the recordings of the same stimulus word into the average duration of the two intonation conditions and scaled the intensity of the recordings at 75 dB. Then, we systematically manipulated the pitch height of the disyllabic stimuli and the pitch range of the second syllables using Praat (Boersma and Weenink, 2021). For each stimulus word, we created 3 height steps and 3 range steps with equal intervals between the question and the statement version, and also added 2 more extra height steps (i.e., one higher than the question and one lower than the statement). Height step 1 is the lowest and 7 is the highest while range step 1 is the statement range and 5 is the question range. For pitch height, to simplify the manipulation, we calculated the average f_0 of all six stimuli and rounded the number to create intervals. The average f_0 of the statement stimuli across tone conditions was 289.96 Hz (SE = 7.65 Hz), about 70.24 Hz lower than that of the question stimuli, 360.20 Hz (SE = 6.59 Hz). Therefore, we set Height step 1 of all stimuli at 245 Hz, and Height step 7 at 380 Hz, with an equal interval of 22.5 Hz in between. The manipulation of range at Step 1 is illustrated in **Figure 4**.

We manipulated the stimuli starting from both the statement and question recordings, resulting in 70 manipulations (5 range steps * 7 height steps * 2 source recordings) for each stimulus and 420 stimuli in total (70 steps manipulations * 2 stimulus words * 3 tones). Forty-eight stimuli from Experiment 1 were added as fillers without manipulation.

Procedure

The experiment was programmed in PsychoPy 3.0 (Peirce et al., 2019) and the procedure was the same as in Experiment 1 except that this time, no time limit was set for key-pressing, though participants were encouraged to give their answers as quickly as possible. This was because during piloting, participants reported that they were distracted by trying to observe the time limit. The participants took part in this experiment in a quiet room. The experiment consisted of nine practice trials which were the same as in Experiment 1, 420 experimental trials and 48 filler trials which were pseudo-randomized with two 5-min breaks. The whole experiment took about 40 min including instructions and practice trials.

Data Analysis

The analysis focused on intonation identification (i.e., question or statement). A binominal ordinary logistic regression model was first established to evaluate whether and how Tone, Pitch height, Pitch range and Original intonation (i.e., manipulated from the original recording of the statement or the question version) affected identification (Question vs. Statement) in each tone condition. Since the complexity of the model influences the degree of uncertainty (Babyak, 2004), we further split the data by Tone, and for each tone condition, a binominal ordinary logistic regression was established to evaluate the effects of Pitch height, Pitch range and Original intonation, and their interactions. The models were established through a process similar to the models established in Experiment 1 using glm in the lmerTest package (Kuznetsova et al., 2017) in R (R Core Team, 2020), and *post-hoc* comparisons were carried out using Tukey tests. The percentage of stimuli identified as questions (henceforth *question identification*) was also calculated by dividing the number of questions chosen by the total stimulus number in each tone and intonation combination in Intrinsic NT, T2, and T4 conditions to enable a visual description of the results.

Results

The binominal ordinary logistic regression model established on the whole dataset showed that Tone, Pitch height, Pitch range and the interactions between all variables (except Pitch range × Original intonation) had a significant effect on the identification of question intonation ($p < 0.0001$; for the full model see **Supplementary Table 1** in **Appendix C**). Tukey *post-hoc* comparisons showed that the perceptual results for Intrinsic NT (43.13% trials identified as question), T2 (86.88% trials identified as question) and T4 (26.01% trials identified as question) all differed significantly from each other (ps < 0.0001).

Binominal ordinary logistic models by tone condition demonstrated that in all three tone conditions, the effects of both Pitch height and Pitch range on intonation identification were significant (ps < 0.0001; for the full models and the interactions between the variables, see **Supplementary Table 2** in **Appendix C**). Moreover, in the Intrinsic NT condition, Original intonation also had a significant effect ($p < 0.0001$). Increases in the average pitch height as well as the manipulation of the range (to the question intonation) both led to more questions identified in all tone conditions, but the specific effects showed tone-specific patterns (**Figure 5**; **Table 6**).

When the **pitch range** of the stimuli became more question-like (i.e., step number increased, illustrated in **Figure 4**), more stimuli were perceived as a question, regardless of lexical tone. However, since there was already a high preference to question identification in the T2 condition at Range Step 1 (i.e., the statement range), the increase in *question identification* brought by changes in range in the T2 condition was restricted compared to the other two conditions. Specifically, differences in intonation identification were significant between the first 3 steps that were more statement-like (i.e., 1 vs. 2, 1 vs. 3 and 2 vs. 3, ps < 0.0005) but not the last 3, more question-like range steps (i.e., 3 vs. 4, 3 vs. 5 and 4 vs. 5). Nevertheless, at each range step, T2 stimuli were interpreted as a question more often than Intrinsic NT and T4 stimuli (ps < 0.0001).

In the Intrinsic NT and T4 conditions, while the question-like manipulation of pitch range led to a much larger increase in *question identification* than in the T2 condition, the trajectories were different. In the Intrinsic NT condition, there was a steady increase in *question identification* with the range steps. Tukey *post-hoc* comparisons demonstrated that the differences in intonation identification between all range steps in the Intrinsic NT condition were significant (ps < 0.0001) except that between Range Step 4 and 5, that is between the most question-like range and the question range. In the T4 condition, a steady increase in *question identification* was observed at Range

Step 4 and 5, but not for the first three more statement-like contours. It is worth mentioning, however, that the differences in intonation identification between all contour steps were statistically significant in the T4 condition (ps < 0.005). Although *question identification* at Range Step 2 was larger than that at Range Step 3, they were all far lower than chance, suggesting that slightly flattened contours still led to a preference for a statement interpretation in the T4 condition. The identification differences between Intrinsic NT and T4 at Range Step 3 and 4 were also significantly different (ps < 0.005).

The increase in **pitch height** also led to more stimuli being identified as a question, and again the increase was much larger in the Intrinsic NT and T4 conditions than in the T2 condition. Tukey *post-hoc* comparisons showed that in the Intrinsic NT condition, except for differences between adjacent steps (i.e., 1 vs. 2, 2 vs. 3, 3 vs. 4, 5 vs. 6, and 6 vs. 7), the identification differences between height steps were all statistically significant (ps < 0.005). In general, an increase in pitch height led to a gradual increase in *question identification* in the Intrinsic NT condition. In contrast, in the T4 condition, significant differences in intonation identification were found between Height Step 1 and Height Steps 4–7 (ps < 0.01), Height Step 2 and Height Steps 4–7 (ps < 0.0001) and Height Step 3 and Height Steps 5–7 (ps < 0.0001), but not within the higher height steps, namely, steps 4–7. In the T2 condition, from Height Step 3, the increasing pitch height seemed to play a predominant role, leading to over 90% *question identification*. The identification difference was significant between Height Step 1 and Height Steps 4–7 (ps < 0.0001), Height Step 2 and Height Steps 4–7 (ps < 0.005), and Height Step 3 and Height Steps 5–7 (ps < 0.0001). Again, at each height step, question interpretations were more frequent in the T2 condition than in the Intrinsic NT and T4 conditions (ps < 0.0001), and at Height Step 2, 3, 6, and 7, question interpretations were more frequent in the Intrinsic NT condition than in the T4 condition (ps < 0.0005).

Discussion of Experiment 2

In Experiment 2, we examined the influence of changes in pitch height and range, and the relative weighting of these cues in intonation perception in Mandarin, using disyllabic Intrinsic NT, T2, and T4 stimuli with T1 as the preceding tone. Although both pitch height and range played important roles in intonation perception, a general effect of the lexical tone on the weightings of the two cues in intonation perception was observed. In general, Intrinsic NT and T4, which were phonetically realized as similar falling contours, showed more similarity to each other than to rising T2. Pitch range played a more important role in question identification in Intrinsic NT and T4, while an increase in pitch height was the primary cue to question intonation in T2. This means that H2 is largely confirmed, though pitch range was a less important cue on T2 than expected. Since the difference in pitch range in T2 was quite small, this finding is not surprising. Unlike the results of Experiment 1, a clear preference for question interpretations was found in the T2 condition in the present experiment, regardless of range or height steps. F_0 manipulation only significantly influenced intonation perception at the first several range and height steps, namely, the more statement-like steps in the T2 condition. From Range Step 3 and Height Step 4 upwards, the percentage of stimuli identified as questions (*question identification)* became higher than 90%, which could be indicative of a ceiling effect. It also possible that the intervals between height steps were not large enough, but further enlargement of the intervals would have made the stimuli sound unnatural, as Height Step 7 in the present study already sounded very high and Height Step 1 very low. This marked preference for question interpretations in the T2 condition may be due to the lack of durational cues in this experiment, or simply, because a rising contour is interpreted as more question-like than other contours. In contrast, a preference for statement interpretations existed in the other two tone conditions, especially in the T4 condition (overall 56.82% in Intrinsic NT and 73.99% in T4), but it was not as strong as the preference for question interpretations in the T2 condition (overall 86.88%). In other words, intonation identification appeared to be especially difficult in utterances ending with a T2 rising contour. On the one hand, the expanded range of the final T2 and the rise in the overall pitch height both led to more stimuli identified as questions. On the other hand, though, the participants were not as sensitive to the more question-like manipulation of f_0 range in the T2 condition as in the other two conditions.

In both the Intrinsic NT and T4 conditions, pitch range played a more important role in intonation perception such that the pairwise identification differences between all range steps reached statistical significance, except Intrinsic NT Range Step 4 vs. 5. Nevertheless, the perception in the Intrinsic NT and T4 conditions also showed some interesting differences. The effects of question-like range manipulation and height were relatively gradual on *question identification* in the Intrinsic NT condition, but showed a sudden rise at step 4 of both the height and range manipulations in the T4 condition. Moreover, it seemed that a stronger flattening of the falling contour was required for a T4 word to be identified as a question than a NT word, as the facilitating effects were more clearly observed in the last two contour steps for T4, while they were already observed at lower steps for NT. In addition, raising pitch height alone did not lead to any preference for question interpretations in the T4 condition. Even when presented with the highest pitch step, participants still tended to identify T1-T4 words as statements unless the falling contour of T4 was reduced at its end, resulting in a reduced f_0 range. Therefore, although f_0 range played a rather important role in intonation perception in the two tone conditions with a phonetically falling movement, it seemed to weigh more heavily as a cue in the T4 condition than in the Intrinsic NT condition. At the same time, other subtle acoustic cues that we did not consider in the present experiment that were hidden in the original recordings (e.g., spectro-temporal differences in sonorous segments between statement and question, see for instance, Coath et al., 2005) might have played a role in the Intrinsic NT condition, since identification in the Intrinsic NT condition was affected by the version of the stimuli that were used for manipulation. In other words, participants showed more sensitivity to more types of cues in the Intrinsic NT condition than in the CT conditions.

To sum up, the findings of Experiment 2 suggested that in short utterances, changes in both overall pitch height and pitch range realized on utterance-final syllables were important cues in question intonation identification in Mandarin. However, the latter cue seemed to weigh more heavily in the identification of intonation for lexical tones that were realized as falling contours than for rising T2, probably because the combined cues made the difference between statements and questions particularly salient, while the intonational contrast is primarily signaled by a difference in height in the T2 condition.

GENERAL DISCUSSION

The present study investigated the perception of intonation type (question vs. statement) on Mandarin NT in comparison to representative CTs, the mid-rising T2 and the high-falling T4. Although Intrinsic NT and Derived NT differ from each other in their phonological representations on some accounts (e.g., Zhang, 2021) as discussed in the introduction, Experiment 1 showed that intonation identification in these two conditions was as highly accurate as in the T4 condition, which was significantly more accurate than that in the T2 condition. The acoustic analyses of the stimuli in Experiment 1 revealed that question intonation was always marked by a higher overall f_0 level, but this was accompanied by a decrease of the f_0 range for the falling contours of NTs and T4 (manifested as a higher ending of the falling contour), while the T2 rise only changed in that it became slightly steeper. Experiment 2 showed that pitch range was the most important cue in the T4 condition and also important in the Intrinsic NT condition, while pitch height played a role in the T2 condition. These results confirm that the reduced f_0 range on the surface plays a more important role in intonation perception in Mandarin NT words than any possible tone-intonation interaction at a phonological level. The present findings shed light on the intonation perception mechanisms in Mandarin as well as the phonetic targets of both types of NT.

That the phonetic tonal realization can be modified to such an extent may be due to the relatively simple tonal system of Mandarin. The flattening of the falling tone or the raising of a level tone would hardly lead to misidentification of lexical tones (Liu et al., 2016), but can be used to alert the listener that there is an intonational event going on. In syllable tone languages with a more complex tonal system like Cantonese, the effect of intonation on the realization of lexical tones often leads native listeners to misidentify them as other lexical tones rather than facilitating intonation perception (Kung et al., 2014). This difference in complexity may also explain why Cantonese (and middle ancient Chinese which had 4 tonal contours and 2 tonal registers) has retained a much richer inventory of modality particles than modern Mandarin, and why they are not reduced to the NT-bearing syllables of Mandarin.

The acoustic analysis of stimuli in Experiment 1 also showed that, despite their difference in phonological tonal representation, both types of NT surface with similar phonetic forms in declarative utterances. In addition, both Intrinsic NT and Derived NT maintain a slightly falling contour in questions as short as disyllabic words, suggesting that they may share a unique phonetic target, namely, a mid static target according to Chen and Xu (2006) that is different from the tonal targets found for the four CTs.

To conclude, the investigation of intonation perception on different types of NT in the present study allows us to attribute the tone-specific pattern found in Mandarin to the phonetic realization rather than the phonological interaction between lexical tone and intonation. It would be interesting to investigate to what extent our findings generalize to preceding tones other than T1 and other intonations, or longer utterances.

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 Ethics Committee of the Faculty of Modern and Medieval Languages and Linguistics (MMLL), University of Cambridge. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YZ, ES, and BP contributed to the conception of the study, experimental design, and manuscript preparation. YZ carried out the experiment and data analysis. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

YZ would like to thank the CHINA Scholarship COUNCIL (CSC) and Cambridge Trust for their doctoral scholarship.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.849132/full#supplementary-material

REFERENCES

Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychos. Med.* 66, 411–421. doi: 10.1097/00006842-200405000-00021

Boersma, P., and Weenink, D. (2021). *Praat: Doing Phonetics By Computer [Computer program]*. Version 6.1.51. Available online at: http://www.praat.org/ (accessed July 22, 2021).

Cao, J. F. (1986). Acoustic features of Neutral Tone in Standard Mandarin. *Appl Phonetics* 4, 1–6.

Chen, Y., and Xu, Y. (2006). Production of weak elements in speech – evidence from f0 patterns of neutral tone in standard Chinese. *Phonetica* 63, 47–75. doi: 10.1159/000091406

Coath, M., Brader, J. M., Fusi, S., and Denham, S. L. (2005). Multiple views of the response of an ensemble of spectro-temporal features support concurrent classification of utterance, prosody, sex and speaker identity. *Network Comput. Neural Syst.* 16, 285–300. doi: 10.1080/09548980500290120

Field, A., Miles, J., and Field, Z. (2012). *Discovering Statistics Using R.* Great Britain: Sage Publications, Ltd, 958.

Hothorn, T., and Everitt, B. S. (2006). *A Handbook of Statistical Analyses Using R.* Boca Raton: CRC Press.

Huang, T., and Johnson, K. (2010). Language specificity in speech perception: Perception of Mandarin tones by native and nonnative listeners. *Phonetica*. 67, 243–267.

Kung, C., Chwilla, D. J., and Schriefers, H. (2014). The interaction of lexical tone, intonation and semantic context in on-line spoken word recognition: an ERP study on Cantonese Chinese. *Neuropsychologia* 53, 293–309. doi: 10.1016/j.neuropsychologia.2013.11.020

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *J. Statist. Software* 82, 1–26. doi: 10.18637/jss.v082.i13

Lee, W.-S., and Zee, E. (2014). "Chinese phonetics," in *The Handbook of Chinese Linguistics*, eds C. J. Huang, Y. A. Li, and A. Simpson (Hoboken: John Wiley & Sons), 367–399.

Li, Z. (2003). *The phonetics and phonology of tone mapping in a constraint-based approach*. (Doctoral dissertation). Massachusetts Institute of Technology.

Liang, J., and Heuven, V. J. (2009). "Chinese tone and intonation perceived by L1 and L2 listeners," in *Experimental Studies in Word and Sentence Prosody* (De Gruyter Mouton), 27–62.

Lin, M. C., and Yan, J. Z. (1980). Acoustic features of Neutral Tone in Mandarin. *Dialects* 3, 166–178.

Lin, T. (1983). "A primary test on neutral tones in Beijing Mandarin," in *Phonetic experiments on Beijing Dialect*, eds T. Lin and L. J. Wang (Beijing: The Peking University Publishing House), 1–26.

Liu, F., and Xu, Y. (2005). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica* 62, 70–87. doi: 10.1159/000090090

Liu, F., and Xu, Y. (2007). "The neutral tone in question intonation in Mandarin," in *Eighth Annual Conference of the International Speech Communication Association*.

Liu, M., Chen, Y., and Schiller, N. O. (2016). Online processing of tone and intonation in Mandarin: evidence from ERPs. *Neuropsychologia* 91, 307–317. doi: 10.1016/j.neuropsychologia.2016.08.025

Macmillan, N. A., and Creelman, C. D. (2004). *Detection Theory: A User's Guide.* New York, NY: Psychology Press.

Pallier, C. (2002). *Computing Discriminability and Bias With the R Software*. Available online at: http://www.pallier.org/ressources/aprime/aprime (accessed April 22, 2022).

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., et al. (2019). PsychoPy2: experiments in behavior made easy. *Behav. Res. Methods*. 51, 195–203. doi: 10.3758/s13428-018-01193-y

Peng, S. H., Chan, M. K., Tseng, C. Y., Huang, T., Lee, O. J., and Beckman, M. E. (2005). "Towards a Pan-Mandarin system for prosodic transcription," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, ed S. A. Jun (OUP Oxford), 230–270. doi: 10.1093/acprof:oso/9780199249633.003.0009

Pollack, I., and Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Sci.* 1, 125–126. doi,: 10.3758/BF03342823

R Core Team (2020). *R: A Language and Environment for Statistical Computing.*Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/ (accessed April 22, 2022).

Ren, G. Q., Tang, Y. Y., Li, X. Q., and Sui, X. (2013). Pre-attentive processing of Mandarin tone and intonation: evidence from event-related potentials. *Funct. Brain Mapp. Endeav. Understand Working Brain* 6, 95–108. doi: 10.5772/56503

Shen, J. (1992a). "Hanyu yudiao moxing chuyi" [On Chinese intonation model]. *Yuwen Yanjiu* 45, 16–24.

Shen, X. (1989). *The Prosody of Mandarin Chinese*. Berkeley: University of California Press, 9–30.

Shen, X. S. (1992b). Mandarin neutral tone revisited. *Acta Linguistica Hafniensia* 24, 273. doi: 10.1080/03740463.1992.10412273

Smith, W.D. (1995). Clarification of sensitivity measure A. *J. Math. Psychol.* 39, 82–89. doi: 10.1006/jmps.1995.1007

Stanislaw, H., and Todorov, N. (1999). Calculation of signal detection theory measures. *Behav. Res. Methods Instr. Comput.* 31, 137–149. doi: 10.3758/BF03207704

The National Bureau of Quality and Technical Supervision (1996). *National Standard of the People's Republic of China: The Usage of Punctuation*. Beijing: Standards Press of China.

Tukey, J. W. (1977). *Exploratory Data Analysis.* 2, 131–160. Available online at: http://theta.edu.pl/wp-content/uploads/2012/10/exploratorydataanalysis_tukey.pdf (accessed June 17, 2022).

Wang, J. (1996). "An acoustic study of the interaction between stressed and unstressed syllables in spoken Mandarin," in: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96 (Vol. 3)*. Philadelphia: IEEE, 1616–1619.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Commun*. 46, 220–251. doi: 10.1016/j.specom.2005.02.014

Yip, M. (2002). *Tone*. Cambridge, MA: Cambridge University Press.

Yuan, J. (2004). "Perception of Mandarin intonation," in *2004 International Symposium on Chinese Spoken Language Processing* (Paris: IEEE). 45–48.

Yuan, J. (2006). "Mechanisms of question intonation in Mandarin," in *International Symposium on Chinese Spoken Language Processing.* Berlin, Heidelberg: Springer, 19–30.

Yuan, J. (2011). Perception of intonation in Mandarin Chinese. *J. Acoust. Soc. Am.* 130, 4063–4069. doi: 10.1121/1.3651818

Yuan, J., Shih, C., and Kochanski, G. P. (2002). "Comparison of declarative and interrogative intonation in Chinese," in *Speech Prosody 2002, International Conference* (Aix-en-Provence).

Zhang, J., and Mueller, S. T. (2005). A note on ROC analysis and non-parametric estimate of sensitivity. *Psychometrika* 70, 203–212. doi: 10.1007/s11336-003-1119-8

Zhang, Y. (2021). *Neutral tone in Mandarin: representation and interaction with utterance-level prosody*. (Doctoral dissertation). University of Cambridge, Cambridge, United Kingdom.

Zhang, Y. (2018). "Anticipatory dissimilation in (non-clitic) neutral tones in Mandarin," in: *Tone and Intonation in Europe 2018, International Conference* (Stockholm).

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: 12 July 2022
doi: 10.3389/fpsyg.2022.902569

The Sequence Recall Task and Lexicality of Tone: Exploring Tone "Deafness"

Carlos Gussenhoven[1,2]*, Yu-An Lu*[2]*, Sang-Im Lee-Kim*[2]*, Chunhui Liu*[3]*, Hamed Rahmani*[1]*, Tomas Riad*[4] *and Hatice Zora*[5]*

[1]*Centre for Language Studies, Radboud University, Nijmegen, Netherlands,* [2]*Department of Foreign Languages and Literatures, National Yang Ming Chiao Tung University, Hsinchu, Taiwan,* [3]*College of Literature and Journalism, Sichuan University, Chengdu, China,* [4]*Department of Swedish Language and Multilingualism, Stockholm University, Stockholm, Sweden,* [5]*Department of Neurobiology of Language, Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands*

Many perception and processing effects of the lexical status of tone have been found in behavioral, psycholinguistic, and neuroscientific research, often pitting varieties of tonal Chinese against non-tonal Germanic languages. While the linguistic and cognitive evidence for lexical tone is therefore beyond dispute, the word prosodic systems of many languages continue to escape the categorizations of typologists. One controversy concerns the existence of a typological class of "pitch accent languages," another the underlying phonological nature of surface tone contrasts, which in some cases have been claimed to be metrical rather than tonal. We address the question whether the Sequence Recall Task (SRT), which has been shown to discriminate between languages with and without word stress, can distinguish languages with and without lexical tone. Using participants from non-tonal Indonesian, semi-tonal Swedish, and two varieties of tonal Mandarin, we ran SRTs with monosyllabic tonal contrasts to test the hypothesis that high performance in a tonal SRT indicates the lexical status of tone. An additional question concerned the extent to which accuracy scores depended on phonological and phonetic properties of a language's tone system, like its complexity, the existence of an experimental contrast in a language's phonology, and the phonetic salience of a contrast. The results suggest that a tonal SRT is not likely to discriminate between tonal and non-tonal languages within a typologically varied group, because of the effects of specific properties of their tone systems. Future research should therefore address the first hypothesis with participants from otherwise similar tonal and non-tonal varieties of the same language, where results from a tonal SRT may make a useful contribution to the typological debate on word prosody.

Keywords: word prosody, lexicon-based memory, tone contrast salience, tone language, semi-tonal language, sequence recall task

Edited by:
Juhani Järvikivi,
University of Alberta, Canada

Reviewed by:
Ting Zou,
Beijing Foreign Studies University, China
Ratree Wayland,
University of Florida, United States

Correspondence:
Hatice Zora
hatice.zora@mpi.nl

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

Received: *23 March 2022*
Accepted: *16 May 2022*
Published: *12 July 2022*

Citation:
Gussenhoven C, Lu Y-A, Lee-Kim S-I, Liu C, Rahmani H, Riad T and Zora H (2022) The Sequence Recall Task and Lexicality of Tone: Exploring Tone "Deafness". Front. Psychol. 13:902569. doi: 10.3389/fpsyg.2022.902569

INTRODUCTION

Lexical tone has been investigated in a large body of perception research and is a prominent traditional typological concept in phonology, perhaps more so than word stress, which until recently was often treated as a universal (*cf.* van Heuven and Turk, 2020). Tones can form a great variety of subsystems in the phonologies of languages. There can be few or many of them and contrasts will vary in salience. Functionally, they could share the phonological specification of morphemes with vowels and consonants ("lexical tones") or be their sole exponents ("grammatical tones," Hyman, 2011, 2016). While the linguistic and cognitive evidence for lexical tone is beyond dispute, as indicated by the results of dichotic listening, categorical perception, ABX designs, and brain response registrations (Lau et al., 2020), the word prosodic systems of many languages continue to escape the categorizations of typologists, with frequent debates about the categorization of tone languages (Hyman, 2006; Kehrein et al., 2017; Steien and Yakpo, 2020; Gooden, 2022). The present paper aims to contribute to the understanding of the lexical status of tone by comparing non-tonal, semi-tonal, and tonal languages in a Sequence Recall Task (SRT). It was developed by Emmanuel Dupoux and colleagues as a diagnostic for the presence of word stress in a language (Dupoux et al., 2001). It followed their earlier speculations on why French listeners underperformed in an ABX task relative to Spanish listeners, where A and B were trisyllabic non-words differing in the location of stress (Dupoux et al., 1997). An SRT trial presents participants with a sequence of some 4 to 6 disyllabic non-words which have a prominence on either one or another of its syllables, as in the disyllabic non-word sequence *númi – numí – númi – númi*. Participants are asked to reproduce the order of the two non-words on a keyboard (in this case 1–2–1–1) after hearing a distracting sound immediately after the sequence, intended to prevent them from relying on their acoustic memory (cf. Baddeley, 2010). Speakers of Spanish, a language with contrastive word stress, outperformed speakers of French on this task, which language has phrasal stress (Dupoux et al., 2001). The effect survives language contact as in L2 learning (Dupoux et al., 2008).

Explanations of the inability of French listeners to perform the task as effectively as Spanish listeners first addressed the exposure to meaningful word prosody during language acquisition, but later shifted to the resulting abstract lexical representation of stress (Peperkamp, 2004; Dupoux et al., 2008). Providing support for this interpretation, Rahmani et al. (2015) showed that the presence of syllabic prominence in lexical representations, whether from tone or stress, explained the results of an experiment with five language groups, Dutch, Japanese, French, Indonesian, and Persian. As hypothesized, Dutch and Japanese participants outperformed the participants in the other three language groups, who for that reason are "stress-deaf" (the term is due to Dupoux et al., 1997). The explanation the authors give is that Dutch and Japanese participants could engage their lexicon-based memory on the basis of the contrastive location of a syllabic prominence in words, stress in Dutch and a HL melody in Japanese. The interpretation of stress as tone by the Japanese listeners was also evident in Qin et al. (2017), in which Standard Mandarin, Taiwan Mandarin, and English participants achieved comparable SRT performance on disyllabic English stress pairs. None of the other three languages in Rahmani et al. (2015) possesses lexically contrastive word prosody, whether due to stress or tone, so that any reliance on a "lexical memory" is not an option open to them.

The similar effects of stress and tone in the Dutch and Japanese accuracy scores in Rahmani et al. (2015) must not lead us to lose sight of the profoundly different character of tone from stress. Tones can form a great variety of subsystems in the phonologies of languages. There can be few or many of them and contrasts will vary in salience. And they could be lexical as well as morphological or syntactic ('grammatical'). Stress, by contrast, is usually taken to be the head of a constituent of the prosodic hierarchy, the foot, in which unstressed syllables may additionally occur in non-head positions (Selkirk, 1980; Hayes, 1995). Since all words are footed, and hence stressed, no stress contrasts are possible on monosyllables if a language has feet ("obligatoriness," Hyman, 2006). This is why the non-words in a stress-based SRT are disyllabic: stressed–unstressed or unstressed–stressed. At the same time, this makes it necessary to use monosyllabic contrasts in the case of tone, in order to guarantee tonal interpretations of the pitch contrasts. It is true that stress systems too vary across languages, for instance in the degree of exceptionality of stress locations. Moreover, stressed syllables may or may not have an intonational pitch accent, as in Germanic languages (*cf.* "primary stress," Domahs et al., 2008), and stress may correlate with syllable quantity or vowel reduction (Hayes, 1995). Such differences have not affected the results of SRTs much. In Peperkamp and Dupoux (2002), an experiment with six language groups, Polish, which has regular penultimate stress with few words having ultimate or antepenultimate stress, came out as intermediate between a stress-deaf and a non-stress-deaf group. Also, the categorical interaction between vowel quality and stress in European Portuguese explains why listeners are stress-deaf if they cannot rely on the vowel quality differences (Correia et al., 2015; Lu et al., 2018).

Because of the more varied complexity of lexical tone systems compared to stress systems, we may reasonably expect the results of a tonal SRT to be affected by relevant features of a language's phonology (Best, 2019). First, the number of monosyllabic tone melodies may vary from 2 to as many as 9 (e.g., Hyman, 2011). A high functional load of lexical pitch contrasts may well affect recall accuracy. Moreover, tone contrasts may be restricted to certain positions in the word, like the final syllable in Ma'ya (Remijsen, 2002) or a non-final syllable in Swedish (Riad, 2014: 182). This means that in addition to a simple discrete concept of lexical "tonality," that is, the presence of a pitch specification in the phonological form of at least some morphemes (Hyman, 2006), it will be necessary to test for effects of relative "tonality," that is, the complexity of lexical tone systems. Second, the choice of the pitch contrast in the experiment may favor participants that happen to have that contrast in their tonal grammar. We take this potential benefit to be independent of the lexical or intonational status of the pitch contrast. An experiment that intends to include this factor in its design, will need to test for a number of pitch contrasts, such that each of them fails to turn up in at least one language under investigation. Third, pitch contrasts vary in salience, that is, in the perceptual difference between the two contrasting pitch shapes. If sequences of less salient contrasts are harder to recall than contrasts with larger differences, the size of the contrast will need to be included as a variable in our experiment.

We selected one unambiguously non-tonal language (Indonesian), one borderline case (Stockholm Swedish), and two unambiguously tonal languages (Taiwan Mandarin and Zhumadian Mandarin). The inclusion of two similar tone languages served as a sanity check, as it predicts that their scores will be quite similar as well as quite different from the non-tonal language. A heuristic element in our choice of languages is the ambiguous "semi-tonal" language, which might statistically side with either the non-tonal language or the tonal ones, or appear as a category in between.

Indonesian has neither tone nor stress on any syllable, whether word-based or phrase-based (Odé, 1994; Goedemans and van Zanten, 2007; Maskikit-Essed and Gussenhoven, 2016). The performance of the Indonesian participant group should provide a lower baseline. The language has an intonational contrast between a phrase-final rise, used in pre-final intonational phrases and in final interrogative phrases, and a rise–fall, used in final declarative phrases. The contrast between these right-edge melodies will show up in stated and questioned monosyllabic words. **Figure 1** shows this contrast as spoken by a 28-year-old male speaker from East Java. This pitch contrast is the main intonational contrast in the language and there may therefore be a fair bit of variation in the phonetic shapes.

Stockholm Swedish has a lexical tone contrast in non-final syllables with word stress, Accent 1 vs. Accent 2, as occurring in *anden* "the duck" and *anden* "the spirit," respectively. Accent 1 is a rise in the stressed syllable, followed by low pitch when occurring in the nuclear position, as illustrated by the solid line of an isolated pronunciation of the expression meaning "the duck" in **Figure 2**. Accent 2 has an early fall in the stressed syllable, which in the nuclear position is followed by a pitch peak in the phrase-final syllable, as shown by the dashed line for an isolated pronunciation of the expression meaning "the spirit" in **Figure 2**. Both have an intonational melody LHL%, which is preceded by a lexical H in the case of Accent 2, effectively shifting the intonational f0 peak onto the final syllable (Riad, 2014). Arguably, the different contours in the unstressed phrase-final syllables represent contrasting phonetic cues to the tone contrast on the penultimate syllable. However, such contextual cues abound in languages generally, so that we cannot interpret the phrase-final pitch difference as a contrast of the language, whether lexical or intonational.

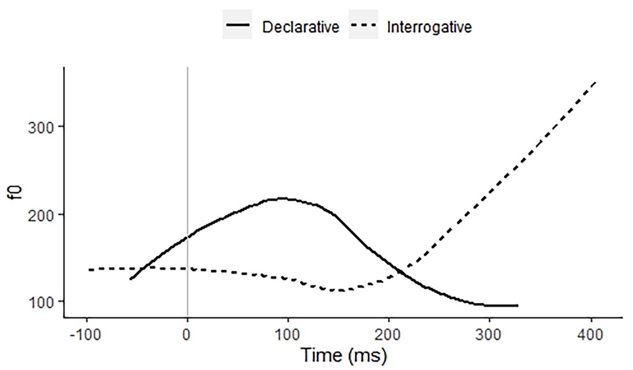


FIGURE 1 | f0 contours of declarative (solid line) and interrogative (dashed line) citation pronunciations of the monosyllabic word *gong* ("gong"), recorded by a 28-year-old male speaker of Standard Indonesian.

Zhumadian Mandarin, spoken in Henan Province, China, has four lexical tones, two rises, and two falls, which contrast for temporal alignment, leading to a late rise (Tone 1), a late fall (Tone 2), an early rise (Tone 3), and an early fall (Tone 4). The early rising Tone 3 tends to rise only a little, thus resembling Tone 1 of Standard Mandarin, while the late rising Tone 1 may sound like a final, dipping Tone 3 of Standard Mandarin (Gussenhoven and van de Ven, 2020). The language has a Fourth Tone Sandhi rule, changing 4+4 into 1+4, as well as toneless morphemes, that is, neutral tone. **Figure 3** presents examples of the four tones on the syllable /mae/. Younger speakers are bilingual with Standard Mandarin. Except in educational contexts, speakers use the Zhumadian dialect.

Taiwan Mandarin is a standard variety of Mandarin. It has four lexical tones, a high level tone, a rising tone, a low tone, and a high falling tone, Tones 1 to 4, respectively (**Figure 4**). In addition, it has the Third Tone Sandhi rule (3+3 → 2+3) as well as syllables with neutral tone, whose pitch contours are derivative from a preceding toned syllable. The most striking difference with Standard Chinese is the shorter duration of Tone 3, which typically lacks or significantly reduces the rising part in phrase-final position (Kubler, 1985; Fon and Chiang, 1999; Torgerson, 2005; Deng et al., 2006). Its tonal complexity is quite comparable to that of Zhumadian Mandarin.

MATERIALS AND METHODS

We included three-pitch contrasts in the experiment, EarlyFall vs. LateFall, EarlyRise vs. LateFall, and RiseFall vs. EarlyRise. None of these are pitch levels, which are likely to sound like a melody when occurring in a sequence, which would be more memorable than sequences of pitch shapes. In addition, we used a "phoneme" contrast of the type that has served as a control variable in SRT experiments (Peperkamp et al., 2010; Rahmani et al., 2015; Qin et al., 2017). A phonetically trained speaker of Dutch in his early 70s recorded each of these seven syllable types at least eight times in a sound-treated booth. Three tokens of each syllable type were selected that sounded natural and seemed good exemplars of the intended pitch shape. **Figure 5** displays these tokens for all five-pitch shapes figuring in these contrasts, all pronounced on the syllable [la], aligned at the onset-vowel boundary indicated by the gap in the figure, which corresponds to 0 ms in the signal. The phoneme contrast was between the syllables [ta] and [la], both pronounced with level midpitch. We avoided adjustments of the original durations, unlike Peperkamp et al. (2010), who drastically shortened the original recordings of disyllables. Largely depending on pitch shape, tones require a certain duration to produce (Xu and Sun, 2002) and shortened syllables may as a result sound distorted. Across pitch shape types, durations varied from

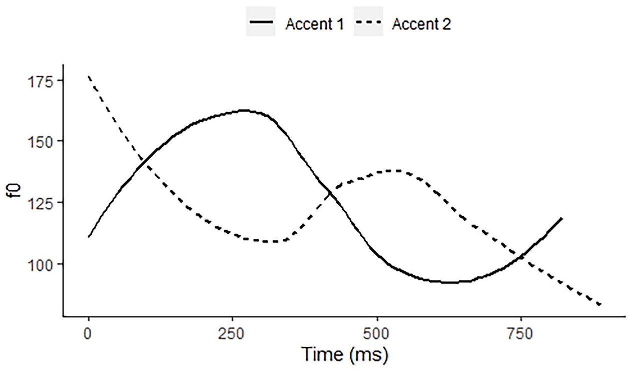


FIGURE 2 | f0 contours of citation pronunciations of Accent 1 on anden "the duck" (solid line) and Accent 2 on anden "the spirit" (dashed line) by a 60-year-old male speaker of Stockholm Swedish.

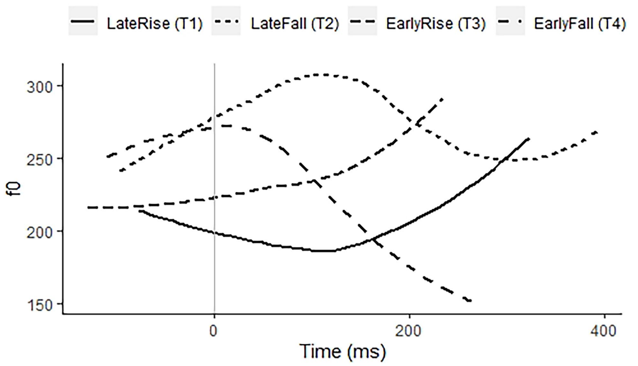


FIGURE 3 | f0 contours of citation pronunciations of a late rise/Tone 1 on 麥 "cereal," a late fall/Tone 2 on 埋 "bury," an early rise/Tone 3 on 買 "buy," and an early fall/Tone 4 on 賣 "sell," all with the segmental syllable /mae/, recorded by a 22-year-old female speaker of Zhumadian Mandarin.

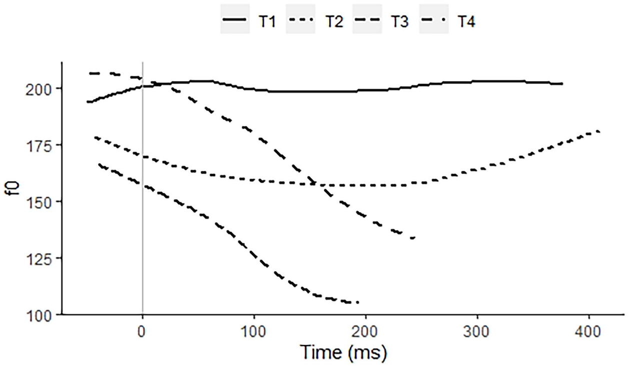


FIGURE 4 | f0 contours of citation pronunciations of a high level/Tone 1 on 媽 "mother," a rise/Tone 2 on 麻 "hemp," a low tone/Tone 3 on 馬 "horse," and a high falling/Tone 4 on 罵 "scold," all with the segmental syllable/ma/, recorded by a 40-year-old female speaker of Taiwan Mandarin.

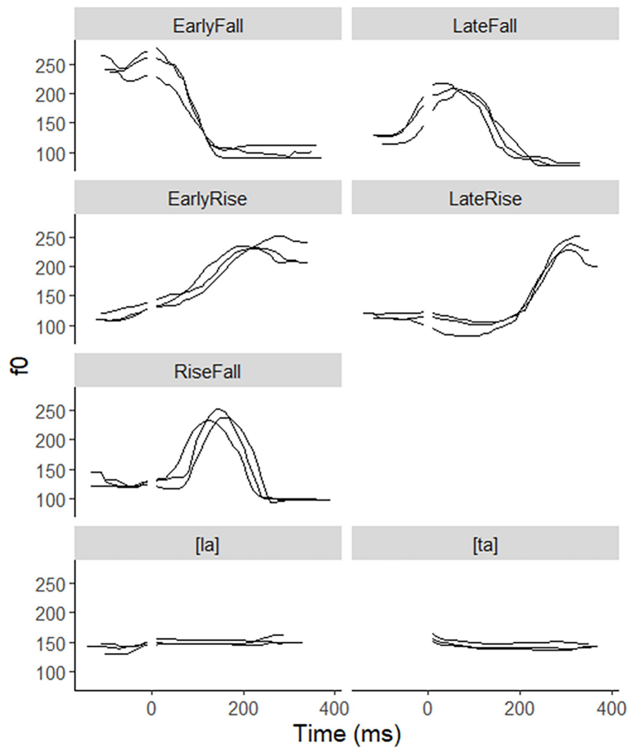


FIGURE 5 | f0 tracks of the three tokens for each of 7 syllables types, with the onset-vowel boundaries indicated by an interruption.

430 ms for a token of the EarlyRise to 569 ms for a token of the RiseFall. The three tokens had very similar durations in three of the five-pitch shape types. Only the triplets for the EarlyFall and the LateFall varied more noticeably, for which reason we standardized the three exemplars to the rounded mean duration in each triplet, 440 ms and 460 ms, respectively, using Praat (Boersma and Weenink, 1992–2020). **Figure 6** shows acoustic durations of all 15 pitch shape stimuli and the 6 stimuli for the phoneme contrast, for onset consonant and vowel separately; in the case of [ta], the burst duration is shown.

A number of independent variables were included in the analysis. Sex and aptitude were the two participant variables, of which aptitude was motivated by the expectation that participants may vary in their aptitude for carrying out an SRT. For this variable we used each participant's mean accuracy score on the phoneme contrast. Rather than controlling for pitch discrimination and categorization abilities, which have been shown to explain variation in pitch-related learning and identification tasks (*cf.* Sadakata and McQueen, 2014; Zhao and Kuhl, 2015; Bowles et al., 2016; Qin et al., 2021; Rhee et al., 2021), we intended to control for a more general ability to perform the experimental task of remembering sequences of tokens of two sound categories. Earlier research had taken this effect for granted, by subtracting phoneme accuracy scores from stress contrast scores (e.g., Peperkamp et al., 2010). We felt we needed to have a better understanding of the relation between the control and experimental contrasts in view of the prospect of continued research on languages with older populations of speakers.

Four language variables figured in our investigation, lexicality, tonecomplexity, salience, and havecontrast.

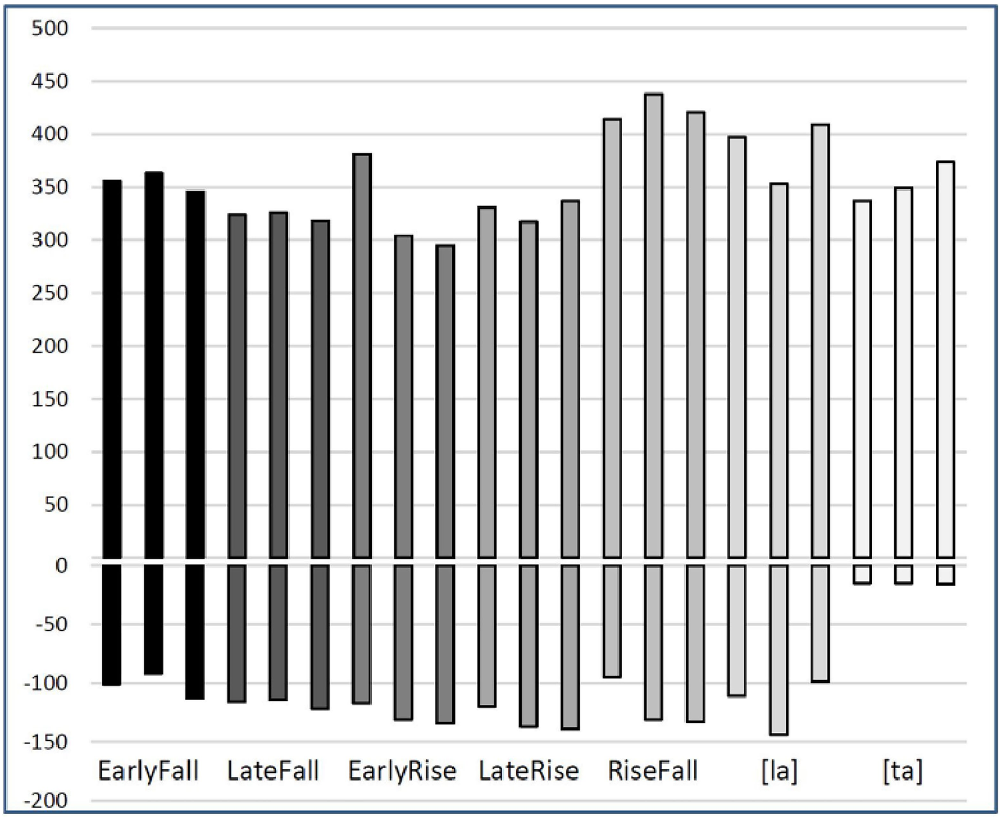


FIGURE 6 | Durations of onset [l] (negative bars) and rhyme [a] (positive bars) of the 27 stimuli in the experiment. For [ta], the negative bars give the positive VOTs. The value for the onset in [ta] is the burst and friction of the released [t].

Since our main hypothesis was that participants with tonal language backgrounds will outperform participants with non-tonal language backgrounds, we interpret LEXICALITY as a binary variable characterizing any language with a lexical marking of pitch as tonal (Hyman, 2006), which includes "semi-tonal" Swedish. While the distribution of the two Swedish tone categories is highly predictable from the phonology and morphology of words (Bruce, 1977: 18; Wetterlin et al., 2007; Riad, 2014: 183), there are exceptions, most obviously in disyllables with penultimat stress. For instance, many loan words have Accent 1, like *ketchup* and *solo*, in contrast to other words, like *senap* "mustard" and *pizza*, which have Accent 2. Moreover, in a priming experiment, Althaus et al. (2021) have shown that native speakers use the contrast in lexical access. Accordingly, only Indonesian was coded as −1 and the other three as 1 for this variable. At the same time, a gradient characterization of lexical tone complexity might provide a better predictor of accuracy scores than binary LEXICALITY, for which reason we coded the two Mandarin varieties as 4.0 for TONECOMPLEXITY, to reflect the number of tone categories. While Swedish has two tone categories, it has no tone contrast on monosyllabic words and hence not for the monosyllabic non-words in our experiment. We coded it as 0.5, while Indonesian was coded 0.0. Because LEXICALITY and TONECOMPLEXITY amount to discrete and gradual interpretations of a language's status as a tone language, we will not include both variables in the same analysis.

Our experiment involved pitch contrasts that obviously varied in salience. Because sequences of similar pitch shapes may be harder to recall than sequences of more different pitch shapes, we measured subjective phonetic differences among six-pitch shapes, one token of each of the five-pitch shapes in our experiment plus a FallRise, spoken by the same speaker, for the sake of symmetry in the set of pitch shapes to be measured. The 6 × 5 pairs were included in a Praat Multiple-Forced Choice experiment together with two filler pairs, presented in a per participant randomized order. Eight phonetically trained judges were asked to rate all pairs for phonetic distance on a 10-point scale, after listening to recordings of all six-pitch shapes and rating three trial pairs. Pearson's correlation coefficients between the scores of each judge and the mean score over all judges showed that the scores by two judges failed to reach significance at a 5% level. Of the other six, two judges had $r < 0.55$ and four $r > 0.83$.[1] They were native speakers of Dutch, English, Korean, and Mandarin

[1] In a methodologically comparable experiment with 40 participants, no effect of order of presentation within a pair was found (Fournier and Gussenhoven, 2010). In that experiment, the scores of all participants correlated with the mean scores over all judges, with a range of 0.56–0.88.

TABLE 1 | Mean subjective phonetic distances per pair of pitch shapes.

	EarlyFall	LateFall	EarlyRise	LateRise	RiseFall	FallRise
LateFall	**2.3**					
EarlyRise	9.5	**8.3**				
LateRise	9.5	8.7	4.3			
RiseFall	7.5	6.8	8.3	**8.7**		
FallRise	8.6	8.0	8.8	8.3	7.2	

Experimental contrasts are printed in bold.

TABLE 2 | Experimental pitch contrasts functioning as phonological contrasts.

Contrast	Indonesian	Swedish	Zhumadian Mandarin	Taiwan Mandarin
EarlyFall vs. LateFall	−1	−1	1	−1
LateFall vs. EarlyRise	−1	−1	1	1
LateRise vs. RiseFall	1	−1	−1	−1

(3), with ages ranging from 27 to 47. The native speaker of Korean grew up speaking tonal Gyeongsang Korean, but uses Standard Korean in virtually all domains. Their language backgrounds were otherwise evenly divided over tonal and non-tonal languages, which minimized language-specific biases (*cf.* Huang and Johnson, 2010). **Table 1** presents all scores, pooled over the two orders in each pair, which we used as the salience scores.

Finally, in order to be able to assess the extent to which the presence of a contrast in the participant's native language influences accuracy scores, we coded languages for HAVECONTRAST for each pitch contrast. When a language has a contrast in its lexical or postlexical phonology, it is coded 1 for that contrast, otherwise −1. For instance, Zhumadian Mandarin has a lexical contrast between an early aligning and a late-aligning fall, while the other three languages do not, entitling it to a 1 coding for that contrast (see **Table 2**). It also has a contrast between a late fall and an early rise, corresponding to our second experimental contrast. Taiwan Mandarin has a contrast between a fall and a late rise. Native speaker reactions suggest that the EarlyFall and the LateFall are equally good exemplars of the Taiwan Mandarin Fall, while the EarlyRise and the LateRise are both good exemplars of the Taiwan Mandarin Rise. We therefore also coded both Zhumadian and Taiwan Mandarin as 1 for the LateFall vs. EarlyRise contrast. Indonesian has an intonational contrast between a LateRise and a RiseFall, while the other three languages do not. Swedish lacks monosyllabic contrasts, so that it is harder to define the occurrence of our experimental contrast in the phonology of Swedish. Even if we were to interpret the f0 shapes of the first syllables as a RiseFall for Accent 1 and an EarlyFall for Accent 2 (see **Figure 2**), this would not correspond to any of the experimental pitch contrasts. Accordingly, all three contrasts are coded as −1 for Swedish.

We employed two sequence lengths for the two non-words, a 4-non-word and a 5-non-word sequence length, giving a binary variable SEQUENCELENGTH. Piloting with 6-non-word sequences made it clear that these were too difficult to deal with. In addition, we found that the task required a high level of concentration, which we felt put strict limits on the time participants could be asked to perform it. In a further attempt to make the task easier, we blocked the 4-non-word and presented these before moving on the block of 5-non-word sequences. Finally, GROUP and CONTRAST were the variables of central interest in the investigation. A summary of the independent variables introduced above appears in **Table 3**. Sequences of non-words avoided regular alternations (e.g., 1,212) and maximized the number of switch points (1 to 2, 2 to 1), following Rahmani et al. (2015), which led us to use 1211, 1221, 2112, 2122, 2212 and 1121 for 4-word sequences and 11221, 12112, 12212, 22112, 21221 and 21121 for 5-word sequences. With four contrasts and twice six sequences the total number of trials was 48. The total duration of the experiment was about 30 min.

We recruited minimally 20 participants for each language who were between 18 and 30 years old and attended or had attended institutes of tertiary education. **Table 4** lists the numbers per language split over the sexes, their age ranges, mean ages, and recruitment locations. We presented the experiment on a desktop computer with E-Prime 3.0 for the Zhumadian Mandarin participants and E-Prime 2.0 for the other participants (Schneider et al., 2012). Participants listened individually to the stimuli through headphones. Instructions were provided in English on the screen, supplemented with oral instructions in each native language. The experiment consisted of four blocks, one for each of the four contrasts with breaks in between, in a randomized order for each participant. Each block started with a training session. For the phoneme contrast, participants were trained to associate the syllable [la] with key "1" and [ta] with key "2," while for the three-pitch contrasts they were trained to associate [LateFall] with key "1" and [EarlyFall] with key "2," [LateFall] with key "1" and [EarlyRise] with key "2," and [LateRise] with key "1" and [RiseFall] with key "2."

Participants were told at the beginning of each block that they were going to learn two words in a foreign language. First, they heard all three tokens of one non-word with a "1" displayed on the screen, and then heard all three tokens of the other non-word with a "2" displayed on the screen. This cycle was repeated three times, exposing participants to 3 tokens x 2 non-words x 3 repetitions, or 18 non-words, before they proceeded to the second training stage, during which they heard each of the 6 tokens, together with a display of the corresponding key

TABLE 3 | Independent variables in the investigation.

Variable		Description
Experimental design	GROUP	Indonesian, Swedish, Zhumadian Mandarin, Taiwan Mandarin
	CONTRAST	EarlyFall vs. LateFall, LateFall vs. EarlyRise, LateRise vs. RiseFall, [la] vs. [ta]
	SEQUENCELENGTH	4-word sequence—1, 5-word sequence 1
Participant	SEX	Female—1, Male 1
	APTITUDE	Accuracy score [la]-[ta]
Linguistic structure	LEXICALITY	Indonesian—1, all other groups 1
	TONECOMPLEXITY	Indonesian 0.0, Swedish 0.5, Zhumadian Mandarin 4.0, Taiwan Mandarin 4.0
	HAVECONTRAST	See detailed coding in Table II.
	SALIENCE	EarlyFall vs. LateFall 2.3, LateFall vs. EarlyRise 8.3, LateRise vs. RiseFall 8.7

TABLE 4 | Participants in four language groups.

	N	Age range	Mean age	Location
Indonesian	10F, 10M	19–30	24.4	National Yang Ming Chiao Tung University (Hsinchu, Taiwan)
Swedish	11F, 10M	20–29	24.1	Stockholm University (Sweden)
Zhumadian M	15F, 10M	18–23	19.8	Huanghuai College (Zhumadian, China)
Taiwan M	10F, 10M	20–22	21.5	National Yang Ming Chiao Tung University (Hsinchu, Taiwan)

on the screen, in a random order. After they had indicated having learned the relevant two-way classification, participants moved on to an identification task in which they heard one of the six tokens in a contrast and were asked to respond by pressing "1" or "2." After each identification trial, they saw either "CORRECT!" or "INCORRECT!" on their screen for 800 ms as feedback. This procedure was repeated four times. The SRT proper was preceded by a warm-up block with six 3-word sequence trials. No feedback of any kind was given in the 4-sequence and 5-sequence experimental blocks. Ignoring the warm-up block, the experimental trials presented participants with all 48 stimulus pairs (6 sequences × 2 sequence lengths × 4 contrasts). Participants confirmed the completion of their response by pressing the ENTER key. The order of presentation of all sequences within all blocks was randomized per participant.

Within each sequence, the non-words were randomly instantiated by one of the three tokens, while no token appeared more than once in a sequence.

Tokens were separated by 120-ms intervals in all sequences. Participants could only register their response after hearing a 1,600-ms recording of four piano chords, played 100 ms after the last token in a sequence. Its function was to reduce the ability of participants to rely on their acoustic memory, similar to that of the recording of "OK!" which has been used for SRTs with stress contrasts. Intervals between trials were 1,500 ms. No response was registered if its sequence length did not match that of the input sequence length.

RESULTS

Two analytical procedures were followed, after Peperkamp et al. (2010), one to answer the question what properties of the pitch contrast, the languages and the participants predict the accuracy scores and another to establish the differences between language groups and any interactions with the contrasts. Thus, we first report two multiple logistic regression analyses of the linguistic variables SALIENCE and HAVECONTRAST, together with the participant variables SEX and APTITUDE. In the first multiple logistic regression analysis, we included the binary variable LEXICALITY, while the gradient variable TONECOMPLEXITY was included in the second. We will next move on to building a mixed-effects model with the experimental design variables, including the phoneme control contrast [la] vs. [ta] (APTITUDE).

The results of the multiple logistic regression analysis on the accuracy scores for the three-pitch contrasts with SALIENCE, HAVECONTRAST, SEX, APTITUDE, and the binary variable LEXICALITY are given in **Table 5**. Significant HAVECONTRAST ($\beta = 0.29$, $p < 0.0001$) shows that participants generally have higher accuracy scores if some pitch difference they are judging is contrastive in their native language ("yes" $M = 0.63$ vs. "no" $M = 0.49$). SALIENCE ($\beta = 0.29$, $p < 0.0001$) indicates that the participants' performance relied to a large extent on how salient a specific contrast is. LEXICALITY ($\beta = 0.3$, $p < 0.0001$) also explained the accuracy results. Participants who speak a (semi-)tonal language ($M = 0.58$) outperformed Indonesian participants, whose native language lacks lexical tone ($M = 0.39$). Lastly, participants' performance on the three-pitch contrasts strongly depended on their scores for the phoneme contrast (APTITUDE, $\beta = 1.12$, $p < 0.0001$). The near-significant effect of SEX ($\beta = -0.07$, $p = 0.079$) weakly indicates that women ($M = 0.55$) performed better than men ($M = 0.52$). The model fit (r^2) is 0.24.

The results of the multiple logistic regression analysis with gradient TONECOMPLEXITY instead of LEXICALITY are given in **Table 6**. With a model fit (r^2) of 0.25, the explained variance is comparable, while the overall results for all identical variables are the same in the two analyses. The range of the accuracy means for TONECOMPLEXITY (0.39 to 0.63) is marginally wider than that for LEXICALITY (0.39 to 0.58) in the first analysis.

Next, two mixed-effects logistic regression analyses were performed on the accuracy scores to establish the effects of contrasts and language groups. The first focused on the tonally intermediate Swedish. With the Swedish participants and the phoneme contrast, [la] vs. [ta], as baselines, the regression model was fitted with CONTRAST * GROUP and SEQUENCELENGTH as variables, where CONTRAST has the three-pitch contrasts and the phoneme contrast as levels. In addition, the model included random intercepts for participant as well as by-participant random slopes for CONTRAST and SEQUENCELENGH. The second analysis was carried out to assess the degree of similarity between the two tonal languages, Taiwan vs. Zhumadian Mandarin. For this analysis, Taiwan Mandarin and the phoneme contrast were set as baselines, with the rest of the model structure remaining the same as that of the first. The analyses were run in R using the *lme4* package (Bates et al., 2015). The results of the two analyses are presented in **Tables 7** and **8**. **Figure 7** gives a box plot with accuracy means and per participant scatter plots.

The results of the first model show that the Swedish participants ($M = 0.88$) performed comparably at the phoneme contrast baseline with the Indonesian ($M = 0.86$) and Zhumadian Mandarin participants ($M = 0.87$), but marginally underperformed compared to the Taiwan Mandarin participants ($M = 0.93$; $\beta = 0.62$, $p = 0.06$). Swedish participants performed less well on the tonal contrasts than on the phoneme contrast (EarlyFall vs. LateFall ($M = 0.25$; $\beta = -3.51$, $p < 0.0001$), LateFall vs. EarlyRise ($M = 0.61$; $\beta = -1.79$, $p < 0.0001$) and the LateRise vs. RiseFall ($M = 0.60$; $\beta = -1.70$, $p < 0.0001$). Importantly, the Group–Contrast interactions indicate that the participants of the two tonal languages, Taiwan and Zhumadian Mandarin, outperformed Swedish participants on the LateFall vs. EarlyRise contrast (TM: $M = 0.90$, $\beta = 1.41$, $p < 0.001$; ZM: $M = 0.73$, $\beta = 0.76$, $p = 0.02$), while Swedish participants, in turn, outperformed non-tonal Indonesian participants on the same contrast ($M = 0.44$, $\beta = -0.92$, $p = 0.05$). Additionally, Zhumadian Mandarin participants performed better at the tonal contrast that is specific to their language, EarlyFall vs. LateFall, than the baseline Swedish participants ($M = 0.36$, $\beta = 0.74$, $p = 0.03$), while the results of the other two groups on this contrast were comparable to those of the Swedish group. Additionally, Taiwan Mandarin participants ($M = 0.86$) performed better on the LateRise vs. RiseFall contrast than the Swedish participants ($M = 0.60$; $\beta = 0.95$, $p = 0.02$). Finally, and unsurprisingly, 4-word sequences ($M = 0.69$) were responded to with higher accuracy than 5-word sequences ($M = 0.56$; $\beta = -0.42$, $p < 0.0001$).

The model with Taiwan Mandarin as the baseline shows that the Taiwan Mandarin group outperformed the Zhumadian Mandarin group on the phoneme baseline contrast ($\beta = -0.80$, $p = 0.01$); the difference with the Swedish group is just shy of significance. The low score for the Indonesian participants is not significantly different from the Taiwan Mandarin group, which is no doubt due to the wider spread of the scores by the Indonesian group compared to the concentration of the Taiwan Mandarin scores around 1 (**Figure 7**). Similar to the Swedish group, the Taiwan Mandarin group performed less well on the EarlyFall vs. LateFall ($M = 0.29$; $\beta = -3.93$, $p < 0.0001$) and the LateRise vs. RiseFall ($M = 0.86$; $\beta = -0.76$, $p = 0.07$) contrasts than on the phoneme contrast. Their performance on the LateFall vs. EarlyRise contrast, however, was as good as that on the phoneme contrast ($M = 0.90$; $\beta = -0.39$, $p = 0.32$). While the Taiwan Mandarin group still outperformed the non-tonal Indonesian and "semi-tonal" Swedish groups on the LateFall vs. EarlyRise and LateRise vs. RiseFall contrasts (Indonesian: $\beta = -2.32$, $p < 0.0001$; Swedish: $\beta = -1.41$, $p < 0.001$), the Zhumadian Mandarin group stood out on the Zhumadian-specific contrast, EarlyFall vs. LateFall ($M = 0.36$; $\beta = 1.16$, $p = 0.002$), the only contrast for which the Taiwan Mandarin group scored below Zhumadian Mandarin (see also **Figure 7**).

TABLE 5 | Results of a multiple logistic regression analysis with TONE COMPLEXITY as the tonality variable.

	$R^2 = 0.24$				
	B	SE	*z*	*p*	Accuracy means
Intercept	–2.709	0.235	–11.515	<0.0001	
HAVECONTRAST	0.288	0.042	6.809	<0.0001	no: 0.49; yes: 0.63
SALIENCE	0.290	0.014	20.533	<0.0001	2.3: 0.27; 8.3: 0.67; 8.7: 0.67
APTITUDE	1.123	0.286	3.927	<0.0001	
LEXICALITY	0.302	0.066	4.61	<0.0001	–1: 0.39; 1: 0.58
SEX	–0.071	0.04	–1.751	0.079	female: 0.55; male: 0.52

TABLE 6 | Results of a multiple logistic regression analysis with TONE COMPLEXITY as the tonality variable.

	$R^2 = 0.25$				
	B	SE	*z*	*p*	Accuracy means
Intercept	–3.17	0.207	–15.289	<0.0001	
HAVECONTRAST	0.181	0.046	3.954	<0.0001	
SALIENCE	0.296	0.014	20.668	<0.0001	
APTITUDE	1.374	0.233	5.905	<0.0001	
TONECOMPLEXITY	0.165	0.025	6.533	<0.0001	0.0: 0.39; 0.5: 0.49; 4.0: 0.63
SEX	–0.07	0.04	–1.74	0.081	

TABLE 7 | Results of mixed-effects logistic regression analysis with Swedish and [la] vs. [ta] as baselines.

	$R^2 = 0.47$			
	B	**SE**	***z***	***p***
Intercept	2.318	0.300	7.716	<0.0001
GroupIndonesian	0.025	0.428	0.059	0.953
GroupZhumadian M.	−0.187	0.274	−0.681	0.496
GroupTaiwan M.	**0.615**	**0.327**	**1.882**	**0.060**
ContrastEarlyFall **vs.** ***LateFall***	**−3.510**	**0.321**	**−10.925**	**<0.0001**
ContrastLateFall **vs.** ***EarlyRise***	**−1.794**	**0.309**	**−5.802**	**<0.0001**
ContrastLateRise **vs.** ***RiseFall***	**−1.703**	**0.354**	**−4.819**	**<0.0001**
Sequence	**−0.421**	**0.044**	**−9.664**	**<0.0001**
GroupIndonesian:ContrastEarlyFall vs. *LateFall*	−0.718	0.473	−1.520	0.129
GroupZhumadian M.:ContrastEarlyFall **vs.** ***LateFall***	**0.738**	**0.337**	**2.193**	**0.028**
GroupTaiwan M.:ContrastEarlyFall vs. *LateFall*	−0.417	0.389	−1.072	0.284
GroupIndonesian:ContrastLateFall **vs.** ***EarlyRise***	**−0.917**	**0.458**	**−2.003**	**0.045**
GroupZhumadian M.:ContrastLateFall **vs.** ***EarlyRise***	**0.761**	**0.337**	**2.258**	**0.024**
GroupTaiwan M.:ContrastLateFall **vs.** ***EarlyRise***	**1.407**	**0.420**	**3.345**	**0.001**
GroupIndonesian:ContrastLateRise vs. *RiseFall*	−0.420	0.510	−0.823	0.410
GroupZhumadian M.:ContrastLateRise vs. *RiseFall*	0.403	0.334	1.206	0.228
GroupTaiwan M.:ContrastLateRise **vs.** ***RiseFall***	**0.948**	**0.405**	**2.342**	**0.019**

Significant results are presented in bold.

DISCUSSION

There are three main results of our experiment on the sequence recall of pitch shapes with Indonesian, Swedish, and Mandarin participants.

1. Accuracy scores were positively influenced by (i) similarities between experimental pitch contrasts and phonological contrasts in the languages, (ii) the phonetic salience of the experimental pitch contrast, and (iii) the participant's aptitude for the experimental task as measured by the score on the phoneme contrast.
2. On one contrast, LateFall vs. EarlyRise, the Swedish group distinguished themselves as intermediate by outperforming the Indonesian group and being outperformed by the two Mandarin groups, with the two Mandarin groups not differing among themselves.
3. On none of the three-pitch contrasts did semi-tonal Swedish participants and the two tonal Mandarin groups outperform the non-tonal Indonesian group without differing among themselves.

We discuss these three findings in this order below.

Dependence of Tone Contrast Sequence Recall Accuracy Scores on Other Factors

Without a doubt, the linguistic effects of our first finding will show up in similar experiments performed with different selections of languages. Given the small size of our experiment, we cannot be confident that the effect sizes will be preserved proportionally in experiments with different sets of pitch contrasts and languages, but our results do show that a tonal SRT will need to address the effects of linguistic properties to a larger extent than a stress-based SRT (*cf.* Best, 2019). Despite the cross-linguistic variation in the distribution of stressed syllables within words outlined in Peperkamp and Dupoux (2002), the cross-linguistic variation in tone systems is larger than that of stress.

The effect of the general ability of participants to perform an SRT, as measured by the accuracy scores of the phoneme contrast (APTITUDE), turned up among four groups of participants with similar age ranges and levels of education. This suggests that for older participants, this task may be more challenging and hence likely to produce lower accuracy scores compared to our participants. Less demanding versions of this experimental task may therefore need to be explored with older participants. As far as we are aware, this is the first time that an SRT aptitude effect has shown up. Rahmani et al. (2015) ignored the phoneme contrast for not being significantly different between language groups. In Peperkamp et al. (2010), the dependent variable was the difference between the accuracy scores for the phoneme contrast and the stress contrast, on the assumption that this effect will exist in absolute terms, while excluding participants showing poor performance from the analysis, resulting in a significant data loss. By including the phoneme contrast scores as a variable in our multiple regression analyses and the model analyses, we were able to retain all participants in the experiment so as to closely model their performance. Various components of aptitude have been addressed in more recent studies as a variable that could potentially modulate tone perception, as in Bowles et al. (2016) and Qin et al., (2021).

The effect of the existence of an experimental pitch contrast in a language's phonology (HAVECONTRAST) is apparent from

TABLE 8 | Results of mixed-effects logistic regression analysis with Taiwanese Mandarin and [la] vs. [ta] as baselines.

	$R^2 = 0.47$			
	B	**SE**	***z***	***p***
Intercept	2.934	0.341	8.604	<0.0001
GroupZhumadian M.	**−0.802**	**0.316**	**−2.542**	**0.011**
GroupSwedish	**−0.615**	**0.327**	**−1.883**	**0.060**
GroepIndonesian	−0.590	0.456	−1.293	0.196
ContrastEarlyFall* vs. *LateFall	**−3.926**	**0.358**	**−10.957**	**<0.0001**
ContrastLateFall vs. *EarlyRise*	−0.387	0.393	−0.987	0.324
ContrastLateRise vs. *RiseFall*	−0.756	0.416	−1.817	0.069
Sequence	**−0.421**	**0.044**	**−9.664**	**<0.0001**
GroupZhumadian M.:ContrastEarlyFall* vs. *LateFall	**1.155**	**0.370**	**3.119**	**0.002**
GroupSwedish:ContrastEarlyFall vs. *LateFall*	0.417	0.389	1.072	0.284
GroupIndonesian:ContrastEarlyFall vs. *LateFall*	−0.302	0.498	−0.606	0.544
GroupZhumadian M.:ContrastLateFall vs. *EarlyRise*	−0.646	0.411	−1.570	0.116
GroupSwedish:ContrastLateFall* vs. *EarlyRise	**−1.407**	**0.420**	**−3.347**	**0.001**
GroupIndonesian:ContrastLateFall* vs. *EarlyRise	**−2.324**	**0.518**	**−4.485**	**< 0.0001**
GroupZhumadian M.:ContrastLateRise vs. *RiseFall*	−0.545	0.395	−1.380	0.168
GroupSwedish:ContrastLateRise* vs. *RiseFall	**−0.948**	**0.404**	**−2.343**	**0.019**
GroupIndonesian:ContrastLateRise* vs. *RiseFall	**−1.368**	**0.557**	**−2.455**	**0.014**

Significant results are presented in bold.

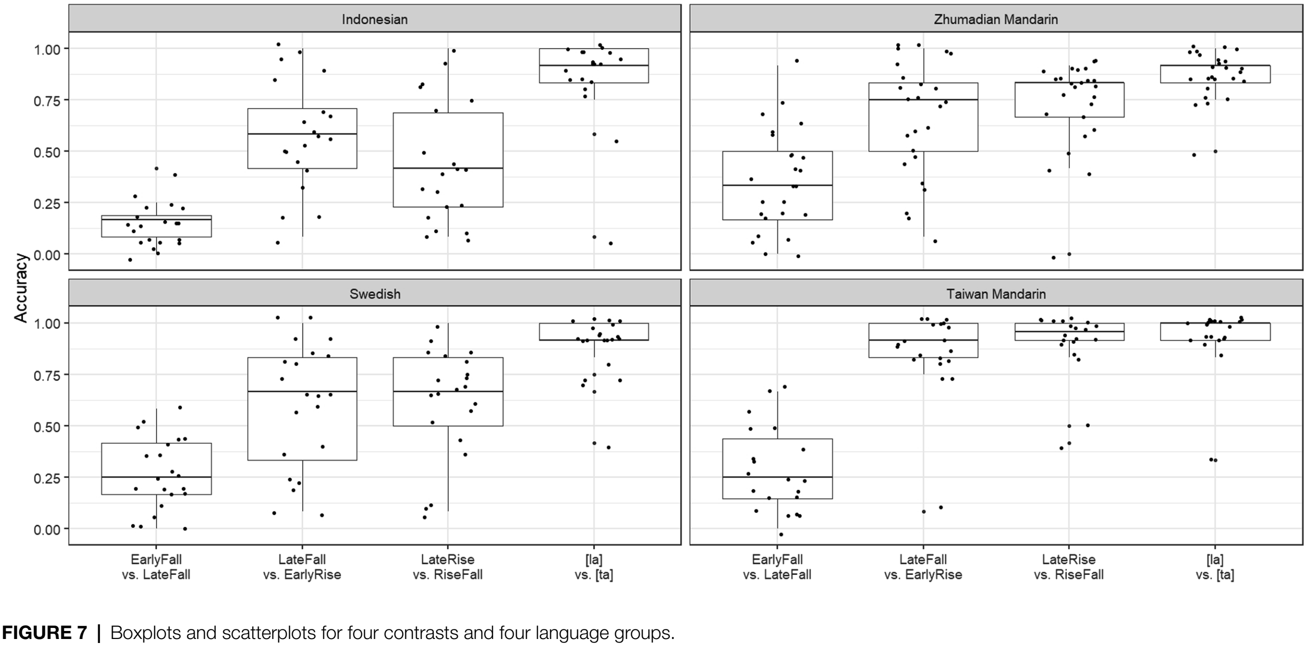


FIGURE 7 | Boxplots and scatterplots for four contrasts and four language groups.

the interactions between the pitch contrasts and the language groups in the mixed-effects models. The Zhumadian group, whose language is the only one to have a temporal alignment contrast for falls, outperformed both the Swedish and Taiwan Mandarin groups on the EarlyFall vs. LateFall contrast, in addition to the low-scoring Indonesian group. The three non-Zhumadian groups did not differ significantly from each other, as shown by the lack of any interaction between Indonesian and the EarlyFall vs. LateFall contrast in either analysis (**Tables 7** and **8**). The effect of contrast salience (SALIENCE) was most clearly in evidence in the overall lower scores for the EarlyFall vs. LateFall contrast compared to the other two pitch contrasts.

Three Typological Groups?

Our second finding was that both Mandarin groups outperformed the Indonesian and Swedish groups on the LateFall vs. EarlyRise contrast, with the Indonesian group scoring below the Swedish group. If we interpret the contrast between rising and falling pitch to be prototypical, the pattern Indonesian < Swedish < Zhumadian and Taiwan Mandarin suggests a three-way distinction between atonal, semi-tonal, and tonal languages. If this result were to be replicated with other mixes of languages, it would imply that a binary diagnostic is unlikely to emerge from a tone-based SRT with a broad typological mix of languages. In turn, this might put experiments with small numbers of languages that have yielded significant results between tonal and non-tonal languages in a different perspective, in the sense that they may represent values on a tone/non-tone continuum rather than as values of a binary variable.

Testing Varieties of the Same Language

Turning the above conclusion around so as to adopt a positive perspective, we might expect tonal and non-tonal varieties of the same language that otherwise have few differences between them to be consistently distinguishable with the help of a tonal SRT. Such languages include Japanese, Korean, Swedish/Norwegian, Franconian varieties of Dutch and German, and Serbian/Croatian (van der Hulst et al., 2011; Gussenhoven and Chen, 2020). Importantly, it is in such cases that the tonal nature of languages has been debated, most notably with respect to two properties, one distributional and the other representational. The first is exemplified by Tokyo Japanese and Northern Bizkayan Basque, which have been characterized as "pitch accent languages," a distinct type by the side of tonal and non-tonal languages. Dominant characterizations of this group indicate the restriction of contrastive tone in a single location of the word or word-like domain. Hyman (2006, 2009) has signaled the absence of a clear definition, in particular that of the demarcation line with tone languages proper. Thus, the single location could be "fixed," like the penultimate syllable of Lekeitio Basque, be restricted to the non-final stressed syllable, as in Swedish, or to one of two syllables at a word edge, as in Kagoshima Japanese and Barasana, or be lexically specified, as in Tokyo Japanese (Elordieta, 1998; Gomez-Imbert and Kenstowicz, 2000; Hualde, 2012; Jun and Kubozono, 2020). Also, there may be two locations for a tone contrast, one at the beginning and one toward the end, as in Osaka and Ibukujima Japanese (Pierrehumbert and Beckman, 1988; Uwano, 1999), while the contrastive tone could be privative, as in the above varieties of Japanese, or represent a contrast between two tone melodies, as in Barasana (*cf.* Hualde, 2012). The other controversy concerns the issue whether surface tone contrasts in varieties of Swedish/Norwegian and Franconian are due to underlying tones (e.g., Bruce, 1977; Riad, 2014; Gussenhoven and Peters, 2019) or to differences in underlying foot structure which generate the different surface tone structures (e.g., Köhnlein, 2011, 2016, 2017; Hermans, 2012; Morén-Duolljá, 2013; Kehrein, 2018). Future explorations of our tone-based SRT might therefore fruitfully compare non-tonal and putatively tonal varieties of the same language.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/uxv4j/?view_only=86c8981b6c9c46c38fe2c2900afd4bcc.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Research Ethics Committee for Human Subject Protection of National Yang Ming Chiao Tung University. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CG: conceptualization, methodology, data curation, supervision, and writing—original draft. Y-AL: data curation, formal analysis, funding acquisition, project administration, resources, visualizations, software, Taiwan Mandarin and Indonesian experiments, and writing—review and editing. S-IL-K: methodology, formal analysis, and writing—review and editing. CL: resources, Zhumadian Mandarin experiment, and writing—review and editing. HR: resources. TR: writing—review and editing. HZ: Swedish experiment and writing—review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

The work by CG was supported by MOST-208-2811-H-009-500 (Ministry of Science and Technology, Taiwan) awarded to Ho-hsien Pan. Taiwan Mandarin and Indonesian participants were run using MOST 109-2410-H-009-048 granted to Y-AL.

REFERENCES

Althaus, N., Wetterlin, A., and Lahiri, A. (2021). Features of low functional load in mono- and bilinguals' lexical access: evidence from Swedish tonal accent. *Phonetica* 78, 175–199. doi: 10.1515/phon-2021-2002

Baddeley, A. (2010). Working memory. *Current Biology* 20, R136–140. doi: 10.1016/j.cub.2009.12.014

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

Best, C. (2019). The diversity of tone languages and the roles of pitch variation in non-tone languages: considerations for tone perception research. *Front. Psychol.* 10:364. doi: 10.3389/fpsyg.2019.00364

Boersma, P., and Weenink, D. (1992-2020). Doing phonetics by computer. Available at: www.praat.org

Bowles, A. R., Chang, C. B., and Karuzis, V. P. (2016). Pitch ability as an aptitude for tone learning. *Lang. Learn.* 66, 774–808. doi: 10.1111/lang.12159

Bruce, G. (1977). *Swedish Word Accent in Sentence Perspective*. Lund: Gleerup.

Correia, S., Butler, J., Vigário, M., and Frota, S. (2015). A stress "deafness" effect in European Portuguese. *Lang. Speech* 58, 48–67. doi: 10.1177/0023830914565193

Deng, D., Shi, F., and Lu, S. (2006). The contrast on tone between Putonghua and Taiwan Mandarin. *Acta Acustica* 31, 536–541.

Domahs, U., Wiese, R., Bornkessel-Schlesewsky, I., and Schlesewsky, M. (2008). The processing of German word stress: evidence for the prosodic hierarchy. *Phonology* 25, 1–36. doi: 10.1017/S0952675708001383

Dupoux, E., Pallier, C., Sebastián, N., and Mehler, J. (1997). A destressing 'deafness' in French? *J. Memory Lang.* 36, 406–421. doi: 10.1006/jmla.1996.2500

Dupoux, E., Peperkamp, S., and Sebastián-Gallés, N. (2001). A robust method to study stress 'deafness'. *J. Acoust. Soc. Am.* 110, 1606–1618. doi: 10.1121/1.1380437

Dupoux, E., Sebastián-Gallés, N., Navarete, E., and Peperkamp, S. (2008). Persistent stress 'deafness': the case of French learners of Spanish. *Cognition* 106, 682–706. doi: 10.1016/j.cognition.2007.04.001

Elordieta, E. (1998). Intonation in a pitch accent variety of Basque. *ASJU: Int. J. Basque Ling. Philology* 32, 511–569.

Fon, J., and Chiang, W.-Y. (1999). What does Chao have to say about tones? A case study of Taiwan Mandarin/赵氏声调系统与声学之联结及量化–以台湾地区国语为例. *J. Chin. Ling.* 27, 13–37.

Fournier, R., and Gussenhoven, C. (2010). Measuring phonetic salience and perceptual distinctiveness: the lexical tone contrast of Venlo Dutch. *Revista Diadorim: Revista de Estudos Linguísticos e Literários do Programa de Pós-Graduação em Letras Vernáculas da Universidade Federal do Rio de Janeiro*, 12. Available at: http://www.revistadiadorim.letras.ufrj.br (Accessed September 24, 2021).

Goedemans, R., and van Zanten, E. (2007). "Stress and accent in Indonesian," in *Prosody in Indonesian Languages*. eds. V. J. van Heuven and E. van Zanten (Utrecht: LOT), 35–62.

Gomez-Imbert, E., and Kenstowicz, M. (2000). Barasana tone and accent. *Int. J. Am. Ling.* 66, 419–463. doi: 10.1086/466437

Gooden, S. (2022). Intonation and prosody in creole languages: an evolving typology. *Annu. Rev. Ling.* 8, 343–364. doi: 10.1146/annurev-linguistics-031120-124320

Gussenhoven, C., and Chen, A. (2020). *The Oxford Handbook of Language Prosody*. Oxford: Oxford University Press.

Gussenhoven, C., and Peters, J. (2019). Franconian tones fare better as tones than as feet: a reply to Köhnlein (2016). *Phonology* 36, 497–530. doi: 10.1017/S095267571900023X

Gussenhoven, C., and van de Ven, M. (2020). Categorical perception of lexical tone contrasts and gradient perception of the statement-question intonation contrast in Zhumadian Mandarin. *Lang. Cogn.* 12, 614–648. doi: 10.1017/langcog.2020.14

Hayes, B. (1995). *Metrical Stress Theory: Principles and Case Studies*. Chicago, IL: Chicago University Press.

Hermans, B. (2012). "The phonological representation of the Limburgian tonal accents," in *Phonological Explorations: Empirical, Theoretical and Diachronic Issues*. eds. B. Botma and R. Noske (Berlin: Mouton de Gruyter), 223–239.

Hualde, J. I. (2012). Two Basque accentual systems and word-prosodic typology. *Lingua* 122, 1335–1351. doi: 10.1016/j.lingua.2012.05.003

Huang, T., and Johnson, K. (2010). Language specificity in speech perception: perception of Mandarin tones by native and nonnative listeners. *Phonetica* 10, 243–267. doi: 10.1159/000327392

Hyman, L. M. (2006). Word prosodic typology. *Phonology* 23, 225–257. doi: 10.1017/S0952675706000893

Hyman, L. M. (2009). How (not) to do phonological typology: the case of pitch accent. *Lang. Sci.* 31, 213–238. doi: 10.1016/j.langsci.2008.12.007

Hyman, L. M. (2011). "Tone: is it different?" in *The Handbook of Phonological Theory. 2nd Edn*. eds. J. A. Goldsmith, J. Riggle and A. Yu (Malden: Wiley Blackwell), 197–239.

Hyman, L. M. (2016). "Lexical vs. Grammatical Tone: Sorting out the Differences." in *Proceedings of the 5th International Symposium on Tonal Aspects of Languages* (Buffalo, NY: TAL 2016), 6-11.

Jun, S.-A., and Kubozono, H. (2020). "Asian Pacific Rim," in *The Oxford Handbook of Language Prosody*. eds. C. Gussenhoven and A. Chen, (Oxford: Oxford University Press), 355–369.

Kehrein, W. (2018). "There's no tone in Cologne: against tone-segment interactions in Franconian," in *Segmental Structure and Tone*. eds. W. Kehrein, B. Köhnlein, P. Boersma and M. van Oostendorp (Berlin: Mouton de Gruyter), 147–194.

Kehrein, W., Köhnlein, B., Boersma, P., and van Oostendorp, M. (2017). *Segmental Structure and Tone*. Berlin: Mouton de Gruyter, 147−194.

Köhnlein, B. (2011). Rule reversal revisited: synchrony and diachrony of tone and prosodic structure in the Franconian dialect of Arzbach. PhD dissertation: Leiden. LOT Dissertation series 274.

Köhnlein, B. (2016). Contrastive foot structure in Franconian tone-accent dialects. *Phonology* 33, 87–123. doi: 10.1017/S095267571600004X

Köhnlein, B. (2017). "Synchronic alternations between monophthongs and diphthongs in Franconian tone accent dialects: a metrical approach," in *Segmental Structure and tone*. eds. W. Kehrein, B. Köhnlein, P. Boersma and M. van Oostendorp (Berlin: Mouton de Gruyter), 211–235.

Kubler, C. C. (1985). The influence of southern min on the mandarin of Taiwan. *Anthropol. Ling.* 27, 156–176.

Lau, J. C. Y., Xie, Z., Chandrasekaran, B., and Wong, P. C. M. (2020). "Cortical and subcortical processing of linguistic pitch patterns," in *The Oxford Handbook of Language Prosody*. eds. C. Gussenhoven and A. Chen (Oxford: Oxford University Press), 499–508.

Lu, S., Vigário, M., Correia, S., Jerónimo, R., and Frota, S. (2018). Revisiting stress "deafness" in European Portuguese: a behavioral and ERP study. *Front. Psychol.* 9:2486. doi: 10.3389/fpsyg.2018.02486

Maskikit-Essed, R., and Gussenhoven, C. (2016). No stress, no pitch accent, no prosodic focus: the case of Ambonese Malay. *Phonology* 33, 353–389. doi: 10.1017/S0952675716000154

Morén-Duolljá, B. (2013). The prosody of Swedish underived nouns: no lexical tones required. *Nordlyd* 40, 196–248. doi: 10.7557/12.2506

Odé, C. (1994). "On the perception of prominence in Indonesian," in *Experimental Studies of Indonesian Prosody*. eds. C. Odé, V. J. van Heuven and E. van Zanten (Leiden University: Rijksuniversiteit te Leiden, Vakgroep Talen en Culturen van Zuidoost-Azië en Oceanië), 27–107.

Peperkamp, S. (2004). Lexical exceptions in stress systems: arguments from early language acquisition and adult speech perception. *Language* 80, 98–126. doi: 10.1353/lan.2004.0035

Peperkamp, S., and Dupoux, E. (2002). "A typological study of stress 'deafness'" in *Laboratory Phonology 7*. eds. C. Gussenhoven and N. Warner (Berlin: Mouton de Gruyter), 203–240.

Peperkamp, S., Vendalin, I., and Dupoux, E. (2010). Perception of predictable stress: a cross-linguistic investigation. *J. Phon.* 38, 422–430. doi: 10.1016/j.wocn.2010.04.001

Pierrehumbert, J. B., and Beckman, M. E. (1988). *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Qin, Z., Chien, Y.-F., and Tremblay, A. (2017). Processing of word-level stress by Mandarin-speaking second language learners of English. *Appl. Psycholinguist.* 38, 541–570. doi: 10.1017/S0142716416000321

Qin, Z., Zhang, C., and Wang, W. S.-Y. (2021). The effect of mandarin listeners' musical and pitch aptitude on perceptual learning of Cantonese level-tones. *J. Acoust. Soc. Am.* 149, 435–446. doi: 10.1121/10.0003330

Rahmani, H., Rietveld, T., and Gussenhoven, C. (2015). Stress "deafness" reveals absence of lexical marking of stress or tone in the adult grammar. *PLoS One* 10:e0143968. doi: 10.1371/journal.pone.0143968

Remijsen, B. (2002). "Lexically contrastive stress accent and lexical tone in Ma'ya," in *Laboratory Phonology. Vol. 7*. eds. C. Gussenhoven and N. Warner (Berlin/New York: Mouton de Gruyter), 585–614.

Rhee, N., Chen, A., and Kuang, J. (2021). Musicality and age interaction in tone development. *Front. Neurosci.* 16:804042. doi: 10.3389/fnins.2022.804042

Riad, T. (2014). *The Phonology of Swedish*. Oxford: Oxford University Press.

Sadakata, M., and McQueen, J. M. (2014). Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Front. Psychol.* 5:1318. doi: 10.3389/fpsyg.2014.01318

Schneider, W., Eschman, A., and Zuccolotto, A. (2012). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools Inc.

Selkirk, E. O. (1980). The role of prosodic categories in English word stress. *Ling. Inq.* 11, 563–605.

Steien, G. B., and Yakpo, K. (2020). Romancing with tone: on the outcomes of prosodic contact. *Language* 96, 1–41. doi: 10.1353/lan.2020.0000

Torgerson, R. C. (2005). A comparison of Beijing and Taiwan Mandarin tone register: An acoustic analysis of three native speech styles. A comparison of Beijing and Taiwan Mandarin tone register: An acoustic analysis of Three native speech styles. PhD dissertation. Provo, UT: Brigham Young University.

Uwano, Z. (1999). "Classification of Japanese accent systems," in *Cross-linguistic Studies of Tonal Phenomena: Tonogenesis, Typology, and Related Topics*. ed. S. Kaji (Tokyo: ILCAA, Tokyo University of Foreign Studies), 151–186.

van der Hulst, H., Goedemans, R., and van Zanten, E. (2011). *A Survey of Word Accentual Patterns in the Languages of the World*. Berlin: Mouton de Gruyter.

van Heuven, V. J., and Turk, A. (2020). "Phonetic correlates of word and sentence stress," in *The Oxford Handbook of Language Prosody*. eds. C. Gussenhoven and A. Chen (Oxford: Oxford University Press), 1501–1565.

Wetterlin, A., Jönsson-Steiner, E., and Lahiri, A. (2007). "Tones and loans in the history of Scandinavian," in *Tones and Tunes. Volume 1: Typological Studies in Word and Sentence Prosody*. eds. T. Riad and C. Gussenhoven, (Berlin: Mouton de Gruyter), 353–375.

Xu, Y., and Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *J. Acoust. Soc. Am.* 111, 1399–1413. doi: 10.1121/1.1445789

Zhao, T. C., and Kuhl, P. K. (2015). Effect of musical experience on learning lexical tone categories. *J. Acoust. Soc. Am.* 137, 1452–1463. doi: 10.1121/1.4913457

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 Hypothesis and Theory
PUBLISHED 28 July 2022
DOI 10.3389/fpsyg.2022.910787

OPEN ACCESS

EDITED BY
Hatice Zora,
Max Planck Institute
for Psycholinguistics, Netherlands

REVIEWED BY
Ulrike Domahs,
University of Marburg, Germany
Tomas Riad,
Stockholm University, Sweden

*CORRESPONDENCE
Mikael Roll
Mikael.Roll@ling.lu.se

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

RECEIVED 01 April 2022
ACCEPTED 28 June 2022
PUBLISHED 28 July 2022

CITATION
Roll M (2022) The predictive function of Swedish word accents.
Front. Psychol. 13:910787.
doi: 10.3389/fpsyg.2022.910787

The predictive function of Swedish word accents

Mikael Roll*

Centre for Languages and Literature, Lund University, Lund, Sweden

Swedish lexical word accents have been repeatedly said to have a low functional load. Even so, the language has kept these tones ever since they emerged probably over a thousand years ago. This article proposes that the primary function of word accents is for listeners to be able to predict upcoming morphological structures and narrow down the lexical competition rather than being lexically distinctive. Psycho- and neurophysiological evidence for the predictive function of word accents is discussed. A novel analysis displays that word accents have a facilitative role in word processing. Specifically, a correlation is revealed between how much incorrect word accents hinder listeners' processing and how much they reduce response times when correct. Finally, a dual-route model of the predictive use of word accents with distinct neural substrates is put forth.

KEYWORDS

phonology, prediction, prosody, morphology, speech processing

Introduction

Swedish words are lexically associated with tonal *word accents* (Elert, 1964). However, the word accent contrast has a questionable phonological function. From a traditional contrastive perspective (Trubetzkoy, 1958), the word accent distinction is often said to have a low *functional load* (Elert, 1972; Riad, 2014; Althaus et al., 2021). Specifically, in Swedish, although word accents are in principle lexically distinctive, in practice, they do not have any relevant role in distinguishing words from each other. The number of minimal pairs is only in the order of a few hundred. Elert (1972) presented a list of 357 minimal pairs, but noted that many were based on archaic word forms, like the 2nd person imperative accent-2 word ^{2}*träden* "step!/thread!" contrasting with accent-1 ^{1}*träden* "the trees." Further, as the previous example illustrates, distinctive pairs often involve different word classes. Their members are hence unlikely to occur in the same syntactic context. Lastly, even the few within-word-class contrasts are questionable as minimal pairs in the traditional sense since their morphological structure differs consistently (Riad, 2014). Typically, the accent-1 words have monosyllabic stems (^{1}*and-en* "the duck"), whereas the accent-2 words have disyllabic stems (^{2}*ande-n*), involving a stem vowel like *-e* (Riad, 2015).

In Norwegian, there is a much higher number of minimal pairs: at least 2,432 (Jensen, 1958) and possibly 3,000 or more, depending on the criteria used (Leira, 1998). It is thus easy to agree with the view that Swedish word accents have a low functional load. However, since lexical word accents are thought to have been in use already in Late Proto Norse, somewhere between the years 600 and 800 (Riad, 1998), an inevitable question arises. Why has the language kept this apparently useless distinction for over a thousand years and shows no signs of losing it? There does not seem to have been any previous stage with a higher functional load of word accents. On the contrary, the larger extension of the contrast in Norwegian is mainly due to a diachronically fairly late change of unstressed vowels into /e/ and a general reduction of unstressed syllables to [ə], making previously different forms become segmental homophones (Elert, 1964, 1981). It is hence not Swedish that has lost contrasts, but Norwegian that has gained them (Riad, 1998).

Swedish word accents

Swedish word accents consist of two distinct word melodies, *accent 1* and *accent 2* (Elert, 1964; **Figure 1**). Accent 1 is often assumed to be the default intonation of a stressed syllable in the absence of a lexical specification (Riad, 2014). In Central Swedish, it is realized as a low tone associated with the stressed syllable of a word (L*). If the word is in a semantically focused context, a rise to a focal high (H) tone is added, giving L*H (**Figure 1**, example 1a). Accent 2 can be assigned lexically or post-lexically. Post-lexical accent 2 is found in all words with secondary stress, involving compounds like ^{2}ˈ*lejon*ˌ*man* "lion's mane" and words with stressed suffixes, such as the derivational suffix -ˌ*het* "-ness" in ^{2}ˈ*när*ˌ*het* "closeness." The secondary stressed syllable, *man* "mane" in example (1b) has a pattern similar to that of the stressed syllable of accent 1 (1a): a L*, which can be followed by a focal H, yielding L*H. Specifically for accent 2, however, the primary stressed syllable—the *lej* of *lejon* "lion" in example (1b)—has a H*L pattern, producing a two-peaked H*L*H pitch contour in focused words. Lexically assigned accent 2 is phonetically similar to post-lexical accent 2, but occurs in words without secondary stress, such as *manar* "manes" (1c). It is also pronounced as a high tone followed by a fall in the stressed syllable (H*L). Since the only stressed syllable is already associated with a non-focal H*, the focal H is realized in the posttonic syllable, producing a two-peaked H*LH sequence in focused words but without secondary stress (Riad, 1998).

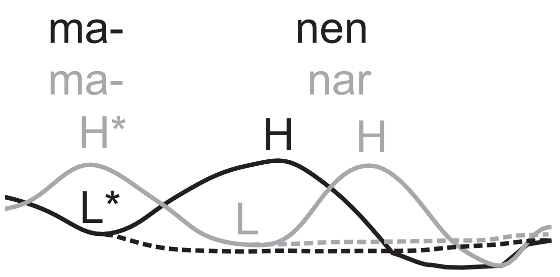


FIGURE 1
Pitch contour of the accent-1 (L*H) word *manen* "the mane" (black lines) and the accent-2 (H*LH) word *manar* "manes" (gray lines). Solid lines represent focused realizations. Dashed lines show unfocused realizations, L* for accent 1 and H*L for accent 2.

(1)

a. *Accent 1*	b. *Post-lexical accent 2*	c. *Lexical accent 2*
L*H	H*L L*H	H*L H
ˈman -en	ˈlejon - ˌman	ˈman-ar
mane -DEF.SG	lion mane	mane-PL

"Lexically" assigned does not mean that the word accent is marked for lexemes. Instead, it is conditioned by the word's morphology (Rischel, 1963). Although word accents are realized on the stressed stem syllable, lexical accent 2 is specified for the stem by a specific set of unstressed suffixes such as *-ar* "-PL (2nd declension)" and *-te/de* "-PST (2nd conjugation)," or stem vowels, as the *-e* of *ande* "spirit." In contrast, accent 1 is a post-lexical realization of a prosodic word not involving any accent 2-inducing morpheme or secondary stress (Riad, 2012). All monosyllabic words have accent 1, but there are also many suffixes that are unmarked for word accent, like *-(e)n* "SG.DEF (2nd declension)" and *-(e)r* "PRS (2nd conjugation)," and therefore occur in words with the default accent 1. In this view, word accents are almost entirely redundant, derivable from stress patterns and suffix information, if not affected by additional phonological processes altering the specified accent (Elert, 1972; Myrberg and Riad, 2015). The accents' redundancy explains their low functional load in the traditional sense; word accents are not used for lexical contrasts but are rather a morphological bi-product. However, whereas post-lexical word accents follow transparent rules, the set of suffixes that triggers accent 2 seems more arbitrary from a synchronic perspective. The next section provides an account for the origin of lexical word accents (Riad, 1998), and sharpens the question of why they have been conserved.

The rise of lexical word accents

Several hypotheses have been advanced about the origin of Scandinavian lexical word accents (Kock, 1878; Öhman, 1967). Riad (1998) particularly well explained the relation between post-lexical and lexical accents and the morphological conditioning of lexical accent 2. Simply put, Riad (1998) derived lexical accent 2 from the still present post-lexical accent 2. His explanation built on the observation that accent 2 without secondary stress is mainly found in words with suffixes that are likely to have been stressed in Early Proto Norse. For example, the Modern Swedish accent-2 word *satte* "put.PRT" contains the past tense suffix -de/-te. All words with this suffix have accent 2, like ^{2}*följ-de* "followed" and ^{2}*köp-te* "bought." Nonetheless, as explained in the previous section, their stems can have accent 1 if combined with an unspecified suffix, such as the present tense conjugation *-er* in ^{1}*följ-er* "follows" or the hypocoristic derivational nominalizer *-is* (Riad, 2012) in the neologism ^{1}*köp-is* "shopping center" of *Valbo Köpis* "Valbo Shopping Center." The reconstruction of the suffix corresponding to past tense *-de/-te* in Early Proto Norse is **-dee* "-3SG.PRT," as in *ˈ*sati-*ˌ*dee* "put-3SG.PRT," with primary stress on *sat-* and secondary stress on *-dee*. In focus, this two-stressed pattern would trigger a post-lexical, two-peaked accent-2 pitch pattern in modern Central Swedish. If post-lexical prominence rules were similar in Proto Norse, words like **satidee* would hence have two pitch peaks. During the syncope period in Late Proto Norse, many intermediate unstressed syllables disappeared, leaving a large number of word forms like *ˈ*sat*ˌ*tee* "put.3PRT" with two adjacent stressed syllables. This led to stress clash resolution removing the secondary stress, giving the Modern Swedish form ˈ*satte* "put.PRT," with only one stressed syllable (Riad, 1992). However, while reducing the length and weight of what had been the secondary stressed syllable, the stress clash resolution left the word melody intact, still with two peaks in a focused position. The pitch contour would then have been reinterpreted as being lexically marked for the specific suffixes rather than the result of applying a post-lexical rule (Riad, 1998). This is where the main question of this article takes shape: Why was the pitch pattern kept when its motivating secondary stress disappeared and why has it been conserved as a lexical accent ever since?

The processing perspective

Elert (1964, 1972, 1981) mentioned two alternative potential functions of word accents besides the lexically contrastive. On the one hand, he argued that one function could be to distinguish different morphemes—chiefly grammatical suffixes—from each other. Thus, whereas the participle suffix *-en* in ^{2}*brut-en* "broken" induces accent 2 onto the stem, the singular definite *-(e)n* in ^{1}*bil-en* "the car" is unmarked for word accent and, therefore, occurs with accent 1. The accent 2-marking for the preceding syllable is what distinguishes the participle suffix from the singular definite. Another role he attributed to accent 2 is the *connective* function. Accent 2 never occurs in monosyllabic words since it is conditioned by secondary stress, suffixes, or stem vowels occurring in a syllable following the primary stressed syllable. This characteristic makes for a potential function of accent 2 in indicating that a word is necessarily polysyllabic (Elert, 1964). However, the morphological and connective functions are both largely redundant from a systemic point of view. The association of word accent with suffix leads only to a handful of contrasts like ^{1}*biten* "the piece" and ^{2}*biten* "bitten," which are included in Elert's list of minimal pairs. In most cases, neither definite nouns have participle segmental homophones nor participle forms have nominal homophones. There is no *^{2}*bilen* or *^{1}*bruten* corresponding to ^{1}*bilen* "the car" and ^{2}*bruten* "broken." Furthermore, since nouns and participles are used in different syntactic environments, word accents are unlikely ever to be needed to distinguish the morphemes.

Post-lexical accents might have a connective function. Specifically, accent 2 can show that two stressed syllables belong to the same syntactic word,[1] and thus make a difference between a phrase like ^{1}ˈ*fin* ^{1}ˈ*hatt* "nice hat" and a compound such as ^{2}ˈ*fin*ˌ*hatt* "fine hat." The phrase and the compound are similar in having two stressed syllables, but, in the phrase, both monosyllables have accent 1, whereas the compound has accent 2 due to its secondary-stress pattern. Nevertheless, the connective function is very weak for lexical accent 2. Polysyllabic words with only one stressed syllable can have either accent 1 or 2. Only in a few cases does the word accent actually distinguish between different forms. It happens under particular syntactic conditions when a suffix and an unstressed verb are homophonous (Elert, 1964). In this vein/'rostar/is understood as a disyllabic verb with the accent 2-inducing suffix *-ar* "-PRS" if pronounced with accent 2 as in example (2)a. If the sequence is uttered with accent 1, it will be interpreted as consisting of two words: the name Ross, followed by the verb *tar* "takes," as indicated in (2)b.

(2)

a. ˈ^{2}rost-ar ˈledningen,
rust-PRS the.wire
"Does the wire rust?"

b. ˈ^{1}Ross tar ˈledningen
Ross takes the.lead

Even the few cases of this type are problematic as arguments for a connective function of lexical accent 2. The verb would not need to be deaccented in example (2b). If it were not, the stress pattern would also have differed between the two sentences. The same is true for the sentence presented by Elert (1964).[2] In other words, lexical accent 2 does not seem to have an essential distinctive function in showing that a stressed and an unstressed syllable together form a word.

As we have seen, even word accents' morphological and connective functions are largely redundant when viewing

[1] Phonologically, it shows that the two syllables belong to the same maximal prosodic word (Myrberg and Riad, 2015).

[2] Elert's (1964) example was *vår* ˈ^{2}*svenska flagga* "our Swedish flag" vs. *vår* ˈ^{1}*sven ska flagga* "our swain shall flag."

language statically as a system. They gain a different sense, however, if a dynamic processing approach is taken. The word accent distinction is perceived in the stressed syllable of a word, often word-initially (sometimes perhaps even in the pre-tonic syllable) as a L* or H* tone. At this point of perception, the suffix, stem vowel, or secondary stress that might have induced accent 2 has not yet been perceived. Thus, at the time point when the word accent distinction becomes audible, it offers non-redundant information about the upcoming structure. At this stage, the tone can have more of a distinctive function. To be exact, most psycholinguistic models assume that the initial speech sounds of an unfolding word (pre-)activate the possible words the listener might be perceiving, the *lexical competitors*. The subsequent sounds reduce the lexical competition by inhibiting competitors that are incompatible with the unfolding sequence of speech sounds, narrowing down the selection to a point where there is only one candidate word left (McClelland and Elman, 1986; Marslen-Wilson, 1987; Norris and McQueen, 2008). In this sense, just like the segments, word accents can help the listener determine which word s/he is listening to. If we hear example (3) with a L* accent 1 tone on *ren-* "reindeer-," we know almost for sure that the noun is definite singular even before hearing the *-en* "-DEF.SG" suffix expressing that information, due to the probabilistic connection between accent 1 and the suffix. If the target word instead involved the accent 2-inducing plural suffix *-ar* "-PL," as in example (4), the stem *ren-* "reindeer" would be pronounced with a H* accent-2 tone. Again, upon hearing the H* tone on the stem, we would strongly expect the associated suffix *-ar* "-PL" to follow.

(3) källk-en drogs av ^{1}ren-en
sledge-DEF.SG was.pulled by reindeer-DEF.SG
"The sledge was pulled by the reindeer"

(4) kälk-en drogs av ^{2}ren-ar
sledge-DEF.SG was.pulled by reindeer-PL
"The sledge was pulled by reindeers"

The predictive function

Recent research has highlighted the predictive nature of speech processing (Kuperberg and Jaeger, 2016; Friston et al., 2021). Rather than processing sounds as they arrive, the brain is thought to constantly entertain weighted hypotheses about what it will perceive next. When the auditory evidence arrives, brain areas of lower-level processing pass on information to higher-level areas about what *does not* conform to the hypotheses. The *prediction error* report is used to fine-tune the predictive model to make predictions even better in the future. Since the major part of our perceptual environment is relatively stable, this *predictive coding* is energetically more cost-effective than treating all information as unexpected (Friston, 2009). It is against this backdrop that I argue that the chief function of word accents and the explanation for their millenary survival is to be found. Word accents are good predictors of how words will continue during processing, and their primary function is predictive. Their role in prediction can be related to their morphological and connective functions. From a processing perspective, word accents can have a quasi-distinctive status as cues to their associated upcoming suffixes. Accent 2 is also a cue to a possible upcoming secondary stress.

There is a relatively large body of evidence that word accents influence prediction. Firstly, if they are combined with the wrong suffix, it takes a longer time to respond to the grammatical meaning conveyed by the suffix (Söderström et al., 2012; Roll et al., 2013, 2015; Roll, 2015; Novén, 2021). For example, if listeners hear *ren-* "reindeer" with accent 1 L* and then the word continues with the accent 2-associated suffix *-ar* "-PL," it takes them longer to decide whether the word is singular or plural than if the correct word accent-suffix combination would have been delivered. Secondly, the surprise at a suffix that is unexpected due to the word accent can also be seen in a brain potential called *P600* (Roll et al., 2013, 2015; Roll, 2015; Novén, 2021). The P600 is an electrically positive brain wave typically peaking at 600 ms following syntactically (Osterhout and Holcomb, 1992) or morphologically (Rodriguez-Fornells et al., 2001) unexpected forms. It has been argued to index reanalysis of the unexpected structure (Morris and Holcomb, 2005).

The fact that word accents *can* be used predictively when relevant to the task (judging suffix-based meaning) does not necessarily entail that they have a predictive role in other contexts. However, even using an acceptability judgment task, Roll et al. (2010) observed a P600 effect for incorrect combinations of word accent and suffix. The experiment additionally involved declensionally incorrect words like **minkor* "minks," where the 1st-declension plural *-or* suffix has replaced the correct 2nd-declension plural *-ar* of *minkar* "minks." Although both suffixes induce accent 2, only *-ar* is of the right declension class. Acceptability was only slightly affected by incorrect combinations of word accent and suffix but was mainly based on the correctness of the declension and the semantic characteristics of the sentences. This implies that the association between word accent and suffix was not perceived as particularly relevant for the task. Likewise, in a study with a task where participants pressed a button at the sentence boundary, suffixes that were invalidly cued by the wrong word accent also produced an increased P600 (Gosselke Berthelsen et al., 2018). In sum, the surprise effect when hearing an incorrectly cued suffix seems relatively task-independent.

Accent 1 is generally a better predictor than accent 2. The reason is that accent 1 reduces the lexical competition more at the point where the stressed syllable is perceived (Söderström et al., 2016). When hearing it, the listener can inhibit the wide range of hypotheses of upcoming possibilities associated with accent 2. Since there are fewer possible continuations, the prediction is more certain when a stem has accent 1. Accent 1 is, to put it in another way, more constraining in processing. As mentioned above, all prosodic words with secondary stress are assigned accent 2 post-lexically. Since compounds have secondary stress, all compounds consequently have accent 2. There are also more inflectional (Riad, 1998) and derivational (Riad, 2012) suffixes that are marked for accent 2 than there are unmarked suffixes. In fact, in a corpus, word-initial syllables with accent 2 had 10.5 times as many possible continuations (10.5 times higher lexical competition) as word-initial syllables with accent 1 (Söderström et al., 2016). The difference in the certainty the two word accents entail can be illustrated by examples (3) and (4). Whereas ^{1}*ren-* with accent 1 has only one possible continuation, ^{2}*ren-* with accent 2 has several, for instance, *spannet* "the team," giving the accent-2 compound *renspannet* "the reindeer team." Hence, even if plural *-ar* is the most likely continuation, the listener cannot be as confident upon hearing the accent-2 stem as when hearing the accent-1 stem. The constraining effect of accent 1 is evidenced by listeners' increased surprise when it is invalidly followed by accent 2-inducing suffixes. The P600 has been found to be larger for invalidly cued accent 2-inducing suffixes, indicating greater morphological reanalysis effects (Roll et al., 2010, 2013). Response times have also been relatively longer for accent-2 suffixes incorrectly preceded by an accent 1 tone on the stem than for unmarked suffixes invalidly cued by accent 2 (Söderström et al., 2012; Roll, 2015).

The higher certainty led to an increase for accent 1 in another brain potential already when participants heard the pitch onset of the word-initial syllable: the *pre-activation negativity* (PrAN) (Roll, 2015; Roll et al., 2015; Söderström et al., 2016, 2017). The PrAN has been seen to be greater the more predictively beneficial a speech sound is (Roll et al., 2017), in both suffix meaning-based tasks and acceptability judgment tasks (Söderström et al., 2016). The PrAN effect of accent 1 was absent in early second language learners, who still had not acquired the predictive use of word accents (Gosselke Berthelsen et al., 2018). However, after intense training, this electrically negative brain potential increased for both word accents, but significantly more for accent 1 (Hed et al., 2019). The results indicate that second-language learners acquired a general predictive use of word accents and learned that accent 1 is a better predictor than accent 2. In short, Swedish-speaking listeners can use word accents predictively during active listening and not only when it is beneficial for a particular task.

Presenting the predictive function in terms of which suffixes word accents pre-activate is overly simplistic. Word accents can often reduce the lexical competition before the listener even knows which stem s/he is perceiving. Already when the initial two segments of a word become apparent, an intense reduction of the available lexical candidates can occur (Marslen-Wilson, 1987; Roll et al., 2017). We cannot directly measure this lexical selection as we cannot access each Swedish speaker's mental lexicon. Still, we can estimate a possible mental lexicon by combining a large speech corpus with a pronunciation lexicon (Söderström et al., 2016). An average speaker can be assumed to have been exposed to words with the approximate frequency and distribution in a corpus with sources representing different language registers. Relating the corpus[3] to the pronunciation dictionary (Andersen, 2011) makes it possible to extract the number of words that begin with a particular sequence of phonemes and their relative frequency.

Taking as an example the last word of (3)–(4), we find that 4,261 nouns begin with /r/. Hearing a following /e/ reduces the lexical competitors to 305 candidates, 7.2% of the initial number. If word-accent information is added, even more substantial inhibition of candidates is achieved. Accent 2 lowers the number to 286, whereas Accent 1 decreases the quantity to 19 possibilities. Hence, when perceiving the second segmental phoneme of the word, accent 1 offers an additional 93.8% reduction of the lexical competitors, whereas accent 2 cuts the number by 6.2%. The example illustrates the general tendency for accent 1 to drop lexical candidates to a much greater extent than accent 2. If we inspect the competitors supported by each word accent, we can see that this quantitative generalization is related to the connective and morphological functions of accent 2, but cannot be reduced to them. The accent-2 group contains words with secondary stress (e.g., *researrangören* "the tour operator" and *renhet* "purity"), words derived by specified suffixes like *-are* [e.g., *redare* "ship-owner(s)"] or stem vowels (e.g., the second *e* in *redet* "the nest"), in addition to the plural *-ar* inflection already mentioned (e.g., *renar* "reindeers"). Although the variation among the accent-1 competitors is much more limited, not only singular suffixes, such as the already mentioned singular definite *-(e)n* of the 2nd-declension word (*renen* "the reindeer") and 5th-declension *-(e)t* of *repet* "the rope," appear, but also the 5th-declension plural inflection *-(e)n* in *repen*, which is also unmarked for accent 2. This illustrates the fact that accent 1 drastically limits the number of morphological possibilities but does not exclude all of them. It should be mentioned that, above, I have disregarded semantic factors that are also liable to play a role in constraining the likelihood of different lexical competitors (Marslen-Wilson, 1987). This section has shown that the pre-activation cued by word accents can be assumed to precede the recognition of the full stressed syllable and to consist to a large extent of suppression of irrelevant alternatives.

[3] https://spraakbanken.gu.se/swe/resurs/parole

The facilitative function

It seems likely that Swedish speakers use word accents to predict upcoming morphemes and word structure and narrow down the lexical competition during listening. Nevertheless, the results reviewed so far do not show that word accents actually facilitate processing. They confirm that incorrect word accents hinder the processing of suffixes, producing a *retardation effect*—slower response times for suffix-based judgments. Incorrect combinations of word accent and suffix also call for reanalysis of the word's morphological structure, as seen in the P600 brain potential. Nonetheless, these effects do not show that the tones make processing faster when they are correct. A *facilitative* role in this sense would be necessary to argue that the predictive function of word accents has been decisive for their survival. This indeed finds support in correlations observed between brain structure measures and skills in the native language. Specifically, a correlation has been detected between the cortical thickness of areas related to phonological and word form processing, involving Wernicke's area, and how much the suffix processing is slowed down by invalid word accents (Schremm et al., 2018; Novén et al., 2021). The participants with thicker cortex in Wernicke's area also showed faster response times for suffix-based meaning in words with correct combinations of word accent and suffix (Schremm et al., 2018). The fact that a thicker cortex in Wernicke's area is related to both quicker processing of valid connections between word accent and suffix and increased impediment in handling invalid combinations suggests that enhanced predictive use of word accents indeed implies better performance in terms of rapid processing of words. Even so, for the purpose of establishing a relation between the processing speed of words and the predictive use of word accents, the link involving the cortex is indirect. A direct relation between the use of word accents and word-processing speed has never been tested.

We can formulate a hypothesis for the facilitative function in the following way: If word accents have a facilitative role, a person who gives them more weight during processing should also process words faster than someone who gives less weight to the word accent information. The facilitative hypothesis can be tested using previously collected response time data from Central Swedish (Roll et al., 2015) and South Swedish (Roll, 2015). I will soon return to the empirical support for the hypothesis but will first present the retardation effect, which is necessary to appreciate the evidence. In Roll et al. (2015) and Roll (2015), participants listened to definite singular or indefinite plural nouns presented in short carrier sentences, for example, *hatt-en* "the hat" or *hatt-ar* "hats" in *Kurt fick hatten/hattar till jul* "Kurt got the hat/hats for Christmas." The same sentences were recorded in the two dialects.[4] Thirty different nouns were presented in singular definite and plural indefinite forms. Half of the stimuli were spliced to create invalid combinations of the word accent realized on the stem and the following suffix.[5] For example, *hatt* "hat" was presented in the valid forms ^{1}*hatt-en* "the hat" and ^{2}*hatt-ar* "hats," and in the invalid forms $*^{2}$*hatt-en* "the hat" and $*^{1}$*hatt-ar* "hats." The task was to judge, as quickly as possible, whether the word was singular (*en* "one") or plural (*flera* "several"). To put it differently, the participants judged the suffix-based part of the meaning of the target words.

If word accents are used predictively, the listeners would be thought to create an expectation for the suffix already when hearing the word stem. If they heard a stem with accent 1, they should predict an upcoming *-en* "-SG.DEF" suffix. If they perceived an accent-2 stem, they would expect a following *-ar* "-PL." The listeners' expectation should lead to increased response times if an unexpected suffix was delivered due to invalid combinations of stem tone and suffix. As mentioned in the previous section, this retardation effect on suffix processing for invalid word accents has been extensively shown. The retardation is the increased response time for suffixes that have been invalidly cued by a stem with the wrong word accent compared to the same suffixes when validly cued by a stem with the correct word accent (Söderström et al., 2012, 2017; Roll et al., 2013, 2015; Roll, 2015). **Figure 2** shows the retardation effect for the joint data for nouns in Roll et al. (2015) and Roll (2015). However, the retardation effect *per se* does not tell whether the word accents have a facilitative function in valid words.

In order to test the facilitative hypothesis, we will now reanalyze the previous response time data to assess whether valid word accents improve processing speed. The reasoning is as follows. If an individual relies more on word accents in his/her processing of suffixes than others, that person should show a greater retardation effect for invalid word accents. Further, if word accents have a facilitative effect, the person depending more on the pitch information would benefit more from hearing valid word accents than others who do not rely as much on the pitch. Therefore, s/he should also be faster than others in

[4] In the Central Swedish stimuli, focus was on the adverbial phrase following the nouns (*till jul* in the example) to avoid having focus on the target word. Focus would create an imbalance between accent 1 and accent 2, since the focal H occurs in the first syllable in accent 1 but in the second syllable in accent 2.

[5] Recordings were carefully carried out to avoid pre- and posttonic differences in F0 between accents 1 and 2. Any potential remaining pre-tonic effects were excluded by the balanced insertion of the target words in carrier sentences originally recorded with accent-1 or accent-2 words. The stressed syllable was always surrounded by voiceless segments to improve the splicing. The resulting stimuli sounded like sentences including a word with correct or incorrect word accent, but otherwise natural pronunciation.

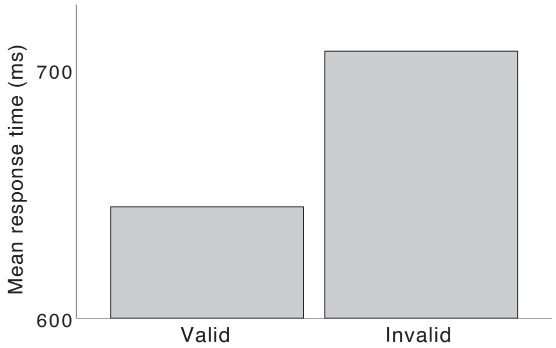


FIGURE 2
Retardation effect. Suffixes in words with invalid combinations of word accent and suffix had longer response times than suffixes in words with valid combinations, $t(79) = 5.93$, $p < 0.001$.

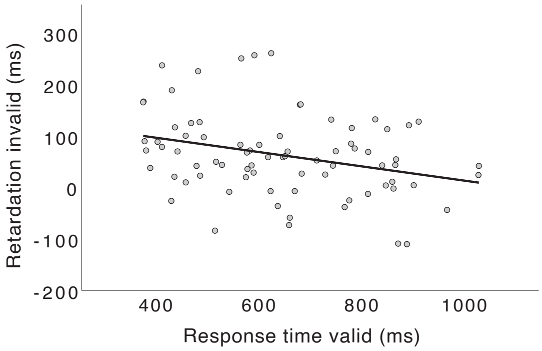


FIGURE 3
Response time for valid words was the best predictor of the retardation effect for invalid words. A regression line with response time in valid words as the only predictor of the retardation effect is shown, $R^2 = 0.093$, $F_{(1,75)} = 7.69$, $p = 0.007$.

processing suffixes of valid words. This will show in a regression model, where an individual participant's response times for valid words should predict the same individual's retardation effect for invalid words. I have tested the hypothesis using linear regression in SPSS (IBM Corp, 2021) on the data in Roll et al. (2015) and Roll (2015) with the retardation effect of invalid word accents as the dependent variable. The retardation effect was calculated as the subtraction of each participant's average response time for a suffix that was validly cued by the correct word accent from the response time for the same suffix when invalidly cued by the incorrect word accent. The response time for validly cued suffixes was entered as an independent variable. Other variables that might explain the retardation effect were also included: word accent and dialect. These variables were dummy coded with values 0 for accent 1 and 1 for accent 2, as well as 0 for Central Swedish, and 1 for South Swedish. Outliers of more than 3 standard deviations over or under the average of each variable were removed. The model was significant, $R^2 = 0.251$, $F_{(3,73)} = 8.15$, $p < 0.001$, explaining 25.1% of the variance in the data. The response time in valid words was the strongest predictor of the retardation effect (standardized $\beta = -0.352$, $p = 0.001$) (**Figure 3**), but word accent (standardized $\beta = -0.292$, $p = 0.005$) and dialect (standardized $\beta = 0.270$, $p = 0.011$), were also significant predictors. To make sure that retardation was specifically related to faster response times for valid words, I also ran the same regression model but included response time for invalid words as an independent variable instead of the response time for valid words. Invalid-word response time did not predict retardation (standardized $\beta = 0.064$, $p = 0.579$).

The regression results show that persons who relied more on word accents during processing also specifically processed valid words faster. In other words, word accents indeed had a facilitative effect. There was also a difference between accents 1 and 2 pointing in the same direction as has previously been found: accent 1 in the stem has a greater retardation effect. As mentioned above, accent 1 is a stronger predictor due to its occurrence in fewer possible words (Roll, 2015; Roll et al., 2015; Söderström et al., 2016). Therefore, it generates stronger activation of its compatible lexical competitors and inhibition of the incompatible candidates, leading to enlarged prediction error and retardation for failed predictions. I will refrain from interpreting the difference between dialects since the speech rate and the focus patterns of the stimuli were not controlled between the two experiments. However, it is worth mentioning that a slightly weaker connective function of accent 2 in South Swedish due to accent 1 also occurring in some compounds (Bruce, 1973; Frid, 2000; Riad, 2015) has been argued to give rise to a somewhat smaller difference in the predictive power of accent 1 and 2 (Roll, 2015).

Dual-route prediction

It now seems clear that word accents have a facilitative function as cues to predict upcoming morphemes and word structure. This implies that Swedish speakers have learned and stored links between tones on stems and specific suffixes. They can further be thought to have associations between an accent-2 tone and an upcoming secondary-stressed syllable. Nevertheless, it is not self-evident how the brain stores the connections between tone and suffix from which the predictions emanate. Based on the dual-route model of morphological processing (Pinker, 1991), there are two chief alternatives for the association between tone and suffix. On the one hand, there can be a more abstract, rule-like connection, something like *H*-ar*, intending to say that all *-ar* "-PL (2nd declension)" suffixes must be preceded by a H* (accent 2) in the stressed syllable. There is also the possibility that words are stored as fully inflected forms, together with their word accent, in representations like $^{\text{H*}}$*bilar* "cars" and ^{H*}*manar* "manes," etc. It is easy to tell that the rule-like option is more parsimonious. Only one association is needed for all words involving the 2nd-declension plural suffix. At the same time, the full-form type storage can be thought to allow for quicker lexical access. When hearing ^{H*}*man*... a listener would immediately activate the full word ^{H*}*manar* "manes" as the most likely option without having to go through a compositional process where, upon hearing the stem, a suffix is selected based on the knowledge about the possible declension and the tone. This option is also more in line with the lexical competition models presented above.

The most apparent evidence of the existence of an abstract association between word accent and suffix comes from a paradigm where pseudoword stems were combined with real suffixes, giving words like *kvup-en* "kvup-SG.DEF" or *kvup-ar* "kvup-PL." As in the experiments reanalyzed in the previous section, the task was to judge the suffix-based meaning, whether the word was in singular or plural form. Sometimes the suffixes were masked by a cough, leaving the word accent as the only cue to the number. Still, it was relatively easy for the participants to perform the task even without hearing the suffixes. The accuracy was as high as 88% for accent 1 and 72% for accent 2 (Söderström et al., 2017), indicating that the participants activated the suffix based only on the word accent since the pseudowords used cannot have had any full-form storage. The lower performance for accent 2 is natural because the words, although less likely, could have been singular compounds, all compounds involving accent 2, or could have had a disyllabic stem with an accent 2-inducing stem vowel like *-e* in *kvup-e*. The word accents were also used predictively. As in real words, invalid combinations of word accents and suffixes led to longer response times and P600 effects for the invalidly cued suffixes. A P600 increase has likewise been observed for invalid combinations of accent 1 with accent 2-inducing suffixes, even if the suffixes were declensionally incorrect, as the $*^{1}$*mink-or* "mink-PL (2nd declension stem-1st declension suffix)" mentioned above (Roll et al., 2010). Furthermore, accent 1 in pseudowords also produced an increased PrAN compared to accent 2 (Söderström et al., 2017). This is because, as mentioned above, the post-lexical accent-2 rule for secondary stressed words applies even in pseudowords, meaning that accent-2 stems yield more possibilities and thus lower certainty. In brief, the word accent-based prediction can proceed combinatorially. Signs of combinatorial processing have also been found for the interaction of stress with suffix in Swedish (Zora et al., 2019). However, it is not evident that this is the preferred route for real nouns (Lehtonen et al., 2009; Schremm et al., 2019).

It has been proposed that word accents of frequent real nouns are stored together with full inflected word forms for quick access (Schremm et al., 2018). Accordingly, as mentioned already, the cortical thickness of Wernicke's area and other temporal brain areas correlated with greater predictive use of word accents in real words (Schremm et al., 2018; Novén et al., 2021). Nonetheless, the same kind of increase in response time for invalid combinations of word accent and suffix in pseudowords did not correlate with cortical thickness in those areas. Instead, there was a correlation with the cortical thickness in Broca's area in the left frontal lobe (Schremm et al., 2018). Broca's area is known for its involvement in combinatorial processing (Ullman et al., 1997). Therefore, Schremm et al. (2018) interpreted the results as showing different neural substrates for the capacity to use word accents predictively in combinatorial and full form-based processing. The brain areas are in line with recent neurolinguistic models situating word processing mainly in the temporal lobe (DeWitt and Rauschecker, 2012) and combinatorial processing in Broca's area (Friederici et al., 2017).

Discussion

The article has asked what the primary function of Swedish word accents is. Lexical word accents have existed in Swedish for probably over a thousand years. Yet, word accents are not really used to distinguish words and hence have a very low functional load in the traditional phonological sense. There is, however, a large body of psycho- and neurophysiological evidence for possible predictive use of word accents. Due to a strong association between the word accents and suffixes, a listener can use the pitch pattern on a stressed word stem to infer properties in the continued speech string. Elert (1964) argued that word accents have a *morphological* function in distinguishing different suffixes. The morphological function can be said to gain relevance when language is viewed from a dynamic processing perspective rather than as a static system. In this sense, the word accent has a quasi-distinctive function at a point in time before the suffix is perceived. At that point, it can be used to predict the suffix. It might be speculated that predicting words' suffixes is crucial in a language where definiteness and number are otherwise mainly expressed in the suffix. Many other languages, involving English, German, and Spanish, express definiteness and number in a pre-nominal article, making the information available before hearing the lexical noun. A preposed definite article is also used in Swedish, but only in complex noun phrases, where no information about the definiteness and number would otherwise be inferable at the phrase onset, being outside the scope of the head noun's word accent pattern. For example, a phrase involving an adjective has an additional initial article doubling the suffix's definiteness and number, as in *den röda boll-en* "SG.DEF red ball-SG.DEF."[6] In these

[6] I thank reviewer 2 for drawing my attention to this fact.

cases, the word accent of the head noun is not perceived at the beginning of the noun phrase, meaning that without the double definiteness marking, information about number and definiteness would only be available upon hearing the noun.

For the first time, this article has shown that the well-known predictive function of word accents, in fact, also involves facilitating word processing. It was revealed that the more listeners relied on word accents in their processing, the faster they processed inflected words with correct word accents. Reliance on word accents was operationalized as the relative increase in response time when judging the meaning of suffixes preceded by the wrong word accent (retardation effect). Finally, the brain can put the predictive function of word accents into practice through two routes with different neural substrates: the combinatorial and holistic routes. In frequent words, word accents seem to be stored and accessed holistically together with fully inflected forms. In essence, upon hearing a stem with a word accent, the listeners activate the linked suffix as part of a word form that is stored with both inflection and word accent as part of the representation. However, the relation between word accent and suffix can also be combinatorially assembled during listening. The combinatorial processing route is probably always activated to some degree, but it is vital in unknown words. More precisely, when hearing an unknown stem with a word accent, the associated suffix is activated through something similar to a grammatical rule, an abstract association between word accent and suffix.

Word accents predict not only suffixes. Since post-lexical accent 2 is used for words with secondary stress, including all compounds, accent-2 stems activate a much larger number of possible continuations. This can be said to be the *connective* function of accent 2 (Elert, 1964) viewed from a speech-processing perspective. Stems with accent 1 can usually only have a limited set of suffixes. Due to the lower number of possible continuations, accent 1 increases the certainty about the continuation of the speech signal and is, therefore, a stronger predictor than accent 2 during listening. In psycholinguistic terms, the pitch-induced certainty is due to a suppression of the lexical competitors that are incompatible with the incoming information. This lexical selection process is likely to gain momentum before the full syllable is recognized, around the point where the first two segmental phonemes become discernable. The higher confidence is indexed by an augmented brain potential, the pre-activation negativity (PrAN), for stems with accent 1. The neural mechanisms underlying the pre-activation of upcoming speech in perception are still being investigated. At present, we do not know to what extent the more prominent neural activity for accent 1 is due to pre-activation of the few alternatives it cues or inhibition of the large number of possibilities associated with accent 2. The most likely scenario is that both processes are involved. Pre-activation can be regarded as a reweighting of hypotheses about the immediate future, strengthening the cued alternatives but inhibiting the uncued. Whether word accents have a low functional load depends on how their function is defined. Here, it is argued that their function is predictive and that they play an essential role in facilitating word processing.

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 Lund Local Ethical Review Board. The patients/participants provided their written informed consent to participate in this study.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

This work was funded by the Swedish Research Council (Grant No. 2018.00632), the Marcus and Amalia Wallenberg Foundation (Grant No. 2018.0021), the Knut and Alice Wallenberg Foundation (Grant No. 2018.0454), Lund University, and the Crafoord Foundation (Grant No. 2017.0006).

Acknowledgments

The author thank the two reviewers, whose comments have helped substantially improve the article. The author also indebted to Anna Hjortdal, Renata Kochančikaitė, and Tuğba Lulaci for discussing a previous version of the manuscript.

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/fpsyg.2022.910787/full#supplementary-material

References

Althaus, N., Wetterlin, A., and Lahiri, A. (2021). Features of low functional load in mono- and bilinguals' lexical access: evidence from Swedish tonal accent. *Phonetica* 78, 175–199. doi: 10.1515/phon-2021-2002

Andersen, G. (2011). *Leksikalsk Database for Svensk*. Oslo: Nasjonalbiblioteket.

Bruce, G. (1973). *Tonal Accent Rules for Compound Stressed Words in the Malmö dialect", Working Papers*. Lund: Phonetics Laboratory. Lund University.

DeWitt, I., and Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proc. Natl. Acad. Sci. U. S. A.* 109, E505–E514. doi: 10.1073/pnas.1113427109

Elert, C.-C. (1964). *Phonologic Studies of Quantity in Swedish Based on Material from Stockholm Speakers*. Stockholm: Almqvist & Wiksell.

Elert, C.-C. (1972). "Tonality in Swedish: rules and a list of minimal pairs," in *Studies for Einar Haugen*, eds K. G. E. S. Firchow, N. Hasselma, and W. O'Neil (The Hague: Mouton).

Elert, C.-C. (1981). *Ljud och Ord i Svenskan 2*. Stockholm: Almqvist & Wiksell International.

Frid, J. (2000). Compound accent patterns in some dialects of Southern Swedish. *Proc. Fonetik* 2000, 61–64.

Friederici, A. D., Chomsky, N., Berwick, R. C., Moro, A., and Bolhuis, J. J. (2017). Language, mind and brain. *Nat. Hum. Behav.* 1, 713–722. doi: 10.1038/s41562-017-0184-4

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Friston, K. J., Sajid, N., Quiroga-Martinez, D. R., Parr, T., Price, C. J., and Holmes, E. (2021). Active listening. *Hear. Res.* 399:107998. doi: 10.1016/j.heares.2020.107998

Gosselke Berthelsen, S., Horne, M., Brännström, K. J., Shtyrov, Y., and Roll, M. (2018). Neural processing of morphosyntactic tonal cues in second-language learners. *J. Neurolinguistics* 45, 60–78. doi: 10.1016/j.jneuroling.2017.09.001

Hed, A., Schremm, A., Horne, M., and Roll, M. (2019). Neural correlates of second language acquisition of tone-grammar associations. *Ment. Lex.* 14, 98–123. doi: 10.1075/ml.17018.hed

IBM Corp (2021). *IBM SPSS Statistics for Macintosh*, 28 Edn. Armonk, NY: IBM Corp.

Jensen, M. K. (1958). *Bokmålets Tonelagspar ("Vippere")*. Bergen: A.S. John Griegs Boktrykkeri.

Kock, A. (1878). *Språkhistoriska Undersökningar om Svensk Akcent*. Lund: Gleerup.

Kuperberg, G. R., and Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* 31, 32–59. doi: 10.1080/23273798.2015.1102299

Lehtonen, M., Vorobyev, V., Soveri, A., Hugdahl, K., Tuokkola, T., and Laine, M. (2009). Language-specific activations in the brain: evidence from inflectional processing in bilinguals. *J. Neurolinguistics* 22, 495–513. doi: 10.1016/j.jneuroling.2009.05.001

Leira, V. (1998). Tonempar i bokmål. *NOR-skrift* 95, 49–86.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition* 25, 71–102. doi: 10.1016/0010-0277(87)90005-9

McClelland, J. L., and Elman, J. L. (1986). The TRACE model of speech perception. *Cogn. Psychol.* 18, 1–86.

Morris, J., and Holcomb, P. J. (2005). Event-related potentials to violations of inflectional verb morphology in English. *Cogn. Brain Res.* 25, 963–981. doi: 10.1016/j.cogbrainres.2005.09.021

Myrberg, S., and Riad, T. (2015). The prosodic hierarchy of Swedish. *Nordic J. Linguist.* 38, 115–147. doi: 10.1080/13682820410001654874

Norris, D., and McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychol. Rev.* 115, 357–395. doi: 10.1037/0033-295X.115.2.357

Novén, M. (2021). *Brain Anatomical Correlates of Perceptual Phonological Proficiency and Language Learning Aptitude*. (Ph.D.thesis). Lund: Lund University.

Novén, M., Schremm, A., Horne, M., and Roll, M. (2021). Cortical thickness of left anterior temporal areas affect processing of phonological cues in native speakers. *Brain Res.* 1750:147150. doi: 10.1016/j.brainres.2020.147150

Öhman, S. (1967). Word and sentence intonation: a quantitative model. *STL-QPSR* 8, 20–54.

Osterhout, L., and Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *J. Mem. Lang.* 31, 785–806. doi: 10.1016/0749-596X(92)90039-Z

Pinker, S. (1991). Rules of language. *Science* 253, 530–535. doi: 10.1126/science.185798

Riad, T. (1992). *Structures in Germanic Prosody*. (Ph.D.thesis). Stockholm: Stockholm University.

Riad, T. (1998). The origin of Scandinavian tone accents. *Diachronica* 15, 63–98. doi: 10.1075/dia.15.1.04ria

Riad, T. (2012). Culminativity, stress and tone accent in Central Swedish. *Lingua* 122, 1352–1379. doi: 10.1016/j.lingua.2012.07.001

Riad, T. (2014). *The Phonology of Swedish*. Oxford: Oxford University Press.

Riad, T. (2015). *Prosodin i Svenskans Morfologi*. Stockholm: Morfem förlag.

Rischel, J. (1963). Morphemic tone and word tone in Eastern Norwegian. *Phonetica* 10, 154–164. doi: 10.1159/000258166

Rodriguez-Fornells, A., Clahsen, H., Lleó, C., Zaake, W., and Münte, T. F. (2001). Event-related brain responses to morphological violations in Catalan. *Cogn. Brain Res.* 11, 47–58. doi: 10.1016/S0926-6410(00)00063-X

Roll, M. (2015). A neurolinguistic study of South Swedish word accents: electrical brain potentials in nouns and verbs. *Nordic J. Linguist.* 38, 149–162. doi: 10.1017/S0332586515000189

Roll, M., Horne, M., and Lindgren, M. (2010). Word accents and morphology—ERPs of Swedish word processing. *Brain Res.* 1330, 114–123. doi: 10.1016/j.brainres.2010.03.020

Roll, M., Söderström, P., Frid, J., Mannfolk, P., and Horne, M. (2017). Forehearing words: Pre-activation of word endings at word onset. *Neurosci. Lett.* 658, 57–61. doi: 10.1016/j.neulet.2017.08.030

Roll, M., Söderström, P., and Horne, M. (2013). Word-stem tones cue suffixes in the brain. *Brain Res.* 1520, 116–120. doi: 10.1016/j.brainres.2013.05.013

Roll, M., Söderström, P., Mannfolk, P., Shtyrov, Y., Johansson, M., van Westen, D., et al. (2015). Word tones cueing morphosyntactic structure: neuroanatomical substrates and activation time course assessed by EEG and fMRI. *Brain Lang.* 150, 14–21. doi: 10.1016/j.bandl.2015.07.009

Schremm, A., Novén, M., Horne, M., and Roll, M. (2019). Brain responses to morphologically complex verbs: an electrophysiological study of Swedish regular and irregular past tense forms. *J. Neurolinguistics* 51, 76–83. doi: 10.1016/j.jneuroling.2019.01.006

Schremm, A., Noveìn, M., Horne, M., Söderström, P., van Westen, D., and Roll, M. (2018). Cortical thickness of planum temporale and pars opercularis in native language tone processing. *Brain Lang.* 176, 42–47. doi: 10.1016/j.bandl.2017.12.001

Söderström, P., Horne, M., Frid, J., and Roll, M. (2016). Pre-activation negativity (PrAN) in brain potentials to unfolding words. *Front. Hum. Neurosci.* 10:512. doi: 10.3389/fnhum.2016.00512

Söderström, P., Horne, M., and Roll, M. (2017). Stem tones pre-activate suffixes in the brain. *J. Psycholing. Res.* 46, 271–280. doi: 10.1007/s10936-016-9434-2

Söderström, P., Roll, M., and Horne, M. (2012). Processing morphologically conditioned word accents. *Ment. Lex.* 7, 77–89. doi: 10.1075/ml.7.1.04soe

Trubetzkoy, N. (1958). *Grundzüge der Phonologie*. Göttingen: Vandenhoeck & Ruprecht.

Ullman, M. T., Corkin, S., Coppola, M., Hickok, G., Growdon, J. H., Koroshetz, W. J., et al. (1997). A neural dissociation within language: evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *J. Cogn. Neurosci.* 9, 266–276. doi: 10.1162/jocn.1997.9.2.266

Zora, H., Riad, T., and Ylinen, S. (2019). Prosodically controlled derivations in the mental lexicon. *J. Neuroling.* 52:100856. doi: 10.1016/j.jneuroling.2019.100856

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 19 August 2022
DOI 10.3389/fpsyg.2022.903879

OPEN ACCESS

EDITED BY
Hatice Zora,
Max Planck Institute for Psycholinguistics,
Netherlands

REVIEWED BY
Ratree Wayland,
University of Florida,
United States
Martin Ho Kwan Ip,
University of Pennsylvania,
United States

*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 Psychology

RECEIVED 24 March 2022
ACCEPTED 19 July 2022
PUBLISHED 19 August 2022

CITATION
Zahner-Ritter K, Zhao T, Einfeldt M and Braun B (2022) How experience with tone in the native language affects the L2 acquisition of pitch accents.
Front. Psychol. 13:903879.
doi: 10.3389/fpsyg.2022.903879

How experience with tone in the native language affects the L2 acquisition of pitch accents

Katharina Zahner-Ritter[1]*, Tianyi Zhao[2], Marieke Einfeldt[2] and Bettina Braun[2]

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

This paper tested the ability of Mandarin learners of German, whose native language has lexical tone, to imitate pitch accent contrasts in German, an intonation language. In intonation languages, pitch accents do not convey lexical information; also, pitch accents are sparser than lexical tones as they only associate with prominent words in the utterance. We compared two kinds of German pitch-accent contrasts: (1) a "non-merger" contrast, which Mandarin listeners perceive as different and (2) a "merger" contrast, which sounds more similar to Mandarin listeners. Speakers of a tone language are generally very sensitive to pitch. Hypothesis 1 (H1) therefore stated that Mandarin learners produce the two kinds of contrasts similarly to native German speakers. However, the documented sensitivity to tonal contrasts, at the expense of processing phrase-level intonational contrasts, may generally hinder target-like production of intonational pitch accents in the L2 (Hypothesis 2, H2). Finally, cross-linguistic influence (CLI) predicts a difference in the realization of these two contrasts as well as improvement with higher proficiency (Hypothesis 3, H3). We used a delayed imitation paradigm, which is well-suited for assessing L2-phonetics and -phonology because it does not necessitate access to intonational meaning. We investigated the imitation of three kinds of accents, which were associated with the sentence-final noun in short *wh*-questions (e.g., *Wer malt denn Mandalas*, lit: "Who draws PRT mandalas?" "Who likes drawing mandalas?"). In Experiment 1, 28 native speakers of Mandarin participated (14 low- and 14 high-proficient). The learners' productions of the two kinds of contrasts were analyzed using General Additive Mixed Models to evaluate differences in pitch accent contrasts over time, in comparison to the productions of native German participants from an earlier study in our lab. Results showed a more pronounced realization of the non-merger contrast compared to German natives and a less distinct realization of the merger contrast, with beneficial effects of proficiency, lending support to H3. Experiment 2 tested low-proficient Italian learners of German (whose L1 is an intonation language) to contextualize the Mandarin data and further investigate CLI. Italian learners realized the non-merger contrast more target-like than Mandarin learners, lending additional support to CLI (H3).

KEYWORDS

second language acquisition, proficiency, tone, intonation, alignment, experience, imitation, general additive mixed models

Introduction

Acquiring a second language (L2) poses many concurrent challenges for the learner: building a lexicon, getting the syntax right, and producing segmental and suprasegmental elements of the language correctly. This paper focuses on the acquisition of prosodic aspects, namely pitch accents in an intonation language. The acquisition of intonation in the L2 is influenced by a large range of factors, such as the native language/variety (for overview, see Mennen, 2015; Trouvain and Braun, 2021), proficiency (Grabe et al., 2003; So and Best, 2010; He et al., 2012; Graham and Post, 2018; Shang and Elvira-García, 2022), musical abilities (Li et al., 2022), language aptitude (Jilka, 2009), etc. Here, we study the roles of *native language* and *proficiency* in the acquisition of L2 intonation. As a test case, we examine how native speakers of a tone language (L1: Mandarin), who are low- or high-proficient learners of an intonation language (L2: German), acquire German pitch accents. Given the prosodic differences between Mandarin and German, this allows us to investigate the crosstalk between tone and intonation in L2 acquisition. As will be shown below, this acquisition setting has hardly been studied. We use an imitation paradigm and test three mutually exclusive hypotheses. The first hypothesis (H1) states that Mandarin speakers produce the two kinds of contrasts similarly to native German speakers – given their increased sensitivity to pitch. The second hypothesis (H2) predicts a reduced ability to produce pitch accent contrasts in an L2 intonation language – given the documented sensitivity to tonal contrasts at the expense of processing phrase-level intonational contrasts. Finally, the third hypothesis (H3) predicts cross-linguistic influence (CLI), which refers to the transfer of native language features into the L2 (e.g., McManus, 2022). In particular, H3 predicts that pitch accent contrasts that are perceived as similar, possibly because they are mapped onto the same tones, are more difficult to imitate than pitch accent contrasts that are perceived as dissimilar.

Generally, L2 learners experience difficulties in the acquisition of a target-like intonation – both in perception and in production (e.g., Mennen, 2004, 2015; Liang and Heuven, 2009; He et al., 2012; Chen, 2014; Graham and Post, 2018; Trouvain and Braun, 2021; Shang and Elvira-García, 2022). In production, deviant intonation contours have been shown to lead to a perceived foreign accent (e.g., Willems, 1982; Anderson-Hsieh et al., 1992; Munro and Derwing, 1995; Magen, 1998; Jilka, 2000; Mennen, 2004; Trofimovich and Baker, 2006; Ulbrich and Mennen, 2016), lower intelligibility (Munro and Derwing, 1995; Holm, 2007), and may even slow down lexical processing (Braun et al., 2011). Although some L2 speakers sound more native-like than others, most L2 speakers still tend to show deviations in intonation patterns – even after having been exposed to their L2 for a long time (e.g., Mennen, 1998, 2004; Atterer and Ladd, 2004; O'Brien and Gut, 2010; Zahner and Yu, 2019; Manzoni-Luxenburger, 2021). Given that foreign accents may lead to reduced intelligibility (Munro and Derwing, 1999; Munro et al., 2006) and to negative attitudes toward the accented speakers (Munro et al., 2006), it is vital to understand the source of these difficulties. Acquiring L2 intonation is a complex endeavor since it involves the acquisition of different components on several linguistic levels. In particular, it requires the acquisition of three main components: (i) the phonological inventory of intonational events (i.e., a set of contrastive units – typically pitch accents and boundary tones), (ii) their phonetic implementation (e.g., tonal alignment), and (iii) their communicative function (semantics/pragmatics), *cf.* Mennen (2015). In the present study, we focus on the first two components.

Particularly in the domain of tone languages, only few studies have examined the acquisition of L2 intonation by L1 speakers of a tone language (He et al., 2012; Liu and Chen, 2016; Yuan et al., 2018; Liu and Reed, 2021; Shang and Elvira-García, 2022). These studies revealed effects of L2 proficiency and cross-linguistic differences, but very few studies have provided direct comparisons between learners whose L1 is a tone language versus learners whose L1 is a non-tone language.[1] Also, prior studies often used tasks that required learners to access the semantic and pragmatic meaning, making it hard to determine genuinely phonetic and phonological factors. It is therefore unclear whether the lexical function of f0 puts learners at an advantage when acquiring L2 intonation, or, conversely, whether L2 intonation acquisition is made even more challenging. The present paper sets out to fill this gap by testing the crosstalk between tone and intonation in the acquisition of pitch accent contrasts by native speakers of a tone language.

In Experiment 1, we elicited L2 imitations of German pitch accent contrasts by speakers of Mandarin Chinese in two proficiency groups – and compared them to the native German productions analyzed in Zahner-Ritter et al. (2022). To gauge the difficulties in L2 acquisition for the two Mandarin proficiency groups and to study the role of L1 tone more directly, we included a control group of low-proficient Italian learners of German, whose L1 is an intonation language (Experiment 2). The paper is structured as follows. In the section "Background," we first provide some background on the phonetics and phonology of pitch in German and Mandarin. Section "Experiment 1" presents the main experiment (Mandarin learners of German) and section "Experiment 2" describes the control experiment (Italian leaners of German). In the "General discussion," we discuss CLI and crosstalk between tone and intonation in our data, as well as the role of proficiency, and end with a "Conclusion."

Background

Prosodic typology differentiates intonation languages and tone languages (Yip, 2002; Hyman, 2006).[2] Broadly speaking,

[1] Cross-linguistic comparisons mostly exist with regard to the acquisition of lexical tone languages (e.g., Gandour, 1983; Chiao et al., 2011; Qin and Mok, 2011; Xu and Mok, 2012, 2014; Braun et al., 2014); for an overview and methodological considerations for native and non-native tone perception see Best (2019).

[2] The typological status of languages such as Japanese or Swedish in which lexical pitch accents occur on certain words but not on others (pitch-accent languages) is not of concern for the present paper. Likewise,

intonation languages use pitch movements to mark words that are prominent on the utterance level and at prosodic boundaries, while in tone languages, pitch movements and/or levels primarily mark lexical or grammatical meaning (Yip, 2002; Gussenhoven, 2004; Ladd, 2008). Since the present study tests the ability of speakers of a tone language to produce pitch accents in an intonation language, we briefly introduce general prosodic properties of intonation languages ("Phonology and phonetics of pitch accent contrasts in German"), with a focus on German rising-falling contours, the test case of this study, and tone languages, with a focus on Mandarin ("Tone and intonation in Mandarin Chinese"). In "Deriving hypotheses on the L2 acquisition of pitch accent contrasts," we briefly survey the state of the art on the acquisition of pitch accents in an L2.

Phonology and phonetics of pitch accent contrasts in German

In intonation languages such as German, English, or Italian, the speech melody comprises pitch accents, which are associated with metrically stressed syllables or prominent words, and boundary tones, which are associated with the edges of intonation phrases (Pierrehumbert, 1980; Beckman and Pierrehumbert, 1986; Ladd, 2008). Each utterance contains at least one intonation phrase and each intonation phrase, in turn, at least one pitch accent (Pierrehumbert, 1980; Nespor and Vogel, 1986). Pitch accents mainly signal post-lexical information, such as the discourse status of referents (given, new, accessible, *cf.* Baumann, 2006) the information structure of an utterance (focus vs. background, D'Imperio, 2001; Ladd, 2008, for overview), and speaker attitudes (Braun et al., 2019; Kutscheid and Braun, 2021; Wochner, Forthcoming); boundary tones mainly signal illocution types (question vs. statement, *cf.* Batliner, 1989; Oppenrieder, 1991; Grice, 1995; Niebuhr et al., 2010; Michalsky, 2017) and discourse organization (Lehiste, 1975; Wichmann et al., 2000).

The present paper focuses on the production of pitch accents. In autosegmental-metrical theory of intonation (Arvaniti and Fletcher, 2020 for overview; Pierrehumbert, 1980; Ladd, 2008), pitch accents are composed of low (L) or high (H) tonal targets (or a combination thereof). These tonal targets are associated with the metrically stressed syllable (e.g., the syllable [man] in <Mandalas> "mandalas"). Differences in the temporal alignment of the tonal targets with regard to the stressed syllable result in different pitch accent types. For instance, in an L + H* accent, the L tone precedes the stressed syllable, and the H tone is realized on the stressed syllable (symbolized by the asterisk), see Figure 1A. In contrast, an L* + H accent has its L tone aligned within the stressed syllable while the H tone is realized on the unstressed syllable following the stressed syllable, see Figure 1C. In the present study, we include a further accent type, termed (LH)*, which acoustically lies between the two and in which both L and H are aligned within the stressed syllable (Kohler, 2005; Zahner-Ritter et al., 2022), see Figure 1B.

A recent imitation study with German participants corroborated this three-way partition (Figures 1A–C) in imitated productions, in particular for speakers from Northern Germany (Zahner-Ritter et al., 2022). Pairwise comparisons of f0 values between these rising-falling contours revealed statistical differences in all cases, with a larger acoustic contrast between (LH)* vs. L* + H (orange contour, Figure 1B vs. blue contour, Figure 1C) compared to (LH)* vs. L + H* (orange contour Figure 1B, vs. gray contour, Figure 1A). The contours further elicit distinct interpretations in native speakers of German and are hence considered phonemic in the German pitch accent system (e.g., Kohler, 1991, 2005; Grice et al., 2005; Kügler and Gollrad, 2015; Lommel and Michalsky, 2017; Braun and Biezma, 2019; Zahner-Ritter et al., 2022). In *wh*-questions, L + H* and L* + H [gray (A) and blue (C) contours in Figure 1] were mostly associated with information-seeking meaning, while (LH)* [orange contour (B) in Figure 1] was interpreted as surprise, negative attitude, aversion, and rhetorical meaning (Zahner-Ritter et al., 2022). In declarative sentences, L + H* has been shown to signal new information, L* + H is associated with established facts, and (LH)* with surprise (Kohler, 1991, 2005; Baumann and Grice, 2006; Wochner, Forthcoming; Zahner-Ritter et al., 2022). Zahner-Ritter et al. (2022) directly compared the meaning attributions of these three accents in *wh*-questions and declarative sentences. While L + H* and L* + H were less distinct in meaning in questions, they were clearly differentiated in declaratives. Crucially, the "intermediate" (LH)* accent was distinct from the two other accent types in both sentence types, mostly being associated with surprise, aversion, or other attitudes. Given these differences in utterance meaning, learners of German eventually need to acquire this contrast in order to successfully communicate in their L2.

Tone and intonation in Mandarin Chinese

In tone languages, such as Mandarin Chinese, tones are used to differentiate lexical meanings. There are four lexical tones in Mandarin: Tone 1 which is high-level, Tone 2 which is high-rising, Tone 3 which is low-rising, Tone 4 which is falling (Chao, 1930, 1956; Lin, 2007, see Figure 2), and a neutral tone, which is prosodically weak and whose shape depends on the preceding tone (Cao, 1992; Yip, 2002;

languages that use intonation to (mostly) mark syntactic phrasing (e.g., French, Japanese, Korean, Bengali, Urdu) are not dealt with here (*cf.* Jun, 2005, for overview).

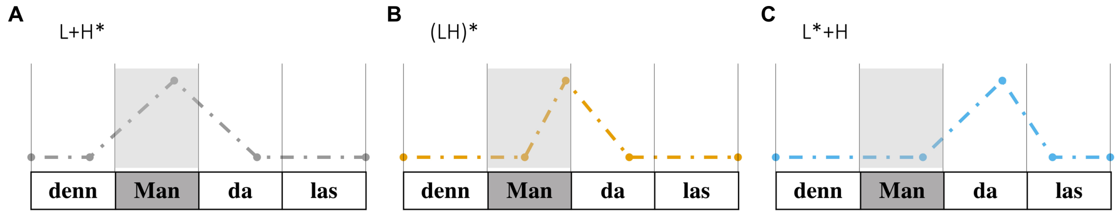


FIGURE 1
Schematic representation of three rising-falling contours in German realized on a four-syllable sequence *denn Mandalas* "PRT mandalas"; gray shading indicates the stressed syllable with which the pitch accent is associated. **(A–C)** show the three different alignment configurations analyzed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

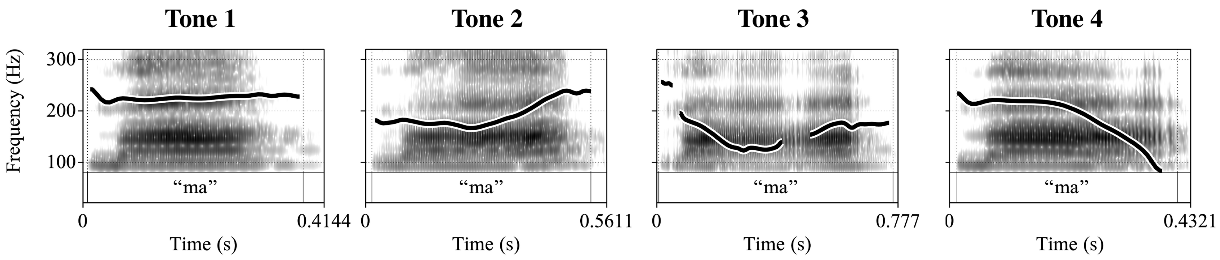


FIGURE 2
Example realization of lexical tones in Mandarin Chinese on the syllable "ma" (Tone 1 to Tone 4, from left to right), produced by a native speaker of Mandarin Chinese.

Zhang et al., 2022).[3] Essentially, each syllable in a phrase carries one of these five tonal specifications, with the syllable boundaries generally serving as anchor points for the alignment of lexical tones (Xu, 1999). Lexical tones essentially determine the shape and height of the f0 contours within the syllable. They also influence the tonal configuration of the adjacent syllables, with tones in the preceding syllables showing stronger influence than the following one. In continuous speech, tones are coarticulated and reach their tonal targets late in the syllable as the f0 contour tends to show less influence and variation caused by the preceding tones toward the end of the syllable (Xu, 1999).

While f0 primarily marks lexical tone in Mandarin, it also conveys post-lexical meaning (Xu, 2019; Zhang et al., 2021, for overviews). The simultaneous existence of the lexical and post-lexical function of f0 (i.e., for tone and intonation) has been referred to as the "multiplexing of the f0 channel" (Zhang et al., 2021, p: 9).[4] For instance, focus is marked by an increase in the f0 range on the focused word and by a compression of the f0 range in the post-focal region (Jin, 1996; Liu and Xu, 2005; Chen and Braun, 2006). Interrogatives are produced with higher overall f0 (Lee, 2005; Liu and Xu, 2005; Yuan, 2006), in particular towards the end of the utterance (Yuan, 2006). The tonal contour at the end of the utterance depends on the tone of the final constituent (Zhang et al., 2021 for overview). For instance, the falling Tone 4 is falling with a smaller range in questions as compared to statements; the rising Tone 2, on the other hand, is realized with an increased f0 range (Zhang et al., 2021) or higher register (Zhang et al., 2022). One approach to model and simulate the effects of lexical tone and intonation on the realization of the f0 contour in Mandarin Chinese is the *Parallel Encoding and Target Approximation model* (PENTA, Xu, 2005). In this model, each tone has an idealized pitch target, which may be static ([high], [low], [mid]), or dynamic ([rise], [fall]). The targets are approximated asymptotically, and,

[3] The neutral tone is typically realised with falling pitch when occurring after Tone 1, 2, and 4, while it is realised with rising pitch after the low-dipping Tone 3 (Yip, 2002, see also Zhang et al., 2022). From a theoretical perspective, the status of lexical tone has been controversially discussed in the linguistic literature and it has not entirely been resolved whether it serves segmental or suprasegmental functions (Yip, 2002; Hyman, 2011; Best, 2019).

[4] Other options to mark information status, information-structure, illocution type, and discourse organisation are word order or particles. In addition, studies have shown that in several tone languages, pitch accents only play a minor role beyond the lexical tone level (Laniran and Clements, 2003; Connell, 2017) and that crosstalk between lexical tone and intonation in tone language may be rather limited (Zerbian, 2016). However, there is also evidence from some other tone languages that intonational effects may be phonetically layered on existing lexical tones, hence revealing crosstalk between the two levels (Xu, 1999; DiCanio et al., 2018).

depending on the communicative function, the f0 range or the strength (speed) of target approximations is adjusted. In any case, this double-function of f0 leads to two processing consequences for Mandarin Chinese listeners: (1) an increased sensitivity to pitch (for tonal contrasts and musical pitch) and (2) a higher sensitivity to lexical tone, which is realized on syllabic units, than for intonation, which spans larger units. We now elaborate on (1) and (2) and formulate hypotheses.

Deriving hypotheses on the L2 acquisition of pitch accent contrasts

The long-term experience with lexical tones leads to a higher sensitivity to musical pitch and improves general pitch processing (Wang et al., 2003; Wong et al., 2007; Zatorre and Gandour, 2008; Pfordresher and Brown, 2009; Bidelman et al., 2011, 2013; Bidelman and Chung, 2015). With regard to musical pitch, Pfordresher and Brown (2009), for instance, showed that L1 speakers of a tone language (Vietnamese, Mandarin Chinese, or Cantonese) outperform L1 speakers of an intonation language (English) on their ability to imitate (*via* singing) four-note sequences of different complexity (level, interval, melodic) and to differentiate between musical notes and musical intervals. The more complex the task, the larger the benefits. Furthermore, behavioral and neurophysiological evidence suggests differences in lexical tone and vowel identification (Gottfried and Suiter, 1997) and in the processing of level tones and contour tones between Chinese and English listeners (e.g., Gandour, 1983; Gandour et al., 2000). Crucially, long-term experience with lexical tones further leads to higher sensitivity toward pitch representations (Gandour et al., 1998, 2000; Krishnan et al., 2005, 2009, 2010; Krishnan and Gandour, 2009) and may enhance neuronal tuning of pitch in the brainstem (Gandour et al., 2000; Krishnan et al., 2009; Krishnan and Gandour, 2009), such that listeners are more accurate in detecting changes in pitch and musical intervals (e.g., interval distances and direction of change, Giuliano et al., 2011). Bidelman and Chung (2015) further showed that Mandarin Chinese listeners showed fine-grained distinctions of pitch encoding between hemispheres and differential processing of pitch contours and intervals, which was different from English listeners. There is also evidence from production supporting the idea that speakers of tone languages may be more sensitive to pitch cues than speakers of intonational languages. For instance, Keating and Kuo (2012) found that Mandarin speakers showed enhanced f0 profiles (higher maxima, larger ranges), especially for one-word utterances, compared to native speakers of American English. This increased sensitivity to pitch in general might therefore generate an advantage for speakers of a tone language when acquiring pitch accent categories in an (intonational) L2.[5] These findings lead to **Hypothesis 1 (H1)** which states a **benefit of general pitch processing**, such that L1 speakers of a tone language are equally good at realizing L2 German pitch accent contrasts (see Figure 1) as native speakers and, crucially, better than learners of an intonation language (Experiment 2).

Meanwhile, several studies have documented that L1 speakers of a tone language are more sensitive to tone, typically restricted to the syllable, than to intonation, typically spanning larger domains (but see Ip and Cutler, 2017, 2020). Yuan (2011), for instance, showed that Mandarin Chinese listeners did not always reach high accuracy in identifying the correct illocution in their L1 (question vs. statement) based on intonational information alone. The lower accuracy occurred with specific lexical tones: Listeners were more accurate in identifying an utterance as a question when it ended in a falling tone (Tone 4, identification rate around 90%) as compared to when it ended in a rising tone (Tone 2, identification rate around 70%). This finding clearly reveals crosstalk between the two domains, see also Liu (2018). Liu et al. (2016a) further tested whether semantic context (neutral vs. providing sufficient information for the (tonal) identity of the final syllable) helped the identification of statement vs. question intonation. Even when the context was informative, identification results were comparable to Yuan (2011), with questions being easier to identify on falling tones. These behavioral findings on the crosstalk between tone and intonation are supported by electrophysiological evidence: Liu et al. (2016b) showed that Mandarin Chinese listeners distinguished between statements and questions based on intonation when the target sentence ended in Tone 4 (as evidenced by a P300 for questions relative to statements), but not when the target question ended in Tone 2 (where no ERP difference between questions and statements was found). This lack of sensitivity to phrase-level intonation also transfers to intonation processing in another tone language (Liang and Heuven, 2009) and to non-native processing (Braun and Johnson, 2011). In particular, Braun and Johnson (2011) used disyllabic nonce-words that had pitch movements resembling Tone 2 and Tone 4 on the first syllable in Experiment 1 or on the second syllable in Experiment 2. Chinese and Dutch listeners performed an ABX match-to-sample task with both sets of contrasts (between-subjects). They showed that Mandarin listeners were more attentive to pitch movements than Dutch listeners as these signaled potential lexical contrasts in Mandarin (but not in Dutch). Dutch listeners, in turn, were more attentive to pitch movements signaling post-lexical information than to pitch movements signaling no meaningful linguistic information. These findings lead to **Hypothesis 2** (H2, crosstalk between tone and intonation) which predicts that L1 speakers of a tone language have generally more difficulties in imitating L2 intonational pitch accent contrasts than learners of an intonation language (who are more used to pitch processing on domains larger than the syllable).

[5] Note that Mandarin and Cantonese listeners also show a great sensitivity to segmental cues, which may even outweigh tonal cues (e.g., Taft and Chen, 1992; Cutler and Chen, 1997).

In the acquisition literature, there are only few studies on the acquisition of pitch accents. Trouvain and Braun (2021) recently summarized that most learner populations align low and high tonal targets differently from native speakers. Later alignment was shown for German learners of English (Atterer and Ladd, 2004; Gut, 2009; Ulbrich, 2013) as well as Japanese and Spanish low-proficient and high-proficient learners of American English (Graham and Post, 2018). Earlier alignment was reported for Dutch learners of Greek (Mennen, 2004) and Basque learners of Spanish, at least in the accents of object phrases (Elordieta, 2003). The fact that some learner groups align tonal targets later and other learner groups earlier suggests an influence of the respective L1. Learners not only deviate from the target in terms of the phonetic realization of pitch accents, but also in terms of the accent type that is used. Ramírez Verdugo (2006) reported that Spanish learners of English produced more rising accents on focused words than native English speakers, who, in turn, produced more falls. Mandarin Chinese learners of Spanish also tended to employ high/rising tunes to substitute Spanish low-pitched accents, along with a general tendency to compress pitch (Shang and Elvira-García, 2022). Most of these studies necessitate access to semantic/pragmatic information, beyond the actual realization of accentual contrasts, which obscures the source of the acquisition difficulties. In the present imitation paradigm, we directly access phonological acquisition. There is only one model on L2 intonation, the L2 Intonation Learning Theory (LILt, Mennen, 2015). In Mennen's model, four aspects are argued to predict successful L2 intonation acquisition, (i) the inventory and distribution of phonological elements, (ii) the phonetic implementation of these elements, (iii) their function and (iv) their frequency of occurrence, hence connecting aspects of form and meaning/usage. Our imitation paradigm allows us to test (i) and (ii). The perceived (dis)similarity between Mandarin tones and the f0 contours on the three target syllables is of relevance for predicting the acquisition success. Native Mandarin Chinese listeners without prior knowledge of German reported that (LH)* and L+H* sounded similar to each other, while L*+H sounded clearly different. Hypothesis 3 (H3) states specific effects of CLI, such that (LH)* and L+H* are more difficult to acquire for L1 tone speakers as they are perceived as similar (**henceforth, "merger contrast"**), while (LH)* and L*+H, which are perceived as dissimilar (**henceforth, "non-merger contrast"**), are easier to acquire.[6] Learners of another intonation language (e.g., Italian), will be exposed to different kinds of CLI and hence produce different intonational patterns.

In the LILt (Mennen, 2015), exposure is a relevant factor to predict successful acquisition of L2 intonation, and indeed empirical studies have shown that higher proficiency is beneficial in pitch accent acquisition (e.g., Baker, 2010; He et al., 2012; Graham and Post, 2018; Shang and Elvira-García, 2022). We therefore predict an effect of proficiency for all three factors described in H1-H3, but potentially in different directions: With respect to CLI (H3) and crosstalk between tone and intonation (H2), proficiency is expected to have a beneficial effect. **CLI** is expected to play a smaller role and the contrasts are produced more target-like (i.e., more similar to the native German realization of the contrast). **Crosstalk** might also be reduced such that learners with more experience of German are able to expand the processing window beyond the syllable, also leading to reduced interference of lexical tone specifics and hence to more target-like productions. In contrast, proficiency might have a reversed effect on the **general, non-linguistic pitch processing skills** in Mandarin learners (H1). Here, high-proficient learners might show a deeper (more linguistic) processing of the contours as compared to low-proficient learners, which might reduce the beneficial effect of general pitch processing advantages. Under this assumption, we predict more distinct contours for low-proficient than for high-proficient learners in production.

Experiment 1

We tested Mandarin Chinese learners of German in two proficiency groups in a delayed imitation paradigm (see Zahner-Ritter et al., 2022, for use with native German speakers). Delayed imitation tasks are particularly suited for tapping into intonational development in phonology because no knowledge of semantics and pragmatics is necessary. In addition, the delay between stimulus and onset of imitation (here of 2.5 s) necessitates some kind of phonological storage, leaving little room for echoic (phonetic) memory (Baddeley, 1986, 2003). When speakers initiate their imitative productions after the delay, the phonetic trace has been decayed and speakers need to recruit phonological processing mechanisms. The paradigm hence directly assesses phonological processing and allows us to shed light on phonological acquisition processes in the L2 acquisition of pitch accent contrasts by L1 tone speakers.

For the analysis, we treat distinct f0 realizations at the group level as evidence for the formation of phonological categories. We processed the f0 contours of the imitations using General Additive Mixed Models (GAMMs, *cf.* Wood, 2006, 2017; Wieling, 2018; van Rij et al., 2019; Sóskuthy, 2021), which allow for a holistic comparison between *intonation condition* and *proficiency* and interactions between these factors over time and in comparison to native German speakers (data from Zahner-Ritter et al., 2022 is used for L1-comparisons). In intonation, there is always some variability, also among native speakers (*cf.* Zahner-Ritter et al., 2022), so we compared the learners' productions to the whole group of native German participants to not disadvantage

[6] Mennen argues that it is crucial to determine whether instances of the L2 category are interpreted as members of the L1 category (Mennen, 2015). Ideally, we wanted to know whether the pitch accents could be transcribed as a certain tone sequence, but this task was too meta-linguistic and caused confusion for native Mandarin speakers (without explicit linguistic knowledge). We instead used the judgements of our informants on the perceived (dis)similarity between the contrasts as a measure to set up the merger vs. non-merger contrast.

TABLE 1 Overview of the Mandarin Chinese participants in Experiment 1. DIALANG scores range from 0 (no knowledge of German) to 75 (excellent knowledge of German). Foreign accent ratings range from 1 (no foreign accent) and 6 (strong foreign accent). For more details on the linguistic background see Supplementary material.

Proficiency	Age in years [mean and (sd)]	DIALANG score [mean and (sd)]	Foreign accent rating [mean and (sd)]
Low-proficiency group	22.4 (2.9)	45.4 (6.3)	4.9 (1.6)
High-proficiency group	23.7 (2.8)	59.9 (3.6)	4.0 (1.4)

learners with less variable input. We focused on the realization of contrasts between pitch accents and tested how these contrasts differ between learners and native speakers. This allows us to reduce differences across participants (e.g., in f0 range) and to focus on the differences across pitch accents.

Methods

Participants

Fifty-five native Mandarin speakers of L2 German participated in an online study *via SoSciSurvey* (Leiner, 2019). Participants filled in a meta-data questionnaire including self-rated proficiency based on the European reference framework ranging from A1 to C2 (Council of Europe, 2020). Proficiency was further measured using the lexical DIALANG test (Alderson, 2005).[7] All participants confirmed to have at least beginner-level knowledge of L2 German (at least A1, otherwise the experiment ended automatically). The mean age of onset in German was 20.6 years (SD = 5.6). Participants participated from various locations in mainland China and Taiwan, with varying proficiency levels across regions. To avoid potential confounds between region and proficiency,[8] we selected 28 participants (see Table 1) based on their proficiency and region of origin. The proficiency grouping was done based on the DIALANG score. Participants with values larger than 53 (i.e., more than 70% of the maximum number of points) were grouped as high-proficient, others as low-proficient. Low-proficient speakers most often indicated their German level as A2, high-proficient speakers as C1. In each proficiency group, 14 participants came from southern regions and 14 from northern regions. However, all of our participants indicated that Mandarin Chinese is their dominant language. Three of the high-proficient participants were living in Germany at the time of testing. None of the participants had any documented speech, hearing, or voice disorders.

After collecting all the data, we randomly selected two utterances from each participant. We asked 12 native speakers of German to rate these utterances for the strength of foreign accentedness on a scale from 1 (no perceivable foreign accent) to 6 (strong foreign accent), see Levi et al. (2007) or Hopp and Schmid (2013). The Mandarin utterances were interspersed with two utterances each from the Italian learners of German (reported in Experiment 2) and 16 utterances from German natives. Agreement among the 12 raters (Cronbach alpha, Cronbach, 1951) was very high (mean $\alpha = 0.97$, 95% CI: [0.96; 0.98]). The last column of Table 1 shows the mean foreign-accent ratings, averaged across the 12 raters and the two recordings of each speaker. The German group had a mean accent rating of 1.0 (SD = 0.0). The difference in DIALANG scores was significant across proficiency groups ($t = 7.73$, $df = 23.0$, $p < 0.0001$), the same was true for mean foreign-accent ratings ($t = -2.26$, $df = 53.9$, $p = 0.03$).

Materials

We employed the stimuli from Zahner-Ritter et al. (2022), used in an imitation study with L1 German speakers. These were 4 *wh*-questions (e.g., *Wer malt denn Mandalas*, lit. "Who draws PRT mandalas?"), in which the final object noun had lexical stress on the first syllable (e.g., [man] in <Mandalas>). The *wh*-questions were recorded by a native speaker of German in two conditions ("source recordings": L + H* and L* + H) and then resynthesized into three intonation conditions, all with a nuclear accent on the object noun: L + H*, (LH)*, and L* + H, with a final low boundary tone, see Figure 1. Differences in duration were removed by manipulating the stimuli such that each syllable had an average duration (within the four items *Mandalas* "mandalas," *Malibu* "Malibu drink," *Melanie* "Melanie," *Libero* "libero soccer position"). The stimuli were further scaled in intensity to 63 dB. All manipulations were done in Praat (Boersma and Weenink, 2016), see Zahner-Ritter et al. (2022) for further details. In total, there were 24 test sentences (4 *wh*-questions × 3 intonation conditions × 2 source recordings).

Procedure

Participants first filled in a questionnaire before they performed the imitation task. After the imitation task they completed the lexical proficiency test (DIALANG). Participants were asked to prepare a computer (desktop or laptop) and headphones at the beginning of the experiment and were given explicit instructions [in German or Mandarin (self-chosen)] for the set-up of recording on their devices (e.g., browser settings). Participants were invited to take part in a lottery for reimbursement. All participants gave informed consent for participation and data processing. The study was conducted remotely *via SoSciSurvey* (Leiner, 2019), ran on an in-house server.

[7] The DIALANG test served as a proxy for assessing the development on all linguistic areas. The DIALANG test scores have been found to correlate well with self-rated proficiency and general proficiency factors such as age of onset, language use, and language preference (Lloyd-Smith et al., 2020).

[8] In the overall data set (N=55), there were more high-proficient participants from the North and vice versa for the South (Zhao, 2022).

For the actual imitation task, participants were randomly assigned to one of two experimental lists (differing in order of items to avoid position effects). The stimuli were played once followed by 2000 ms silence and a 500 ms sine tone (randomly played at 150 Hz or 450 Hz) to reduce the impact of purely phonetic processing (Plomp, 1964; Baddeley and Hitch, 1974). The experiment started with four practice trials to familiarize participants with the voice and the task. Participants were instructed to imitate the target contours as closely as possible. They were additionally advised to complete the study in one go in a quiet environment to minimize background noise and interference during the recording. The recording process began automatically after the second sine tone and ended when participants clicked a key to move on to the next page. Participants were allowed to repeat themselves in case of mistakes or when they were dissatisfied with the recording. In that case, we analyzed the final production. In terms of variables, *intonation contour* was manipulated within-subjects and within-items, so that each participant imitated 24 *wh*-questions overall (4 *wh*-questions × 3 intonation conditions × 2 source recordings).[9]

Data treatment

The sound files were annotated semi-automatically: The initial segmentation generated by *Web-MAUS* (Kisler et al., 2017) was corrected manually where necessary according to standard segmentation criteria, *cf.* Turk et al. (2006), see Figure 3 for analysis tiers and exemplar realizations in the three conditions. The annotation and analysis focused on the final segment [n] of the particle *denn* and the three syllables in the sentence-final noun (e.g., *Mandalas*), as the study concentrated on the production of nuclear pitch accents (see Tier 2 in Figure 3).

F0 values were extracted using ProsodyPro (Xu, 2013) with 50 measurements per syllable. This was done separately for male ($N = 1$) and female speakers ($N = 27$), with different extraction settings for f0-minima and maxima (male: 50–300 Hz; female: 100–500 Hz). The raw f0 values were down-sampled to 10 values per interval for subsequent statistical analyses and converted to semitones (reference level was set to 100 Hz for male and 175 Hz for female speakers). Figure 4 shows the average f0 contours of the low- and high-proficient Mandarin learners of German along with the German native speakers (Zahner-Ritter et al., 2022).

We merged the German dataset reported in Zahner-Ritter et al. (2022), see right panel in Figure 4, with the Mandarin Chinese dataset (middle and left panel in Figure 4) and coded three language-groups: *low-proficient Mandarin Chinese learners*, *high-proficient Mandarin Chinese learners*, and *German native speakers*. We then used General Additive Mixed Models (GAMMs, Wood, 2006, 2017) to test whether the three groups differ in the realization of the three intonation conditions over time. GAMMs allow for a direct comparison between f0 contours because they can model non-linear dependencies of a response variable (here f0 in semitones) and different predictors (here *intonation condition* and *group* and their interaction) over time *via* smooth functions. They do so by using a pre-specified number of base functions of different shapes (Baayen et al., 2018; Wieling, 2018; van Rij et al., 2019; Sóskuthy, 2021). Such direct comparisons between f0 contours allow us to study the realization of accentual contrasts in different speaker groups. Of particular importance are differences in f0 over time in the realization for two kinds of pitch accent contrasts:

- **Non-merger contrast**: (LH)* vs. L* + H (orange vs. blue contour in Figure 1)
- **Merger contrast**: (LH)* vs. L + H* (orange vs. gray contour in Figure 1)

The dependent variable was the f0 value [in semitones (st)]. 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). This allowed us to test whether the interaction significantly improved the fit of the model, compared to a model without an interaction term. Since autocorrelation between values of a variable is problematic and since f0 values at subsequent timepoints are necessarily correlated, we corrected for this by using an autocorrelation parameter *rho*, determined by the acf_resid() function in the package *itsadug* (van Rij et al., 2017). We modeled separate smooths for subjects and items to account for the experimental structure. Model fits were finally 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) to account for tailed residuals. 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). Given that the interpretation of significant differences is only possible through visualization, we present the visualized model output. The steps of the analyses are available on Mendeley http://doi.org/10.17632/w293n86sjr.2.

Results and discussion

The model with the smooth term for the interaction between *condition* and *group* over time was significantly better than the model without this interaction [$\chi^2(18.00) = 271.240$, $p < 2e{-}16$], suggesting that the groups differ in the realization of the f0 contours. The final model (with the scat-linking function), corrected for autocorrelation, accounted for 68.6% of the variance.

Non-merger-contrast: (LH)* vs. L*+H

We start with the distinction of the **non-merger contrast [(LH)* vs. L* + H]**, see Figure 5 for an overview of results. Figure 5 (Panel A) shows the realization of the non-merger contrast in the different groups. Differences between f0 contours

[9] The source recording did not have an effect in the German data (Zahner-Ritter et al., 2022) and is hence not considered here either.

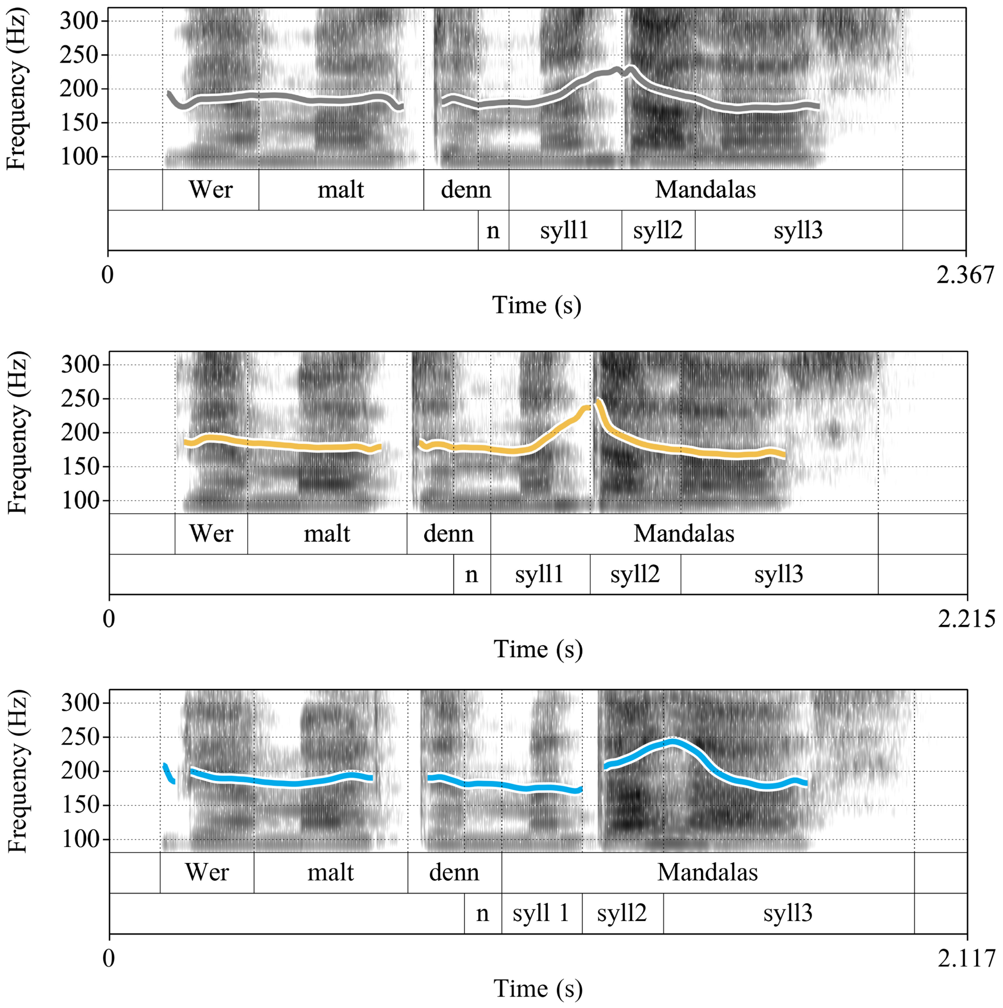


FIGURE 3
Imitative productions of the target question *Wer malt denn Mandalas?* ("Who draws mandalas?") in the three intonation conditions (vp07, Mandarin Chinese low-proficiency group, female, 29 years). Top panel: L+H*, mid panel: (LH)*, bottom panel: L*+H. The filled intervals from tier 2 served as input tier for the extraction of f0 values.

can be assessed in GAMMs with so-called difference curves where one contour is subtracted from the other. In Figure 5 (Panel B), the f0 values of L* + H (blue contour) are subtracted from the f0 values in (LH)* (orange contour). This procedure reveals when in time two f0 contours significantly differ from each other (in case zero is not included in the gray 95% Confidence Interval (CI), indicated by red vertical lines). In terms of L2 acquisition, we interpret distinct f0 contours as evidence for successful category formation. Figure 5 first shows the difference curve for the German L1 data from Zahner-Ritter et al. (2022) in Panel B; Panel C presents the difference curves for the Mandarin Chinese learners of German (low-proficient speakers on the left and high-proficient speakers on the right). Panel D finally presents the difference of the two difference curves shown in B (German) and C (learner groups), hence representing the interaction between *intonation condition* and *group*.

Plotted in terms of such a difference curve (Panel B), the (LH)* contour in **German native speakers** has higher f0 values than the L* + H contour in the stressed syllable (positive difference), and, conversely, the (LH)* contour is lower than the L* + H contour in the post-stressed syllables (negative difference). These differences augment to an absolute value of around 1 st in the stressed and to 2 st in the post-stressed syllable. Also, the L1 German speakers differentiate between (LH)* and L* + H mostly in terms of f0 peak alignment (H tone). The f0 peak occurs late in the stressed syllable of the noun for (LH)* and in the post-stressed syllable for L* + H. Both **learner groups** [Panel C, Mandarin low-proficient learners (left) and Mandarin high-proficient learners (right)] show the same general pattern, but, crucially, tend to make the difference between the two accents acoustically more extreme as compared to the German native speakers (as shown by a larger excursion of the difference curves on the y-axis, compared to the German speakers in Panel B). The contours

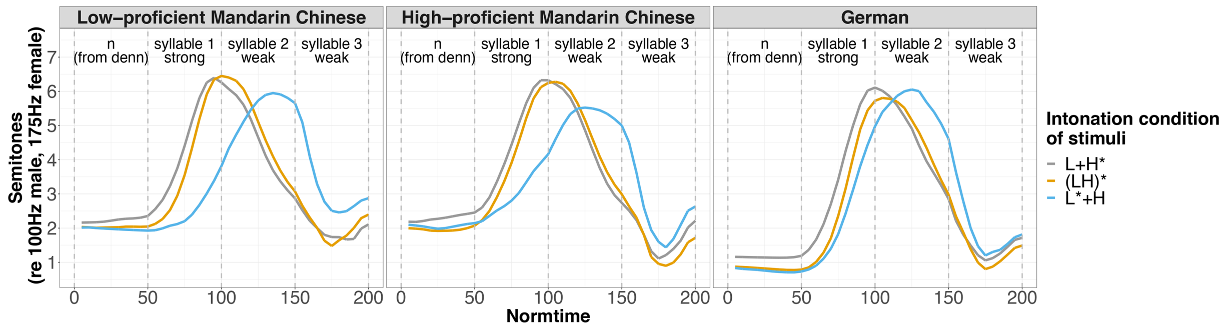


FIGURE 4
Average f0 contours (in st) in the three intonation conditions, split by proficiency group. Left panel shows low-proficient learners of German, middle panel high-proficient learners of German, and right panel German native speakers from Zahner-Ritter et al. (2022).

(LH)* and L* + H are hence further apart in learners than in native speakers. Also, contours in Mandarin Chinese learners start to diverge slightly earlier in the stressed syllable (around Normtime 60) as compared to German native speakers (around Normtime 70), especially for the low-proficiency group. Panel D shows the difference of these differences to pin down group comparisons (German native speakers vs. the two learner groups). To this end, we subtracted the f0 values in the Mandarin groups (low-proficient on the left, high-proficient on the right) from the German group. Since the Mandarin groups show larger f0 differences in the stressed syllable (Normtime 50–100) than the German group, the difference of the difference in Panel D is negative. In the post-stressed syllable (Normtime 100–150), the larger difference of the Mandarin groups reverses. The data hence reveal that learners realize the merger contrast differently from the German native speaker group (more extreme). However, there is also a clear effect of proficiency: The high-proficient Mandarin learners are closer to the German speakers (closer to 0, whereby 0 indicates no deviation from the target) than the low-proficient Mandarin learners, a difference which is significant (see Panel E, which directly compares the contrast in the two learner groups).

Taken together, both high- and low-proficient Mandarin learners produced the pitch accent contrasts in an acoustically more pronounced way than German native speakers. These findings suggest that the perceived difference between the accents (L* + H was judged as different from the other two accents) is clearly measurable in a production experiment, which does not demand conscious judgment. Furthermore, the data show that the effect of perceived (dis)similarity has less effect on high-proficient learners, with high-proficient learners being on average closer to the target than the low-proficient learners. With regard to the realization of the accent types, Mandarin learners realized the rise considerably later in the L* + H accent (compared to the German natives). It is possible that the "late" peak (which was aligned in the post-stressed syllable) was parsed as a tone on the post-stressed syllable, which led to realizations that differed from those of native German speakers.

Merger-contrast: (LH)* vs. L+H*

Figure 6 (Panel A) shows that the merger contrast is acoustically less pronounced than the non-merger contrast across the board (both in native speakers and the two learner groups). In analogy to Figure 5, the f0 values of L + H* (gray contour) are subtracted from the f0 values in (LH)* to arrive at the difference curves (Panels B and C). The difference curves in Panel B show that L1 German speakers differentiate between (LH)* and L + H* such that (LH)* has lower f0 values than L + H*, leading to a negative f0 difference. In the last two thirds of the post-stressed syllable, the (LH)* contour has slightly higher f0 values than L + H*, leading to a positive shift in the difference curve. The two Mandarin Chinese proficiency groups show largely the same pattern as the German native speakers (Panel C), leading to very minor differences of the difference for both speaker groups (Panel D). If anything, the low-proficiency group approached the German native speakers' realization of the contrast more closely than the high-proficient group, evidenced by smaller deviations from 0 in Panel D (left). The low-proficient group, however, showed the differences in the stressed syllable only (i.e., in a smaller time interval than the German native speakers). The accentual differences of the high-proficient learners, in turn, were distributed in the same time intervals as the German native speakers' contrast, but the contrast was smaller for high-proficient leaners than for the German native speakers. The differences between the proficiency groups were numeric only; the interaction between group and proficiency was not significant and is therefore not shown in Figure 6.

Taken together, for the comparison between (LH)* and L + H* (merger contrast), in which the f0 peak (H) was realized in the stressed syllable in both accents, German speakers realized the f0 difference mostly on the stressed syllable (Normtime 50–100), with a slight difference already on the pre-stressed syllable. There were differences in the post-stressed syllable, but these were small. The two learner groups showed a similar pattern, but the difference between the two contours was smaller than in native speakers. For the merger contrast, there was no effect of proficiency.

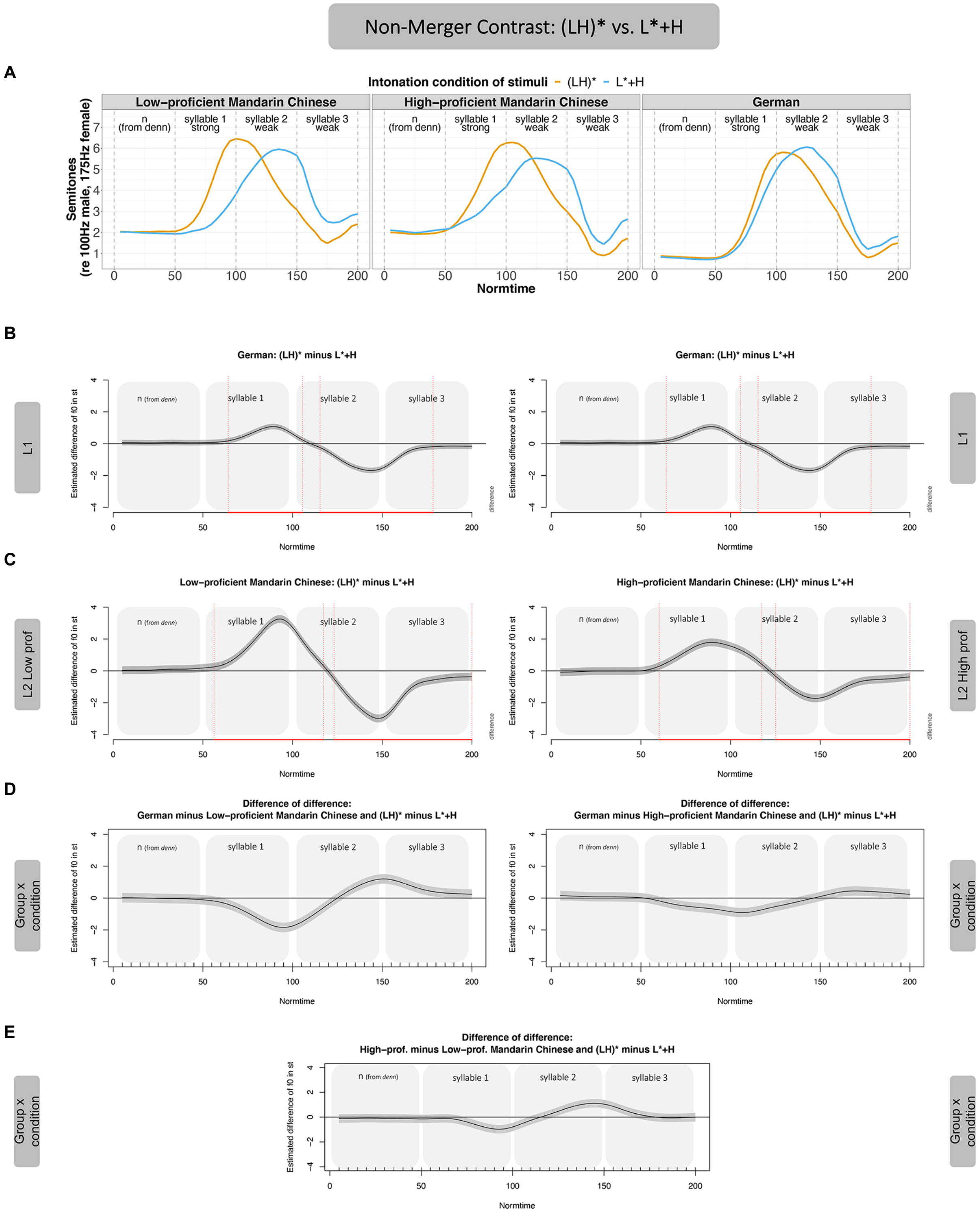


FIGURE 5
GAMM results—Non-Merger Contrast, (LH)* vs. L*+H, by Mandarin learners. Panel **A** shows the contours of the non-merger contrast across participant groups. Panel **B** shows the realization of the contrast in form of difference curves [(LH)* minus L*+H] over time for L1 German (duplicated to make later comparison with the two proficiency groups more transparent, i.e., same figure on left and right). Panel **C** shows the difference curves for the two proficiency groups, Mandarin low-proficient (left) and high-proficient learners (right). Panel **D** shows the difference of the difference between L1 and L2 in the non-merger contrast (i.e., the interaction between condition x group), Panel **E** the difference of the difference between the two proficiency groups.

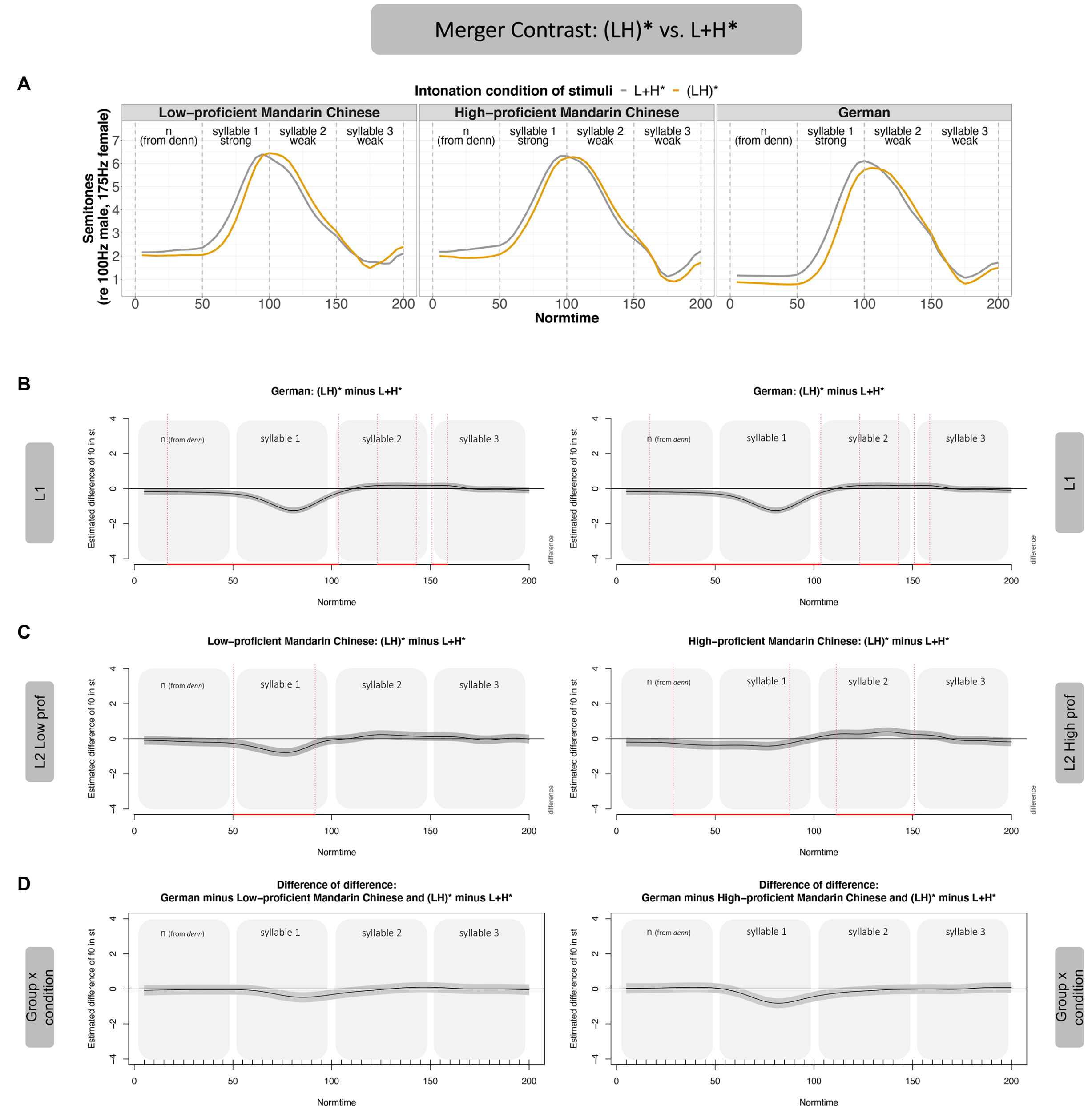


FIGURE 6

GAMM results—Merger Contrast, (LH)* vs. L+H*, by Mandarin learners. Panel **A** shows the contours of the merger contrast across participant groups. Panel **B** shows the realization of the contrast in form of difference curves [(LH)* minus L+H*] over time for L1 German (duplicated to make later comparison with the two proficiency groups more transparent, i.e., same figure on left and right). Panel **C** shows the difference curves for the two proficiency groups, Mandarin low-proficient (left) and high-proficient learners (right). Panel **D** shows the difference of the difference between L1 and L2 in the merger contrast (i.e., the interaction between intonation condition and group).

Interim discussion

Summarizing the results of Experiment 1, the **non-merger contrast [i.e., (LH)* vs. L* + H]** was produced more distinctly by the Mandarin Chinese learners compared to the German native speakers. Auditory impressions by native German speakers even suggested that some of the L* + H realizations in the Mandarin group led to a stress shift, such that the second, unstressed syllable of the noun sounded stressed (instead of the intended first syllable). This perception is most likely driven by the shallower slope of the rise in the stressed syllable, but this needs further investigation. In any case, the larger acoustic contrast in the non-merger contrast by the Mandarin learners (in particular the low-proficient ones) is not target-like. Proficiency seemed to boost the acquisition of this contrast. That is, the high-proficient learners were closer to the native speakers, suggesting that increased experience with an intonation language helps to reduce transfer from the L1. In the **merger contrast [i.e., (LH)* vs. L + H*]**, both learner groups were significantly less distinct than the German native speakers, with no statistical effect of proficiency.[10]

The present data thus reveal an **asymmetrical pattern of pitch accent-contrast acquisition**. Mandarin learners of German are more distinct than German L1 speakers in the non-merger contrast and less distinct in the merger contrast. Such different acquisition outcomes for the two kinds of contrasts had indeed been hypothesized by H3, which based its predictions on CLI. We will return to the discussion of CLI in more detail in the "General Discussion." The imitation data are not in line with H1 (general pitch processing benefit) or H2 (crosstalk), which predicted either **general** benefits or disadvantages for speakers of a tone language in pitch accent processing, i.e., a similar behavior for both kinds of contrasts. With respect to proficiency, our data partly support what has been predicted, since higher proficiency led to more target-like realizations, at least in the non-merger contrast. In the merger contrast, however, the influence of the L1 seems to override effects of proficiency, such that both learner groups produce the contrast in the same way. In Experiment 2, we test whether this pattern of CLI is specific to Mandarin learners or may also be observed in learners whose L1 is an intonation language. In a strong interpretation of H3, L1 speakers of an intonation language (Italian) will produce the contrasts differently than the L1 speakers of a tone language in Experiment 1. These data from learners of a non-tonal language will help us to interpret the type of CLI observed in Experiment 1 better.

Experiment 2

In this control study, we tested a group of low-proficient Italian learners of German using the same paradigm as in Experiment 1. Like German, Italian is an intonation language which highlights words by means of pitch accents (Grice et al., 2001; D'Imperio, 2002; Gili Fivela et al., 2015).[11] Importantly, Italian has a different set of pitch accents and phonetic realizations of these accents than German. It is hence well suited to act as control condition for the performance of the L1 speakers of a tone language who acquire an intonation language (see Experiment 1). If the differences in the realization of the accentual contrasts between the Mandarin learners of German and the German natives is indeed caused by CLI [i.e., that Mandarin participants perceive (LH)* and L + H* as similar, but L* + H as distinct from the two], we expect the Italian learners to produce contrasts closer to the German target (and hence more distinct from the Mandarin learners). If the two learner groups (Mandarin vs. Italian) do not differ, the underlying cause may also be a language-independent psychoacoustic processing mechanism or specific properties of the stimuli.

Note that we keep the terms "non-merger contrast" for (LH)* vs. L* + H and "merger contrast" for (LH)* vs. L + H* also for Italian participants – even though they were established based on the perception of (dis)similarity by Mandarin listeners, since this makes comparison to the Mandarin data easier.

Methods

We used the same online imitation experiment as in Experiment 1, but tested a group of L1 Italian speakers with low proficiency in L2 German.

Participants

We recruited eight low-proficient Italian learners of German (6 female, 2 male; mean age: 29 years, SD: 9.25). They were from the North/Centre of Italy (region of birth: Piedmont: one speaker, Lombardy: four speakers, Veneto: one speaker, Trentino: one speaker, Tuscany: one speaker). One of them lived in Germany and one lived in the US at the time of testing. On average, the participants studied German for 2.9 years (SD: 1.8). Regarding self-rated proficiency based on the European reference framework (Council of Europe, 2020), they most often indicated their level as B1 (A1: two speakers, A2: one speaker, B1: four speakers, B2: one speaker). The Italian low-proficiency group had a mean DIALANG score of 46.3 (SD = 5.4); the score did not differ from the score of the low-proficient Mandarin speakers (45.3, $p > 0.7$). The mean foreign accent rating was 3.9 (SD = 1.5) and did not differ from the Mandarin Chinese participants' rating either ($p > 0.2$).

Materials, procedure and data treatment

The materials, procedure and data treatment were the same as in Experiment 1, except that the segmentation of the four critical intervals ([n] from *denn,* and the three syllables of the sentence-final object) was done manually instead of using *WebMAUS* for an initial segmentation. Six imitations had to be excluded from the analyses due to background noise, hesitations, pauses, or lexical mistakes.

Results and discussion

The data were processed and analyzed as in Experiment 1.

Non-merger contrast: (LH)* vs. L*+H

We first analyzed the data in analogy to the Mandarin Chinese data. For direct comparison between learner groups, we display

[10] Note that in both contrasts, the two Mandarin Chinese groups started with higher f0 than the German group (see Panels A in Figures 5, 6). This difference in pitch scaling is not very relevant, however. In our analysis, we did not compare accent realizations across groups, but the realization of accentual *contrasts* across groups. This allows us to abstract from the generally higher pitch level in the learners' productions prior to the accent.

[11] Unlike German, it lacks post-focus deaccentuation (Swerts et al., 2002, for experimental evidence), but this difference is not relevant as we are dealing with utterances that have the nuclear accent on the last word.

the low-proficient Italian data in the left panel of the figures and the low-proficient Mandarin data in the right panel (Panel C and D, Figure 7). We first combined the Italian data with the German data and tested whether a model with a smooth term for the interaction between *language* and *intonation condition* over time was better than a model with a condition-smooth only, which was the case [$\chi^2(9.00) = 20.011$, $p < 0.001$]. This final model (with the scat-linking function) explained 68.0% of the variance.

Figure 7 shows the two accent conditions of the non-merger contrast (Panel A), followed by difference curves for German (Panel B) and the two learner groups (Panel C). The Italian learners' realizations of the contrast are closer to the German native speakers' than the Mandarin learners' realizations (even though the contours also start to diverge a little earlier than for German native speakers).[12] This difference between learner groups is supported by the difference of the difference plots, which show differences between the Italian and German realization of the contrast (Panel D). These plots (Panel D) also reveal that both groups deviate from German native speakers (both deviate from 0). The accentual realization of Italian learners mostly differed in the post-stressed syllable from the German native speakers, but overall, the contrast was acoustically reduced. As will be discussed in the "General Discussion," this temporal interval for the deviance may potentially be explained by the Italian accentual system, lending further support to CLI (H3).

The descriptive difference in the realization of the contrast between Mandarin Chinese and Italian learners (Panel C and D, left and right) is statistically corroborated as follows: We generated a derived dependent variable that captures the deviance of a learner from the average German speaker. To this end, we averaged the f0 values of the German speakers for each time point and subtracted this value from the learners' f0 values over time. We then run the GAMM with this derived dependent variable, testing whether an interaction term for *condition* and *learner group* is significant. Model comparisons showed that the model with the interaction was significantly better than the model without the interaction term [$\chi^2(9.00) = 118.180$, $p < 2e{-}16$]; it accounted for 64.6% of the variance. The difference between learner groups in the deviance from German native speakers is directly shown in Panel E. Since the realization of the contrast in the post-stressed syllable is opposite in the two learner groups, the difference between these two groups is aggravated in this time interval.

Merger contrast: (LH)* vs. L+H*

The realization of the contrast between (LH)* and L + H* is shown in Figure 8. Italian learners did not realize the contrast but merged the two contours (Panel C left), leading to a significant difference compared to the German native speakers (Panel D left). Recall that the Mandarin learners realized this contrast (Panel C right), but less distinctly than the German native speakers. A direct comparison of the realization of this contrast across L1s (Mandarin Chinese vs. Italian) revealed that the interaction between *language* and *condition* was not significant (and is therefore not shown). Hence, there is no evidence to postulate differences in the realization of the contrast across learner groups (Italian vs. Mandarin Chinese).

The low-proficient Italian learners of German realized both contrasts less distinctly than the German native speakers. The non-merger contrast [(LH)* vs. L* + H] resulted in a significant difference across learner groups (with Mandarin Chinese learners deviating more from the German native speakers than the Italian learners). Given that the proficiency was largely matched across groups, the difference in imitation is likely due to the prosodic system in the native language (tone language vs. intonation language). For the merger contrast, learner groups did not significantly differ from each other; both realized the contrast significantly less distinctly than German native speakers. Note that the average contours of the accents [(LH)* vs. L + H*] for Italian speakers (Panel A) might suggest a difference, but there was great variance (broader confidence intervals in Panel C) and a small number of learners (eight Italian learners as compared to 14 Mandarin learners) – factors that may have prevented this descriptive difference to reach statistical significance.

General discussion

The present study addressed the possibility of crosstalk between tone and intonation by studying the L2 acquisition of pitch accents [German L + H*, (LH)*, and L* + H] by Mandarin Chinese learners of German. Introspective judgements by Mandarin Chinese L1 speakers had suggested that (LH)* and L + H* may be prone to a merger effect because they are perceived as similar, and clearly different from L* + H. We hence based our predictions and analyses on two kinds of pitch accent contrasts, both involving a comparison to the acoustically intermediate condition, i.e., to the (LH)* accent: (1) a "non-merger contrast," (LH)* vs. L* + H, and (2) a "merger contrast," (LH)* vs. L + H*. Based on the literature, we formulated three hypotheses for the realization of these two pitch accent contrasts by L2 speakers. The first two are general hypotheses that are based on the fact that Mandarin is a tone language, the third hypothesis is based on CLI of lexical tone on pitch accents in the L2. H1 stated that Mandarin Chinese learners are equally good in imitating the two pitch accent contrasts as German native speakers because of an enhanced sensitivity to pitch in general (i.e., same pattern for both non-merger and merger contrast) with no effect of proficiency. Our data clearly falsified H1. The data also falsified H2, which stated a general disadvantage for acquiring intonational pitch accents for Mandarin Chinese learners. However, our data are partly compatible with H3, which stated that Mandarin Chinese learners produce the non-merger

[12] The confidence intervals are broader for the Italian group, probably owing to the smaller number of participants.

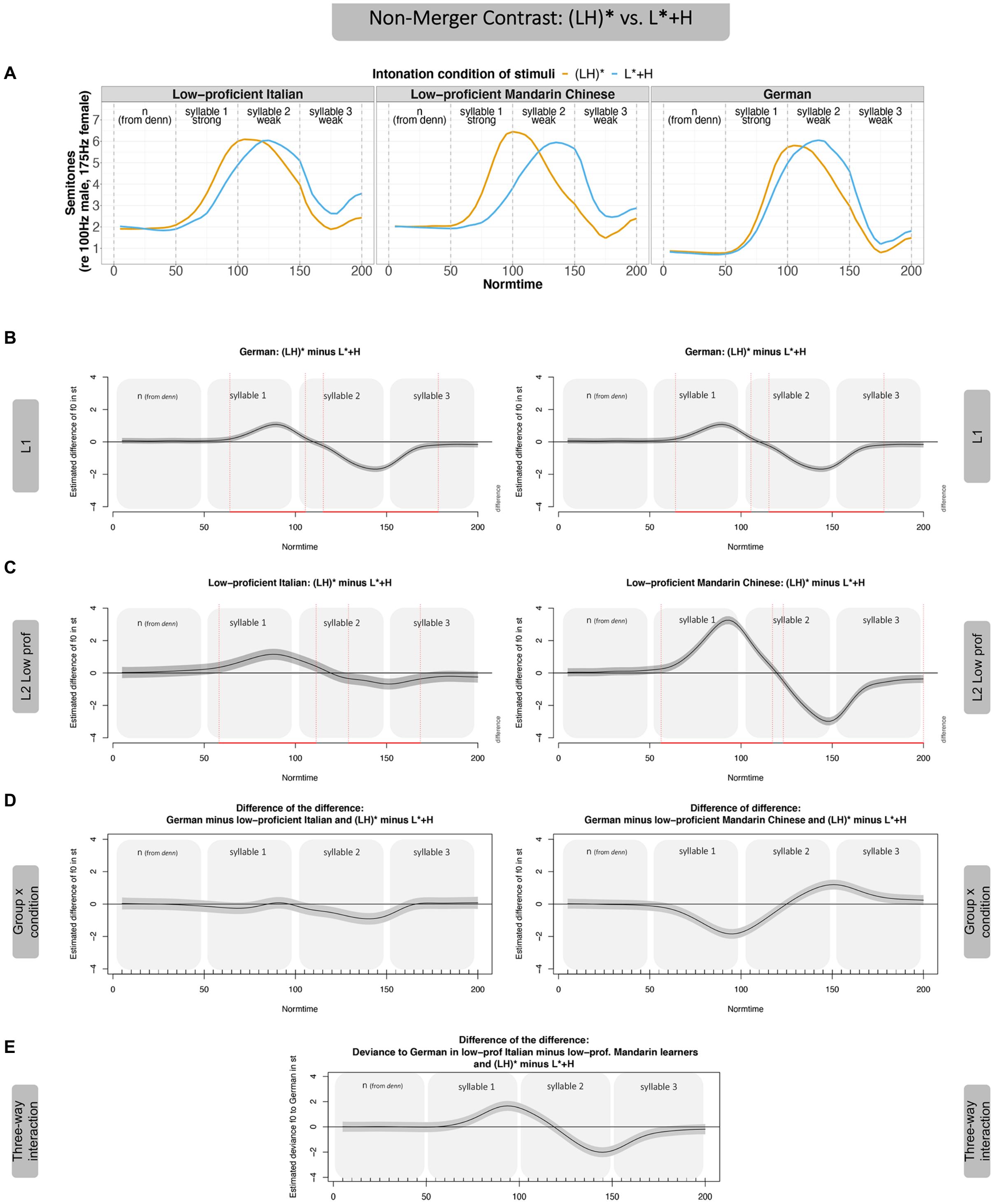


FIGURE 7
GAMM results—Non-Merger Contrast, (LH)* vs. L*+H, by Italian learners. Panel **A** shows the realization of the two contours of the non-merger contrast across participant groups. Panel **B** shows the difference curves for German, Panel **C** for the two low-proficient learner groups (Italian left, Mandarin Chinese repeated right). Panel **D** shows the difference of the difference, directly comparing the realization of the contrast compared to German native speakers. Panel **E** shows a direct comparison of the difference of the L2 groups compared to German.

contrast [(LH*) vs. L* + H] equally distinct as German natives or even more pronounced, and the merger contrast [(LH)* vs. L + H*] less distinct compared to German natives due to CLI. Our findings support Mennen's L2 intonation model (Mennen, 2015) in showing that the perceptual (dis)similarities are a relevant factor for the successful

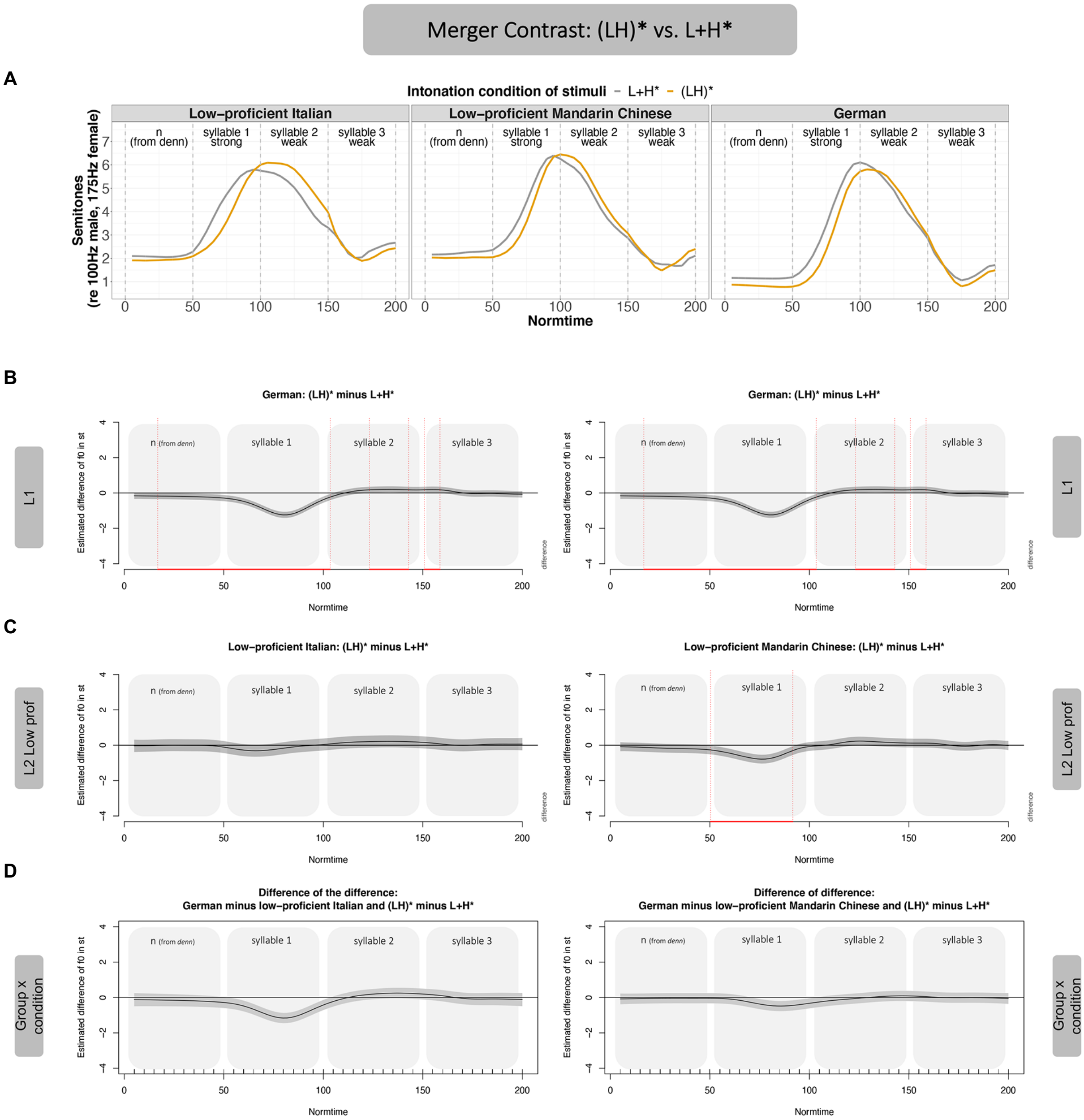


FIGURE 8
GAMM results—Merger Contrast, (LH)* vs. L+H*, by Italian learners. Panel **A** shows the realization of the two contours of the merger contrast across participant groups. Panel **B** shows the difference curves for German, Panel **C** for the two low-proficient learner groups (Italian left, Chinese repeated right). Panel **D** shows the difference of the difference, directly comparing the realization of the contrast compared to German native speakers.

acquisition of pitch accent contrasts. Increased proficiency was claimed to reduce the effect of CLI in the L2 productions, which was the case for the non-merger contrast (but not for the merger contrast). In terms of proficiency, our findings only partly support Mennen's LILt (Mennen, 2015). Experiment 2, a control experiment with native speakers of an intonation language (Italian), corroborated the effects of CLI: Their productions of both contrasts were closer to the German speakers' productions than the productions by L1 Mandarin learners, in particular in the non-merger contrast condition.

Given the comparatively few studies on the phonological acquisition of pitch accents to date, it is difficult to devise models on this kind of CLI at this point. Clearly, more research from other typologically different L1s is necessary to corroborate the cross-linguistic differences and to disentangle the specifics of the L1 influence. In future research, we also plan to complement the imitation data by perceptual tasks (same-difference task) to locate the source of the CLI (in perception or in production). Moreover, it will be useful to test the more general hypotheses H1 and H2 with populations that have no or only very little experience with intonation languages (e.g., school pupils) to minimize effects of exposure. For such an endeavor the delayed imitation paradigm may be too challenging, though. Instead, a simplified version of the task, as has been used in other studies (e.g., an immediate imitation paradigm with multiple exposure to the target utterances, D'Imperio et al., 2014; Zahner-Ritter et al., 2021; Zhao, 2022), may be better suited because it allows participants to directly access the acoustic trace. Another way to simplify the task would be to use shorter utterances (only the object noun), and/or reiterant speech (Larkey, 1983; Rietveld et al., 2004).

In the remainder of this section, we reflect on the nature of CLI and the crosstalk between intonation and lexical tone in our data ("Crosstalk between tone and intonation and cross-linguistic influence") before we briefly turn to the effect of proficiency ("Proficiency").

Crosstalk between tone and intonation and cross-linguistic influence

The type of crosstalk we observe between tone and intonation in L2 acquisition is one of general nature and difficult to disentangle from CLI as our results do not fully support H3. Tone language learners of an intonation language did not generally profit from their enhanced pitch processing abilities shown in other domains (crosstalk, see Wang et al., 2003; Wong et al., 2007; Zatorre and Gandour, 2008; Pfordresher and Brown, 2009; Bidelman et al., 2011, 2013; Bidelman and Chung, 2015). Otherwise, we would have expected target-like realizations of the contrasts, which was not the case. Also, the documented difficulty in intonational processing in their L1, in tonal L2s, and in non-native speech did not transfer to the acquisition of German pitch accents. If this had been the case, we would have expected a poor realization of the contrast across the board. Rather, we observed a nuanced (and asymmetrical) pattern in which the non-merger contrast [(LH)* vs. L* + H] was more distinct in Mandarin Chinese learners compared to the realizations of the German native speakers, while the merger contrast [(LH)* vs. L + H*] was less distinct compared to German native speakers. The merger contrast was closer to the target than the non-merger contrast, which, in turn, was clearly exaggerated. The large distinction of contours of the Mandarin participants for the non-merger contrast [(LH)* vs. L* + H] is likely due to the fact that the L* + H pitch accent was perceived differently from the other two accents, with the late peak (on the post-stressed syllable) being salient for listeners. This increased prominence on the post-stressed syllable for the L* + H might have hindered learners to perceive this contour as a pitch accent associated with the stressed syllable, followed by unstressed syllables, but at times as a pitch accent associated with the post-stressed syllable (*cf.* Kutscheid et al., 2021). Interestingly, some Mandarin productions of L* + H (in particular in the low-proficient group) sounded as if they were stressed on the post-stressed syllable (i.e., resulting in the perception of primary stress on the second syllable, as judged by German native speakers). Hence, crosstalk in our study becomes evident in that learners with a tone language as L1, in particular the low-proficient ones, seem to be influenced by their tonal phonology when processing pitch accents in the L2.

The fact that Mandarin Chinese learners were not generally disadvantaged in imitating pitch accent contrasts (contra H2) may have different explanations. Conceivably, the decreased sensitivity to pitch was mostly documented for the question-statement contrast toward the end of the utterance, while the contrast we tested was a pitch accent contrast in the middle of the utterance. Ip and Cutler (2017) and Ip and Cutler (2020) have shown that Mandarin Chinese listeners are equally good at using intonation to predict an upcoming focus as native English listeners, which suggests that tone and intonation can be integrated in online tasks. More research is needed to determine the conditions that make intonational processing harder for Mandarin speakers and those that are not problematic. Our data show that pitch accent contrasts that sound distinct to Mandarin Chinese listeners can be easily imitated/acquired in the L2.

From a broader perspective, our data show that CLI is the decisive factor in the acquisition of pitch accent contrasts. For both learner groups (i.e., for learners whose language background is either a tone language or another intonation language), specifics of the native language are able to explain the realization of the contrasts in the L2. The influence of the tonal background (Mandarin) was already discussed in the preceding paragraph. We will focus on the Italian system to understand the nature of transfer better. The Italian intonational inventory consists of two monotonal (L* and H*) and seven bitonal accents: H + L*, H* + L, L + H*, L + ¡H*, L+ < H*, L* + H, L*+ > H (Gili Fivela et al., 2015). Note that these accent types occur in a number of varieties across Italy, including varieties of northern and central Italy where our speakers came from. We briefly describe these pitch accents to explain the nature of CLI that can be expected. In L + H*, the H is aligned in the middle or at the end of the stressed syllable. In L + ¡H*, the high tonal target is also aligned at the end of the stressed syllable, but in addition is described as superhigh. In L+ < H*, the starred tone is aligned in the post-stressed syllable or even later. For these three "L + H*-variants," the alignment of the L tone is not described and may therefore not be considered relevant for the characterization of an accent. As evident from schematic representations in Gili Fivela et al. (2015, p: 148), the L alignment seems to be at the beginning of the stressed syllable. In L* + H and L*+ > H (i.e., the "L* + H-variants"), there is a fall to the stressed syllable before the accentual rise. Other than that, both tonal targets are aligned in the stressed syllable. In terms of a potential mapping from L1 to L2 categories, Italian L + H* could be mapped onto German L + H*, Italian L + ¡H* on German (LH)* – if we assume that a superhigh peak results in a steeper slope – and Italian L+ < H* on German L* + H. Given that such a mapping is possible, Italian learners ought to be well equipped to imitate the German accentual contrast. However, we observe some differences in accent realization: In the non-merger contrast [(LH)* vs. L* + H], Italian speakers mainly differed in the post-stressed syllable from the German natives, maybe owing to the fact that there are no rising accents with a late peak [the only rising accents (L* + H and L*+ > H) are preceded by a fall]. The merger contrast [(LH)* vs. L + H*], in turn, was completely mapped onto one contour in Italian learners, with no difference between contours. This finding cannot readily be explained by the Italian phonological system. If anything, it is possible that the actual phonetic alignment differs between Italian and German and that Italian learners of German were not able to perceive a difference between the two accents. We will have to leave this open question to be tested in future research. What is even more important, however, is the comparison of the two learner groups. Here, Italian learners did not differ from Mandarin Chinese learners in the merger contrast, but were closer to the native German speakers in the non-merger contrast.

Contrary to what was predicted by H1 and H2, our data do not suggest that speakers of a tone language may acquire intonational contrasts *generally* more easily or with greater difficulty than speakers of an intonation language. The deviations from the target group realized by learners could – by and large – be explained by the properties of their native language, i.e., CLI (H3). In other words, what we observe is transfer from the L1 to the L2 – a phenomenon that has been shown to occur in various different L2 studies for both segmental (e.g., Flege et al., 1997; Abrahamsson and Hyltenstam, 2009; Hattori and Iverson, 2009; Schmid et al., 2014) and suprasegmental aspects (e.g., Mennen, 1998; Atterer and Ladd, 2004; Arvaniti et al., 2006; Zahner and Yu, 2019; Manzoni-Luxenburger, 2021). As already pointed out by Flege and Bohn (2022), it is difficult to operationalize the perceived phonetic (dis)similarity between L1 and L2 categories on the segmental level. This may hold true even more so for the comparison of tonal and intonational contrasts on the supra-segmental level. We used judgements by L1 Chinese informants without knowledge of German on the distinction between contrasts, resulting in a merger (similar) and non-merger (dis-similar) contrast. Yet, our informants had difficulties mapping the accentual realizations in an unknown L2 onto lexical tone sequences. One possibility to overcome this issue and arrive at a measure of (dis)similarity between L1 and L2 categories would be to have listeners judge how close L2 realizations of pitch accents of the noun (e.g., *Mandalas*) are to trisyllabic tone sequences. However, this kind of data would also rely on metalinguistic judgments. We believe that the imitation paradigm is better-suited to determine phonetic (dis)similarity, as it provides a more direct window into the representations of developing accent categories in the L2. Nevertheless, it stands to reason whether the categories of a tone language might be *per se* more distant than the pitch accents of any other intonation language.

A further factor that may explain differences between Mandarin Chinese and Italian learners (but not the differences in the realization of the two kinds of contrasts for Mandarin Chinese learners) is lexical proximity – the two low-proficient learner groups were matched in proficiency (both when measured in DIALANG and in perceived foreign accentedness). The lexical items, which were chosen to contain mostly sonorant segments, may have been more familiar to Italian than to Mandarin Chinese participants. In particular, the drink "Malibu," the soccer team position "Libero" and the coloring picture "Mandala" are German-Italian cognates and hence exist in the Italian lexicon as well, while they do not exist in Mandarin Chinese. Due to their comparably low lexical frequency[13], it is very unlikely that they are part of the (average) L2 lexicon, so that they must be considered novel words for Mandarin Chinese learners. However, it is not entirely clear how the presence of cognates could have affected the imitation task: On the one hand, the presence of cognates may allow Italian participants to focus on prosody more. For instance, Italian speakers are well able to imitate an alignment pattern of a different Italian variety (D'Imperio et al., 2014; but note that the task may have been easier than the task in the present study because participants did not have to wait before initiating the imitation). On the other hand, the presence of cognates may strengthen L1 transfer, as has been shown for the production of VOT in Spanish learners of English (Amengual, 2012) or phonological /s/ in Spanish-English bilinguals (Brown and Harper, 2009). Note, however, that the main argument of this paper concerns the realization of the two kinds of contrasts for Mandarin Chinese learners, which is unaffected by these lexical considerations, as the items are assumed to be equally unknown to both Mandarin Chinese groups.

The pitch accent contrast between (LH)* and L* + H (non-merger contrast) was acoustically more pronounced than the contrast between (LH)* and L + H* (merger contrast) in all groups. Actually, the terms "non-merger" and "merger" contrast were chosen based on the way Mandarin Chinese speakers perceive the pitch accents. It seems, however, that the merger contrast, for which the pitch peak was aligned with the stressed syllable for both accents, was subtle in terms of f0 differences overall (even for native speakers, Zahner-Ritter et al., 2022). Contexts in which the (LH)* accent occurs in German are attitudinally loaded utterances (rhetorical questions, *cf.* Braun et al., 2019), utterances that signal surprise (Kohler, 2005; Wochner, Forthcoming, for exclamatives) or utterances that mainly signal surprise, aversion, or correction (for declaratives, see Zahner-Ritter et al., 2022). In rhetorical questions and exclamatives, the (LH)* accent is accompanied by further prosodic modification, in particular lengthening and non-modal voice quality (Braun et al., 2019; Wochner, Forthcoming), not necessarily co-occurring with the accented word but occurring across the utterance. Listeners, in turn, do not only use information on the pitch accent type when identifying rhetorical questions, but additionally use durational and voice quality cues (Kharaman et al., 2019). Intensity and voice quality could not be analyzed with the present data set because of remote data collection; the fact that

[13] The words are not even listed in the CELEX corpus (Baayen et al., 1993); dlexDB (Heister et al., 2011) reveals a very low frequency, ranging from 1 for "Mandalas" (0.43 occurrences per million) to 116 for "Melanie" (50.43 o.p.m.).

participants used their own microphones led to great differences in recording quality, which does not lend itself to further phonetic analysis.

It may hence be the case that we are currently overlooking critical aspects when focusing on the analysis of f0 contours only. In future studies, we plan to investigate how tone in the L1 affects the production of pitch accents in their entirety, including durational aspects, voice quality, and intensity in the entire utterance. Post-hoc durational analyses of the object noun of the present data show no effects of group or intonation condition on the duration of the first, stressed syllable and on the last syllable. For the second syllable, however, there was a significant interaction between language group and condition: German and Italian participants did not modulate duration as a function of intonation condition. Mandarin participants (in particular low-proficient speakers), on the other hand, produced longer syllable durations in the L*+H condition than in the other two intonation conditions. This lends further evidence to the observation that some low-proficient Mandarin learners of German may produce a different metrical structure compared to Italian learners or German native speakers.

The present study focused on the phonetic and phonological acquisition of pitch accent contrasts. A further desideratum is to test whether learners can actually use the contrasts in appropriate contexts, which is a key requisite for correct acquisition and successful communication (*cf.* Mennen, 2015).

Proficiency

In this paper, we compared low- and high-proficient Mandarin Chinese learners of German. For the non-merger contrast, the realization of the contrast in the high-proficient group was closer to native speakers than in the low-proficient group; for the merger-contrast, no effect of proficiency was observed.[14] The beneficial effect of proficiency for the non-merger contrast is in line with previous studies (e.g., Baker, 2010; He et al., 2012; Graham and Post, 2018; Shang and Elvira-García, 2022) and experience has been considered in models of L2 acquisition (Flege, 1995; Best and Tyler, 2007; Mennen, 2015; Flege and Bohn, 2022), see Piske (2007) and Tyler (2019) on the relevance of experience in a classroom setting. Importantly, proficiency effects were not observed across the board in our data. In the merger contrast [(LH)* vs. L+H*, which was perceived as similar by Mandarin Chinese listeners], both low- and high-proficient Mandarin learners deviated equally from the native speakers' productions. It hence seems that CLI was stronger than proficiency and may have overwritten the beneficial effect of proficiency. Here, it might be interesting to test an immersed learner group to see whether in such a group, the native Mandarin pattern may be inhibited by German to arrive at target-like productions. It is also conceivable, however, that the realization of the merger contrast, (LH)* vs. L+H*, was already very target-like in the low-proficient group, leaving no room for a positive effect of proficiency (ceiling effect). A way to mathematically model the effect of proficiency is to use the PENTA model (Xu, 2005) and to either remove the targets for tones or to reduce the strength of target approximation with increasing proficiency.

The participants were grouped into high- vs. low-proficient speakers according to a lexical task (DIALANG), which has been argued to be suited to assess L2 proficiency (Alderson, 2005). Furthermore, it has been shown to correlate well with self-rated proficiency and general proficiency factors such as age of onset, language use, and language preference (Lloyd-Smith et al., 2020). To tap more deeply into phonetic/phonological aspects for the current set of sentences, we further solicited perceived foreign accent ratings (Levi et al., 2007; Hopp and Schmid, 2013). Interestingly, for the current data set, the DIALANG scores correlated only weakly with perceived foreign accent ratings, $r = -0.36$ ($t = -2.29$, $df = 34$, $p = 0.03$). It is hence possible that general language skills (such as vocabulary development) are partly dissociated from phonetic and phonological processing. For our purposes, we used the DIALANG score as a measure of proficiency, since otherwise, the argument would have become circular (we cannot exclude that foreign accent ratings are influenced by the intonational realization of the utterances – the very aspect we intend to study). In future research it may be promising to include proficiency as a continuous rather than a categorical variable in the statistical modeling (*cf.* Porretta et al., 2016) to derive a more fine-grained picture. We also leave the factors beyond proficiency, such as motivation, personality, attitude toward the L1 and L2, as well as language aptitude for future research, *cf.* Jilka (2009).

Conclusion

Low- and high-proficient Mandarin Chinese learners of German imitated a three-way pitch accent contrast in an intonational L2. The decisive factor in the realization of pitch accent contrasts was whether the pitch accents were perceived as dissimilar [non-merger contrast, here (LH)* vs. L*+H] or similar [merger contrast, here (LH)* vs. L+H*]. Higher proficiency led to more target-like productions, at least in the non-merger contrast. Comparisons with imitations of Italian learners of German showed that native language experience with a tone language neither yields a general disadvantage in the acquisition of L2 pitch

[14] *Post-hoc* analyses suggested by one of the reviewers indicated that the better performance of the high-proficient group was carried mainly by the speakers from the northern parts of China. Whether or not this is an effect of contact language or due to other aspects (e.g., proficiency in another intonation language, such as English) needs to be left for future research that explicitly manipulates *region* as a factor. In our case, speakers from different regions were chosen to better generalize the results, but regional differences were not our interest.

accent contrasts nor a general advantage, but clearly exhibits crosstalk between lexical tone and intonation (which can be best interpreted as CLI).

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: http://doi.org/10.17632/w293n86sjr.2.

Ethics statement

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

Author contributions

KZ-R, TZ, ME, and BB contributed to the conception of Experiment 1. ME and TZ contributed to the conception of Experiment 2 and carried out the experiment and data annotation for Experiments 1 and 2, respectively. KZ-R and BB led the statistical analyses. All authors contributed to the article and approved the submitted version.

Funding

The research was funded by the DFG as part of research unit “Questions at the Interfaces” (FOR 2111, P6, PIs Bettina Braun and Nicole Dehé), grant numbers BR 3428/4–1,2 and DE 876/3–1,2. We also acknowledge funding from the Trier Center for Language and Communication (TCLC, Patterns, *Forschungsinitiative Rheinland-Pfalz 2019–2023*).

Acknowledgments

We thank Justin Hofenbitzer, Hendrik Behrens-Zemek, Sophie Kutscheid, Naomi Reichmann, Friederike Hohl, Angela James, Elena Schweizer, Julia Kretschmer, Sophia Fünfgeld, Zarah Kampschulte, Melanie Loes, and Vivien Meyer for rating the foreign accentedness of the participants and for help in recruiting participants; Friederike Hohl for help with WebMaus, Hendrik Behrens-Zemek for running ProsodyPro, and to Lena Friedek, Justin Hofenbitzer, Friederike Hohl, Naomi Reichmann, and Elena Schweizer for help with segmenting the Italian data. We also thank Angela James for proof-reading the final version of the manuscript and to Melanie Loes for checking the references. For infrastructural support, we thank the LingLabs (core facilities, University of Konstanz). Finally, we are grateful to the helpful comments provided by two reviewers and the editor.

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.2022.903879/full#supplementary-material

References

Abrahamsson, N., and Hyltenstam, K. (2009). Age of acquisition and nativelikeness in a second language – listener perception vs. linguistic scrutiny. *Lang. Learn.* 59, 249–306. doi: 10.1111/j.1467-9922.2009.00507.x

Alderson, J. C. (2005). *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*. London: A&C Black.

Amengual, M. (2012). Interlingual influence in bilingual speech: Cognate status effect in a continuum of bilingualism. *Biling.: Lang. Cogn.* 15, 517–530. doi: 10.1017/S1366728911000460

Anderson-Hsieh, J., Johnson, R., and Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Lang. Learn.* 42, 529–555. doi: 10.1111/j.1467-1770.1992.tb01043.x

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.

Arvaniti, A., Ladd, D. R., and Mennen, I. (2006). Tonal association and tonal alignment: evidence from Greek polar questions and contrastive statements. *Lang. Speech* 49, 421–450. doi: 10.1177/00238309060490040101

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., Piepenbrock, R., and Gulikers, L. (1993). *The CELEX lexical Database [CD-ROM]: Linguistic Data Consortium*. Philadelphia, PA: University of Pennsylvania.

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.

Baker, R. E. (2010). *The Acquisition of English Focus Marking by non-native Speakers*. United States: Northwestern University.

Batliner, A. (1989). "Wieviel Halbtöne braucht die Frage? Merkmale, Dimensionen, Kategorie," in *Zur Intonation von Modus und Fokus im Deutschen. Vol. 234.* eds. H. Altmann, A. Batliner and W. Oppenrieder (Tübingen: Niemeyer), 111–153.

Baumann, S. (2006). *The Intonation of Givenness: Evidence from German*. Niemeyer: Tübingen.

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

Beckman, M. E., and Pierrehumbert, J. (1986). Intonational structure in English and Japanese. *Phonol. Yearb.* 3, 255–309. doi: 10.1017/S095267570000066X

Best, C. T. (2019). The diversity of tone languages and the roles of pitch variation in non-tone languages: Considerations for tone perception research. *Front. Psychol.* 10:364. doi: 10.3389/fpsyg.2019.00364

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. M. J. Munro and O.-S. Bohn (Amsterdam: John Benjamins), 13–34.

Bidelman, G. M., and Chung, W. L. (2015). Tone-language speakers show hemispheric specialization and differential cortical processing of contour and interval cues for pitch. *Neuroscience* 305, 384–392. doi: 10.1016/j.neuroscience.2015.08.010

Bidelman, G. M., Gandour, J. T., and Krishnan, A. (2011). Musicians and tone-language speakers share enhanced brainstem encoding but not perceptual benefits for musical pitch. *Brain Cogn.* 77, 1–10. doi: 10.1016/j.bandc.2011.07.006

Bidelman, G. M., Hutka, S., and Moreno, S. (2013). Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: Evidence for bidirectionality between the domains of language and music. *PLoS One* 8:e60676. doi: 10.1371/journal.pone.0060676

Boersma, P., and Weenink, D. (2016). Praat: doing phonetics by computer. Version 6.0.23 (version depended on labeller) [computer program]. Available at: https://www.fon.hum.uva.nl/praat/

Braun, B., and Biezma, M. (2019). Prenuclear L*+H activates alternatives for the accented word. *Front. Psychol.* 10:1993. doi: 10.3389/fpsyg.2019.01993

Braun, B., Dainora, A., and Ernestus, M. (2011). An unfamiliar intonation contour slows down online speech comprehension. *Lang. Cogn. Process.* 26, 350–375. doi: 10.1080/01690965.2010.492641

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., Galts, T., and Kabak, B. (2014). Lexical encoding of L2 tones: The role of L1 stress, pitch accent and intonation. *Second. Lang. Res.* 30, 323–350. doi: 10.1177/0267658313510926

Braun, B., and Johnson, E. K. (2011). Question or tone 2? How language experience and linguistic function guide pitch processing. *J. Phon.* 39, 585–594. doi: 10.1016/j.wocn.2011.06.002

Brown, E., and Harper, D. (2009). Phonological evidence of crosslinguistic exemplar connections. *Stud. Hispanic and Lusophone Linguis.* 2, 257–274. doi: 10.1515/shll-2009-1052

Cao, J. (1992). On neutral-tone syllables in Mandarin Chinese. *Can. Acoust.* 20, 49–50.

Chao, Y. R. (1930). A system of tone-letters. *Le Maître Phonétique* 45, 24–27.

Chao, Y. R. (1956). "Tone, intonation, singsong, chanting, recitative, tonal composition and atonal composition in Chinese" in *For Roman Jakobson: Essays on the Occasion of His Sixtieth Birthday*. eds. M. Halle, H. Lunt, H. McLean and C. V. Schooneveld (The Hague, The Netherlands: Mouton Publishers), 52–59.

Chen, A. (2014). "Ultimate attainment in paralinguistic intonational meaning in a second language" in *Where the Principles Fail. A Festschrift for Wim Zonneveld on the Occasion of his 64th Birthday*. eds. R. Kager, J. Grijzenhout and K. Sebregts (Utrecht: UiL-OTS)

Chen, Y., and Braun, B. (2006). Prosodic realization in information structure categories in Standard Chinese. *Proceedings of the 3rd International Conference on Speech Prosody*, Dresden, Germany.

Chiao, W.-H., Kabak, B., and Braun, B. (2011). When more is less: Non-native perception of level tone contrasts. *Proceedings of the Psycholinguistic Representation of Tone Conference 2011*, Hong Kong, China, 42–45.

Connell, B. (2017). Tone and intonation in Mambila. *Intonation in African Tone Lang.* 24, 131–166. doi: 10.1515/9783110503524-005

Council of Europe (2020). Framework of reference for languages: learning, teaching, assessment - companion volume. Available at: https://www.coe.int/en/web/common-european-framework-reference-languages

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334. doi: 10.1007/BF02310555

Cutler, A., and Chen, H. C. (1997). Lexical tone in Cantonese spoken-word processing. *Percept. Psychophys.* 59, 165–179. doi: 10.3758/BF03211886

D'Imperio, M. (2001). Focus and tonal structure in Neapolitan Italian. *Speech Comm.* 33, 339–356. doi: 10.1016/S0167-6393(00)00064-9

D'Imperio, M. (2002). Italian intonation: An overview and some questions. *Probus* 14, 37–69. doi: 10.1515/prbs.2002.005

D'Imperio, M., Cavone, R., and Petrone, C. (2014). Phonetic and phonological imitation of intonation in two varieties of Italian. *Front. Psychol.* 5:1226. doi: 10.3389/fpsyg.2014.01226

DiCanio, C., Benn, J., and García, R. C. (2018). The phonetics of information structure in Yoloxóchitl Mixtec. *J. Phon.* 68, 50–68. doi: 10.1016/j.wocn.2018.03.001

Elordieta, G. (2003). The Spanish intonation of speakers of a Basque pitch-accent dialect. *Catalan J. Linguis.* 2, 67–95. doi: 10.5565/rev/catjl.44

Flege, J. E. (1995). "Second language speech learning: Theory, findings, and problems" in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. ed. W. Strange (Baltimore: York Press), 233–276.

Flege, J. E., and Bohn, O.-S. (2022). "The revised speech learning model (SM-r)," in *Second Language Speech Learning: Theoretical and Empirical Progress*. ed. R. Wayland. (Cambridge University Press), 3–83.

Flege, J. E., Bohn, O.-S., and Jang, S. (1997). The effect of experience on nonnative subjects' production and perception of English vowels. *J. Phon.* 25, 437–470. doi: 10.1006/jpho.1997.0052

Gandour, J. (1983). Tone perception in far eastern languages. *J. Phon.* 11, 149–175. doi: 10.1016/S0095-4470(19)30813-7

Gandour, J., Wong, D., Hsieh, L., Weinzapfel, B., Van Lancker Sidtis, D., and Hutchins, G. (2000). A crosslinguistic PET study of tone perception. *J. Cogn. Neurosci.* 12, 207–222. doi: 10.1162/089892900561841

Gandour, J., Wong, D., and Hutchins, G. (1998). Pitch processing in the human brain is influenced by language experience. *Neuroreport* 9, 2115–2119. doi: 10.1097/00001756-199806220-00038

Gili Fivela, B., Avesani, C., Barone, M., Bocci, G., Crocco, C., D'Imperio, M., et al. (2015). "Intonational phonology of the regional varaieties of Italian," in *Intonation in Romance*. eds. S. Frota and P. Prieto (Oxford: Oxford University Press), 140–197.

Giuliano, R. J., Pfordresher, P. Q., Stanley, E. M., Narayana, S., and Wicha, N. Y. Y. (2011). Native experience with a tone language enhances pitch discrimination and the timing of neural responses to pitch change. *Front. Psychol.* 2:146. doi: 10.3389/fpsyg.2011.00146

Gottfried, T. L., and Suiter, T. L. (1997). Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. *J. Phon.* 25, 207–231. doi: 10.1006/jpho.1997.0042

Grabe, E., Rosner, B. S., García-Albea, J. E., and Zhou, X. (2003). Perception of English intonation by English, Spanish, and Chinese listeners. *Lang. Speech* 46, 375–401. doi: 10.1177/00238309030460040201

Graham, C., and Post, B. (2018). Second language acquisition of intonation: Peak alignment in American English. *J. Phon.* 66, 1–14. doi: 10.1016/j.wocn.2017.08.002

Grice, M. (1995). *The Intonation of interrogation in Palermo Italian*. Tübingen: Niemeyer.

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., D'Imperio, M., Savino, M., and Avesani, C. (2001). "Towards a strategy for ToBI labelling varieties of Italian," in *Prosodic Typology and Transcription: A Unified Approach. Collection of Papers from the ICPhS 1999 satellite Workshop on "Intonation: Models and ToBI Labeling"*. ed. S.-A. Jun (San Francisco: California)

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

Gut, U. (2009). *Non-native speech. A corpus-based Analysis of Phonological and Phonetic Properties of L2 English and German*. Frankfurt: Peter Lang.

Hattori, K., and Iverson, P. (2009). English /ɹ/–/l/ category assimilation by Japanese adults: Individual differences and the link to identification accuracy. *J. Acoust. Soc. Am.* 125, 469–479. doi: 10.1121/1.3021295

He, X., Hanssen, J., van Heuven, V. J., and Gussenhoven, C. (2012). Mandarin-accented fall, rise and fall-rise f0 contours in Dutch. *Proceedings of the 6th International Conference on Speech Prosody*, Shanghai, China, 358–361.

Heister, J., Würzner, K. R., Bubenzer, J., Pohl, E., Henneforth, T., Geyken, A., et al. (2011). dlexDB: Eine lexikalische Datenbank für die psychologische Forschung [A lexical database for research in psychology]. *Psychol. Rundsch.* 62, 10–20. doi: 10.1026/0033-3042/a000029

Holm, S. (2007). The relative contributions of intonation and duration to intelligibility in Norwegian as a second language. *Proceedings of the XVIth International Congress of the Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, 1653–1656.

Hopp, H., and Schmid, M. S. (2013). Perceived foreign accent in first language attrition and second language acquisition: The impact of age of acquisition and bilingualism. *Appl. Psycholinguist.* 34, 361–394. doi: 10.1017/S0142716411000737

Hyman, L. M. (2006). Word-prosodic typology. *Phonology* 23, 225–257. doi: 10.1017/S0952675706000893

Hyman, L. M. (2011). "Tone: Is it different?" in *The Handbook of Phonological Theory*. eds. J. A. Goldsmith, J. Riggle and A. C. L. Yu (Chichester: Wiley-Blackwell), 179–239.

Ip, M. H. K., and Cutler, A. (2017). Intonation facilitates prediction of focus even in the presence of lexical tones. *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, 1218–1222.

Ip, M. H. K., and Cutler, A. (2020). Universals of listening: Equivalent prosodic entrainment in tone and non-tone languages. *Cognition* 202:104311. doi: 10.1016/j.cognition.2020.104311

Jilka, M. (2000). *The Contribution of Intonation to the Perception of Foreign accent*. Germany: University of Stuttgart.

Jilka, M. (2009). "Talent and proficiency in language," in *Language Talent and Brain Activity*. eds. D. Grzegorz and R. Susanne Maria (Germany: De Gruyter Mouton), 1–16.

Jin, S. (1996). *An Acoustic Study of Sentence Stress in Mandarin Chinese*. Columbus: OSU.

Jun, S.-A. (2005). "Prosodic typology," in *Prosodic Typology: The Phonology of Intonation and Phrasing*. ed. S.-A. Jun (Oxford: Oxford University Press), 430–458.

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

Kharaman, M., Xu, M., Eulitz, C., and Braun, B. (2019). The Processing of Prosodic cues to Rhetorical Question Interpretation: Psycholinguistic and Neurolinguistics Evidence. *Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech)*, Graz, Austria, 1218–1222.

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. (1991). 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

Krishnan, A., and Gandour, J. T. (2009). The role of the auditory brainstem in processing linguistically-relevant pitch patterns. *Brain Lang.* 110, 135–148. doi: 10.1016/j.bandl.2009.03.005

Krishnan, A., Gandour, J. T., and Bidelman, G. M. (2010). The effects of tone language experience on pitch processing in the brainstem. *J. Neurolinguistics* 23, 81–95. doi: 10.1016/j.jneuroling.2009.09.001

Krishnan, A., Swaminathan, J., and Gandour, J. T. (2009). Experience-dependent enhancement of linguistic pitch representation in the brainstem is not specific to a speech context. *J. Cogn. Neurosci.* 21, 1092–1105. doi: 10.1162/jocn.2009.21077

Krishnan, A., Xu, Y., Gandour, J., and Cariani, P. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Brain Res. Cogn. Brain Res.* 25, 161–168. doi: 10.1016/j.cogbrainres.2005.05.004

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

Kutscheid, S., and Braun, B. (2021). Be careful what you wish for – how German speakers indirectly communicate what they want. *Talk at the Phonetik und Phonologie im deutschsprachigen Raum (P&P)*. Frankfurt, Germany.

Kutscheid, S., Zahner-Ritter, K., Leemann, A., and Braun, B. (2021). How prior experience with pitch accents shapes the perception of word and sentence stress. *Lang. Congn. Neurosci.* 37, 103–119.

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

Laniran, Y. O., and Clements, G. N. (2003). Downstep and high raising: interacting factors in Yoruba tone production. *J. Phon.* 31, 203–250. doi: 10.1016/S0095-4470(02)00098-0

Larkey, L. S. (1983). Reiterant speech: An acoustic and perceptual validation. *J. Acoust. Soc. Am.* 73, 1337–1345. doi: 10.1121/1.389237

Lee, O. J. (2005). *The Prosody of Questions in Beijing Mandarin*. United States: The Ohio State University.

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. (2019). SoSci Survey (current version: Version (3.1.06), Available at: http://www.soscisurvey.com

Levi, S. V., Winters, S. J., and Pisoni, D. B. (2007). Speaker-independent factors affecting the perception of foreign accent in a second language. *J. Acoust. Soc. Am.* 121, 2327–2338. doi: 10.1121/1.2537345

Li, P., Zhang, Y., Fu, X., Baills, F., and Prieto, P. (2022). Melodic perception skills predict Catalan speakers' speech imitation abilities of unfamiliar languages. *Proceedings of the International Conference on Speech Prosody 2022*, Lisbon, Portugal, 876–880.

Liang, J., and Heuven, V. J. (2009). "Chinese tone and intonation perceived by L1 and L2 listeners," in *Volume 2 Experimental Studies in Word and Sentence Prosody*. eds. G. Carlos and R. Tomas (Germany: De Gruyter Mouton), 27–62.

Lin, Y.-H. (2007). *The Sounds of Chinese*. Cambridge: Cambridge University Press.

Liu, M. (2018). *Tone and Intonation Processing: From Ambiguous Acoustic signal to linguistic Representation. Leiden University*. Utrecht, Utrecht: LOT.

Liu, X., and Chen, X. (2016). The acquisition of English pitch accents by Mandarin Chinese speakers as affected by boundary tones. *Proceedings of International Conference on Speech Prosody 2016*, Boston, USA, 956–960.

Liu, M., Chen, Y., and Schiller, N. O. (2016a). Context effects on tone and intonation processing in Mandarin. *Proceedings of the Speech Prosody 2016*, Boston, USA, 1056–1060.

Liu, M., Chen, Y., and Schiller, N. O. (2016b). Online processing of tone and intonation in Mandarin: Evidence from ERPs. *Neuropsychologia* 91, 307–317. doi: 10.1016/j.neuropsychologia.2016.08.025

Liu, D., and Reed, M. (2021). Exploring the complexity of the L2 intonation system: An acoustic and eye-tracking study. *Front. Commun.* 6:7316. doi: 10.3389/fcomm.2021.627316

Liu, F., and Xu, Y. (2005). Parallel encoding of focus and interrogative meaning in mandarin intonation. *Phonetica* 62, 70–87. doi: 10.1159/000090090

Lloyd-Smith, A., Gyllstad, H., Kupisch, T., and Quaglia, S. (2020). Heritage language proficiency does not predict syntactic CLI into L3 English. *Int. J. Biling. Educ. Biling.* 24, 435–451. doi: 10.1080/13670050.2018.1472208

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.

Magen, H. (1998). The perception of foreign-accented speech. *J. Phon.* 26, 381–400. doi: 10.1006/jpho.1998.0081

Manzoni-Luxenburger, J. (2021). *Luxembourgish intonation: System and Language Contact*. Luxemburg: Melusina Press.

McManus, K. (2022). *Crosslinguistic Influence and second Language Learning*. London, New York: Routledge, Taylor & Francis Group.

Mennen, I. (1998). "Second language acquisition of intonation: The case of peak alignment," in *Chicago Linguistic Society 34*. eds. M. C. Gruber, D. Higgins, K. Olson and T. Wysocki, The panels, vol. *II* (Chicago: University of Chicago), 327–341.

Mennen, I. (2004). Bi-directional interference in the intonation of Dutch speakers of Greek. *J. Phon.* 32, 543–563. doi: 10.1016/j.wocn.2004.02.002

Mennen, I. (2015). "Beyond segments: towards an L2 intonation learning theory," in *Prosody and Language in Contact - L2 Acquisition, Attition and Languages in Multilingual Situations*. eds. E. Delais-Roussarie, M. Avanzi and S. Herment (Berlin: Springer), 171–188.

Michalsky, J. (2017). *Frageintonation im Deutschen. Zur intonatorischen Markierung von Interrogativität und Fragehaltigkeit. [Question intonation in German. On the marking of interrogativity]*. Tübingen: Niemeyer.

Munro, M. J., and Derwing, T. M. (1995). 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. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Lang. Learn.* 49, 285–310. doi: 10.1111/0023-8333.49.s1.8

Munro, M. J., Derwing, T. M., and Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Stud. Second. Lang. Acquis.* 28, 111–131. doi: 10.1017/S0272263106060049

Munro, M. J., Derwing, T. M., and Sato, K. (2006). Salient accents, covert attitudes: Consciousnessraising for pre-service second language teachers. *Prospects* 21, 67–79.

Nespor, M., and Vogel, I. (1986). *Prosodic Phonology*. Dordrecht [u.a.]: Foris.

Niebuhr, O., Bergherr, J., Huth, S., Lill, C., and Neuschulz, J. (2010). Intonationsfragen hinterfragt - die Vielschichtigkeit der prosodischen Unterschiede zwischen Aussage- und Fragesätzen mit deklarativer syntax [Questioning intonation questions. On the complexity of prosodic differences in declarative and interrogatives with declarative syntactic structure]. *Zeitschrift für Dialektologie und Linguistik* 77, 304–346. Available at: http://www.jstor.org/stable/41309786

O'Brien, M. G., and Gut, U. (2010). "Phonological and phonetic realization of different types of focus in L2 speech" in *Achievements and Perspectives in the Acquisition of second Language speech: New Sounds*. eds. K. Dziubalska-Kołaczyk, M. Wrembel and M. Kul (Frankfurt: Peter Lang), 205–215.

Oppenrieder, W. (1991). "Zur intonatorischen form deutscher Fragesätze [On the intonational form of German questions]" in *Fragesätze und Fragen [Questioning sentences and questions]*. eds. M. Reis and I. Rosengren (München: De Gruyter), 243–262.

Pfordresher, P. Q., and Brown, S. (2009). Enhanced production and perception of musical pitch in tone language speakers. *Atten. Percept. Psychophysiol.* 71, 1385–1398. doi: 10.3758/APP.71.6.1385

Pierrehumbert, J. B. (1980). The phonology and phonetics of English intonation. PhD Thesis, Massachusetts Institute of Technology, Dept. of Linguistics and Philosophy, Boston, USA.

Piske, T. (2007). "Implications of James E. Flege's research for the foreign language classroom" in *Language Experience in second Language speech Learning. In honor of James Emil Flege*. eds. M. J. Munro and O.-S. Bohn (Amsterdam: John Benjamins), 301–314.

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

Porretta, V., Tucker, B. V., and Järvikivi, J. (2016). The influence of gradient foreign accentedness and listener experience on word recognition. *J. Phon.* 58, 1–21. doi: 10.1016/j.wocn.2016.05.006

Qin, Z., and Mok, P. (2011). Perception of Cantonese tones by Mandarin, English and French speakers. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS)*, Hong Kong, 1654–1657.

Ramírez Verdugo, D. (2006). Prosodic realization of focus in the discourse of Spanish learners and English native speakers. *Estudios Ingleses de la Universidad Complutense* 14, 9–32. Available at: https://www.researchgate.net/publication/39365282_Prosodic_realization_of_focus_in_the_discourse_of_Spanish_learners_and_English_native_speakers

Rietveld, T., Kerkhoff, J., and Gussenhoven, C. (2004). Word prosodic structure and vowel duration in Dutch. *J. Phon.* 32, 349–371. doi: 10.1016/j.wocn.2003.08.002

Schmid, M. S., Gilbers, S., and Nota, A. (2014). Ultimate attainment in late second language acquisition: Phonetic and grammatical challenges in advanced Dutch–English bilingualism. *Second. Lang. Res.* 30, 129–157. doi: 10.1177/0267658313505314

Shang, P., and Elvira-García, W. (2022). Second language acquisition of Spanish prosody by Chinese speakers: Nuclear contours and pitch characteristics. *Vigo Intern. J. App. Linguis.* 19, 129–176. doi: 10.35869/vial.v0i19.3762

So, C. K., and Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Lang. Speech* 53, 273–293. doi: 10.1177/0023830909357156

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

Swerts, M., Krahmer, E., and Avesani, C. (2002). Prosodic marking of information status in Dutch and Italian: a comparative analysis. *J. Phon.* 30, 629–654. doi: 10.1006/jpho.2002.0178

Taft, M., and Chen, H.-C. (1992). "Judging homophony in Chinese: The influence of tones" in *Language Processing in Chinese*. eds. H. C. Chen and O. J.-L. Tzeng (Amsterdam, The Netherlands: Elsevier), 151–172.

Trofimovich, P., and Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Stud. Second. Lang. Acquis.* 28, 1–30. doi: 10.1017/S0272263106060013

Trouvain, J., and Braun, B. (2021). "Sentence prosody in a second language" in *The Oxford Handbook of Language Prosody*. eds. C. Gussenhoven and A. Chen (Oxford: Oxford University Press)

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 and I. Mleineket al. (Berlin, New York: Walter de Gruyter)

Tyler, M. D. (2019). "PAM-L2 and phonological category acquisition in the foreign language classroom," in *A Sound Approach to Language Matters*. eds. A. M. Nyvad, M. Hejná, A. Højen, A. B. Jespersen and M. H. Sørensen (Honor: Honor of Ocke-Schwen Bohn), 607–630.

Ulbrich, C. (2013). German pitches in English: Production and perception of cross-varietal differences in L2. *Bilingualism - Lang. Cognition* 16, 397–419. doi: 10.1017/S1366728912000582

Ulbrich, C., and Mennen, I. (2016). When prosody kicks in: The intricate interplay between segments and prosody in perceptions of foreign accent. *Int. J. Biling.* 20, 522–549. doi: 10.1177/1367006915572383

van Rij, J., Hendriks, P., An Rijn, H., Baayen, R. H., and Wood Simon, N. (2019). Analyzing the time course of pupillometric data. *Trends in 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, available at: https://cran.r-project.org/web/packages/itsadug/index.html

Wang, Y., Sereno, J. A., Jongman, A., and Hirsch, J. (2003). fMRI evidence for cortical modification during learning of mandarin lexical tone. *J. Cogn. Neurosci.* 15, 1019–1027. doi: 10.1162/089892903770007407

Wichmann, A., House, J., and Rietveld, T. (2000). "Discourse constraints on F0 peak timing in English" in *Intonation: Analysis, Modelling and Technology, Text, Speech, and Language Technology*. ed. A. Botinis, vol. *15* (Dordrecht, Boston, London: Kluwer), 163–182.

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

Willems, N. (1982). *English Intonation from a Dutch Point of View*. Dordrecht: Foris Publications.

Wochner, D. (Forthcoming). *Prosody Meets Pragmatics: A Comparison of Rhetorical Questions, Information-Seeking Questions, Exclamatives, and Assertions*. PhD Thesis, University of Konstanz.

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., and Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nat. Neurosci.* 10, 420–422. doi: 10.1038/nn1872

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton [u.a.]: CRC Press.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73, 3–36. doi: 10.1111/j.1467-9868.2010.00749.x

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (*2nd ed.*). Boca Raton [u.a.]: CRC press.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f(0) contours. *J. Phon.* 27, 55–105. doi: 10.1006/jpho.1999.0086

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Comm.* 46, 220–251. doi: 10.1016/j.specom.2005.02.014

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

Xu, Y. (2019). "Prosody, tone and intonation" in *The Routledge Handbook of Phonetics*. eds. W. F. Katz and P. F. Assmann (New York: Routledge), 314–356.

Xu, B. R., and Mok, P. (2012). Cross-linguistic perception of intonation by Mandarin and Cantonese listeners. *Proceedings of the Sixth International Conference on Speech Prosody 2012*, Shanghai, China, 99–102.

Xu, B. R., and Mok, P. (2014). Cross-linguistic perception of Mandarin intonation. *Proceedings of the Seventh International Conference on Speech Prosody (Speech Prosody 2014)*, Dublin, Ireland, 638–642.

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Yuan, J. (2006). "Mechanisms of question intonation in mandarin" in *Chinese Spoken Language Processing*. eds. Q. Huo, B. Ma and H. Li (Berlin: Springer), 19–30.

Yuan, J. (2011). Perception of intonation in Mandarin Chinese. *J. Acoust. Soc. Am.* 130, 4063–4069. doi: 10.1121/1.3651818

Yuan, J., Dong, Q., Wu, F., Luan, H., Yang, X., Lin, H., et al. (2018). Pitch characteristics of L2 English speech by Chinese speakers: A large-scale study. *Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech)*, Hyderabad, India, 2593–2597.

Zahner, K., and Yu, J. (2019). Compensation strategies in non-native English and German. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 2019)*, Melbourne, Australia.

Zahner-Ritter, K., Einfeldt, M., Wochner, D., James, A., Dehé, N., and Braun, B. (2021). Testing the distinctiveness of nuclear rising-falling contours in German - integrating evidence from form and function. *Talk at the Phonetik und Phonologie im deutschsprachigen Raum (P&P2021).*

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. *Frontiers and Communication* 7:955. doi: 10.3389/fcomm.2022.838955

Zatorre, R. J., and Gandour, J. T. (2008). Neural specializations for speech and pitch: Moving beyond the dichotomies. *Philosop. Transact. Biolog. Sci.* 363, 1087–1104. doi: 10.1098/rstb.2007.2161

Zerbian, S. (2016). "Sentence intonation in Tswana (Sotho-Tswana group)" in *Intonation in African Tone Languages*. eds. L. Downing and A. Rialland (Berlin, Boston: De Gruyter Mouton), 393–434.

Zhang, J., Duanmu, S., and Chen, Y. (2021). "Prosodic systems: China and Siberia" in *The Oxford Handbook of Language Prosody*. eds. C. Gussenhoven and A. Chen (Oxford, UK: Oxford University Press)

Zhang, Y., Schmidt, E., and Post, B. (2022). Perception of intonation on neutral tone in Mandarin. *Front. Psychol.* 131. doi: 10.3389/fcomm.2022.849132

Zhao, T. (2022). *Analysis of the alignment differences of L2 German speakers in German pitch accents - Evidence from L2 German speakers of Mandarin Chinese on wh-question intonation*. Master thesis, University of Konstanz: Germany.

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 26 August 2022
DOI 10.3389/fnins.2022.909796

OPEN ACCESS

EDITED BY
Fang Liu,
University of Reading, United Kingdom

REVIEWED BY
Jing Shao,
Hong Kong Baptist University,
Hong Kong SAR, China
Yang Zhang,
University of Minnesota Health Twin
Cities, United States

*CORRESPONDENCE
Lin Chen
linchen@ustc.edu.cn
Xianming Fu
fuxianmingah@163.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience

RECEIVED 31 March 2022
ACCEPTED 01 August 2022
PUBLISHED 26 August 2022

CITATION
Wei Y, Liang X, Guo X, Wang X, Qi Y, Ali R, Wu M, Qian R, Wang M, Qiu B, Li H, Fu X and Chen L (2022) Brain hemispheres with right temporal lobe damage swap dominance in early auditory processing of lexical tones. *Front. Neurosci.* 16:909796.
doi: 10.3389/fnins.2022.909796

Brain hemispheres with right temporal lobe damage swap dominance in early auditory processing of lexical tones

Yarui Wei[1,2†], Xiuyuan Liang[3†], Xiaotao Guo[4], Xiaoxiao Wang[1], Yunyi Qi[3], Rizwan Ali[1], Ming Wu[5], Ruobing Qian[6], Ming Wang[3], Bensheng Qiu[1], Huawei Li[7], Xianming Fu[6]* and Lin Chen[3,7]*

[1]Biomedical Engineering Center, School of Information Science and Technology, University of Science and Technology of China, Hefei, China, [2]Department of Magnetic Resonance Imaging, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [3]Department of Neurobiology and Biophysics, School of Life Sciences, University of Science and Technology of China, Hefei, China, [4]Department of Otolaryngology-Head and Neck Surgery, The First Affiliated Hospital, University of Science and Technology of China, Hefei, China, [5]Department of Rehabilitation Medicine, The First Affiliated Hospital, University of Science and Technology of China, Hefei, China, [6]Department of Neurosurgery, The First Affiliated Hospital, University of Science and Technology of China, Hefei, China, [7]Clinical Hearing Center, Affiliated Eye and ENT Hospital, Fudan University, Shanghai, China

Labor division of the two brain hemispheres refers to the dominant processing of input information on one side of the brain. At an early stage, or a preattentive stage, the right brain hemisphere is shown to dominate the auditory processing of tones, including lexical tones. However, little is known about the influence of brain damage on the labor division of the brain hemispheres for the auditory processing of linguistic tones. Here, we demonstrate swapped dominance of brain hemispheres at the preattentive stage of auditory processing of Chinese lexical tones after a stroke in the right temporal lobe (RTL). In this study, we frequently presented lexical tones to a group of patients with a stroke in the RTL and infrequently varied the tones to create an auditory contrast. The contrast evoked a mismatch negativity response, which indexes auditory processing at the preattentive stage. In the participants with a stroke in the RTL, the mismatch negativity response was lateralized to the left side, in contrast to the right lateralization pattern in the control participants. The swapped dominance of brain hemispheres indicates that the RTL is a core area for early-stage auditory tonal processing. Our study indicates the necessity of rehabilitating tonal processing functions for tonal language speakers who suffer an RTL injury.

KEYWORDS

hemisphere dominance, lexical tone, mismatch negativity, stroke, brain lesion

Introduction

The hemispheric specialization of language has been evaluated in multiple studies over the years, and the general consensus is that the left hemisphere is specialized for the processing of speech (Broca, 1861; Wernicke, 1874), whereas the right hemisphere is specialized for the processing of pitch such as musical tones

(Zatorre et al., 2002; Poeppel, 2003). As to the factors responsible for this labor division between the two hemispheres, the functional hypothesis claims that the division depends on the auditory cues that serve as input signals (Whalen and Liberman, 1987; Liberman and Whalen, 2000), whereas the acoustic hypothesis claims that the division depends on the acoustic properties of input signals. Thus, the functional hypothesis predicts that linguistic pitch such as lexical tones is preferentially processed in the left hemisphere, whereas the acoustic hypothesis predicts that pitch is preferentially processed in the right hemisphere (Zatorre and Belin, 2001; Zatorre et al., 2002; Albouy et al., 2020). Moreover, a recent study reported increased activation in the right hemisphere when comprehending noisy spoken sentences in Mandarin Chinese (Song et al., 2020). As a matter of fact, neither of these two competing hypotheses can account for the full range of experimental data (Shankweiler and Studdert-Kennedy, 1967, 1975; Shtyrov et al., 2000).

Tonal languages such as Mandarin Chinese deploy lexical tones together with consonants and vowels to define word meaning. Previous neuroimaging studies, including positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), have reported that the bilateral superior temporal gyri (STG), the left anterior insula cortex, and the left middle temporal gyrus, as well as the right lateralized cortical activations in the posterior inferior frontal gyrus, are activated during the processing of lexical tone in Mandarin Chinese (Klein et al., 2001; Wong et al., 2004; Liu et al., 2006; Xi et al., 2010; Chang et al., 2014). In our previous study, spectrograms of the syllable /bai/ pronounced in four lexical tones (bai1, bai2, bai3, and bai4) illustrated that the lexical tones are characterized by varying frequencies with time, and that lexical tones have minimal effects on the voice onset time of the consonant /b/; moreover, spectrograms of the syllables /bai, /dai, and /tai/ pronounced in a flat tone (bai1, dai1, and tai1) illustrated that the syllables show relatively unchanged frequencies with time and that the consonants in the upper syllables are characterized by temporal variations as reflected by the voice onset time. Therefore, lexical tones and consonants are ideal materials for testing the two hypotheses. In our previous study (Luo et al., 2006), we proposed that the processing of a lexical tone carrying semantic information is lateralized to the right hemisphere at an early stage, but to the left hemisphere at a late stage. This so-called two-stage model (Luo et al., 2006) claims that hemisphere labor division initially depends on the acoustic properties of input signals and then depends on the functional cues in the processing from sound to meaning. Thus, the acoustic hypothesis and the functional hypothesis are not mutually exclusive, with each representing a different temporal stage of processing (Ren et al., 2009; Zhou et al., 2021). Our two-stage model resolves the debate over the cues that are used by the brain for the processing of speech sound and tonal sound.

The role of the right hemisphere in speech comprehension (Gainotti et al., 1981; Posner and Petersen, 1989; Mitchell and Crow, 2005; Lam et al., 2016; Gajardo-Vidal et al., 2018) and the speech impairments in patients with right brain damage (Gandour et al., 1988; Hagoort et al., 1996; Mitchell and Crow, 2005; Kadyamusuma et al., 2011; Gajardo-Vidal et al., 2018) have been explored since the 1980s. However, issues related to brain labor division for linguistic processing became more complicated in clinical observations. Some studies have shown that patients with left brain damage show less left lateralization in early auditory processing of consonants (Becker and Reinvang, 2007) and impairments in tone tasks for tonal languages (Gandour et al., 1992, 1996). Other studies have shown impairments of tone identification and production (Kadyamusuma et al., 2011) as well as the acoustic pattern (Gandour et al., 1988) in patients with right brain lesions. Patients with brain lesions are ideal subjects for investigating these issues. However, previous studies were mostly conducted at a behavioral level, which reflects auditory processing at a late stage, not an early stage. Thus, the impairments in the processing of lexical tones at an early stage, or a preattentive stage, and the influence of injury on the labor division of the brain hemispheres for auditory tonal processing at the electrophysiological level remain unclear. We believe that the right hemisphere dominance in early auditory processing of lexical tones would be impaired at the electrophysiological level in patients with right brain lesions. Considering the critical role of the right temporal lobe (RTL) for lexical-tone processing (Ge et al., 2015; Si et al., 2017; Liang and Du, 2018), we predicted that the impairments would be apparent in patients with RTL lesions but not in those with right non-temporal lobe (RNTL) lesions.

In the present study, we explored the hemisphere dominance in early auditory processing of lexical tones by using whole-head electric recordings of mismatch negativity (MMN) obtained from native Mandarin Chinese-speaking patients with RTL or RNTL lesions under a passive auditory oddball paradigm (Picton et al., 2000). The MMN is an index of the brain's automatic processing at an early stage (Naatanen et al., 1978), and it has been used as a probe in several studies related to the realm of pitch and music, as well as language (Luo et al., 2006; Chobert et al., 2012; Wang et al., 2013). For the source localization of MMN, many neuroimaging studies such as PET, fMRI, and magneto/encephalography (M/EEG) have proposed that beyond the bilateral STG, the right inferior frontal gyrus (IFG) contributes to MMN generation (Rinne et al., 2000; Opitz et al., 2002; Dura-Bernal et al., 2012). For the purpose of comparison, we also measured the MMN evoked with pure tones with varied frequencies, which are non-speech stimuli and are known to be dominantly processed in the right brain (Schonwiesner et al., 2005). We used the shortened Mandarin Chinese version of the Token test (De Renzi and Faglioni, 1978) to measure whether right brain injury impairs the ability of speech comprehension for Mandarin Chinese.

Materials and methods

Informed consent was obtained from the participants in accordance with the Declaration of Helsinki. The research protocols used in this study were approved by the Ethics Committee of the First Affiliated Hospital, University of Science and Technology of China.

Participants

Patients participating in this study were recruited from the First Affiliated Hospital, University of Science and Technology of China and screened using the following criteria: (i) provided informed consent after the procedure had been fully explained; (ii) native speakers of Mandarin Chinese; (iii) lesions were restricted to the right hemisphere, i.e., RTL and RNTL; (iv) right-handed before stroke onset, and musically untrained; (v) no contraindications to magnetic resonance imaging (MRI); and (vi) no medical history of audiological, mental, or neurological problems before stroke. These criteria were met in 24 patients: 11 patients with stroke in the RTL (age, 37–76 years; mean age, 57 years; two females) and 13 patients with stroke in the RNTL (age, 37–63 years; mean age, 53 years; two females). Brain lesions in these 24 patients were caused by a stroke with cerebral infarction (10 RTL: 11 RNTL) or cerebral hemorrhage (one RTL: two RNTL). Demographic, clinical, and lesion data of each patient are shown in Supplementary Table 1. Fourteen healthy age- and sex-matched control participants (age, 42–64 years; mean age, 52 years; five females) with Mandarin Chinese as their native language also volunteered to participate in this study (Supplementary Table 1). These participants were not musically trained and did not have a medical history of audiological, mental, or neurological diseases. All participants were right-handed in the assessment performed with the Edinburgh Handedness Inventory (Oldfield, 1971). The hearing thresholds of the three groups were tested by pure-tone audiometry at 500, 1, 2, and 4 kHz, as in a previous study (Robson et al., 2014). We first performed a Shapiro–Wilk test to identify whether the hearing thresholds for each group were normally distributed, and the results showed that the hearing thresholds were not normally distributed for some parts of the groups. Then, we performed the Kruskal–Wallis H-test to determine whether the hearing thresholds among the groups were matched, and the results showed no significant difference among the groups (the left ear: $X^2_{(2)} = 5.40$, $P > 0.05$; the right ear: $X^2_{(2)} = 3.63$, $P > 0.05$). The data for the hearing thresholds of subject No. 8 in the RTL group and subject No. 3 in the RNTL group were not collected. The patients reported no hearing problems before stroke.

Lesion overlay map

Structural high-resolution MRI scans of 13 patients (four RTL: nine RNTL) and CT scans of 11 patients (seven RTL: four RNTL) were acquired. The MRI scans were acquired on a 3T Philips Achieva scanner and included good T2-weighted or DWI B0 images. The CT scans were acquired on a Siemens scanner or a Philips scanner. Lesions were manually delineated by an experienced neurologist in the axial plane on each slice of the T2-weighted (eight RNTL patients; slice thickness, 5 or 5.5 mm; in-plane resolution, 1 mm), DWI (four RTL and one RNTL patients; slice thickness, 5 or 5.5 mm; in-plane resolution, 2 mm), or CT images (slice thickness, 5 mm; in-plane resolution, ≤0.5 mm) by using MRIcron (Rorden and Brett, 2000). Lesion volume was computed by multiplying the damaged area on each delineated slice by the slice thickness. The T2, DWI, and CT images of patients were transformed into standard stereotactic space (MNI) by using a clinical toolbox (www.nitrc.org/projects/clinicaltbx/) and SPM12 (www.fil.ion.ucl.ac.uk/spm). These images were resampled to yield the same voxel resolution, i.e., 1 mm^3. The lesion overlay maps for the RTL (Figure 1A) and RNTL (Figure 1B) groups displayed a distributed profile at the group level. Inspection of the lesion overlay maps and individual MRI/CT scans indicated that the patients in the RTL group mainly had lesions in the RTL, the right insular, and the right frontal lobe (Supplementary Table 1 and Figure 1A), and those in the RNTL group mainly had lesions in the right periventricular white matter, the right basal ganglia, and the right occipital lobe (Supplementary Table 1 and Figure 1B).

Stimuli

Lexical tones and pure tones were used as stimuli in this study. Lexical-tone stimuli were obtained and slightly modified from those used in our previous study (Luo et al., 2006), in which the Mandarin consonant-vowel (CV) syllables /bai1/ and /bai4/ were employed and originally pronounced by an adult male Mandarin speaker (Sinica Corpus, Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China). Pure tones were generated by Audition 3.0 (Adobe Systems Inc., Mountain View, CA, USA). The duration of each lexical tone was normalized to 350 ms, the duration of each pure tone was 200 ms, and both included a 5-ms linear rise and fall time. The lexical-tone contrast was created by a sequence of /bai1/ frequently presented as the standard stimuli and /bai4/ infrequently presented as the deviant stimuli during the auditory stream. The pure-tone contrast was created by a sequence of

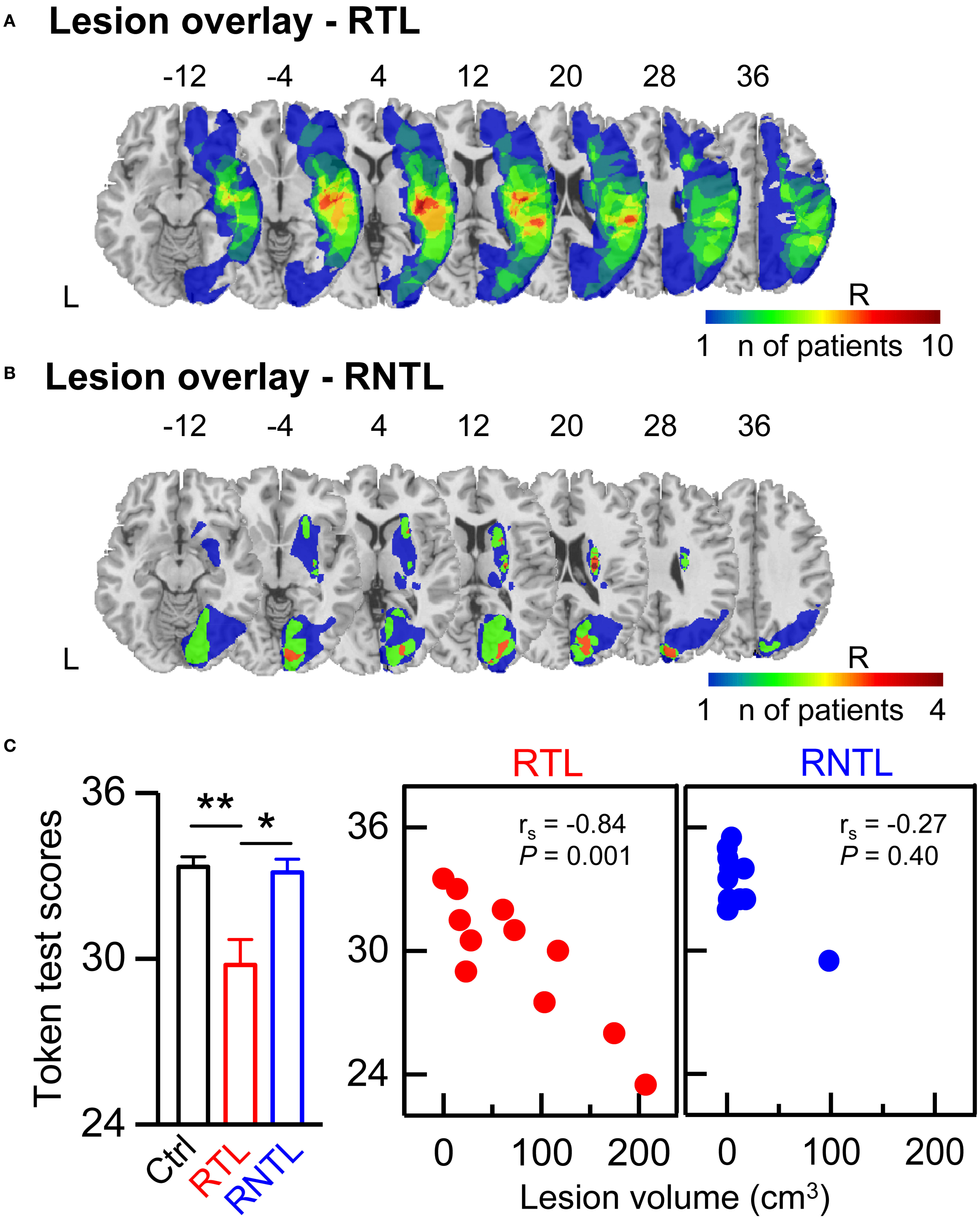


FIGURE 1
Lesion overlay map and speech comprehension ability. **(A,B)** The lesion overlap map of patients with right temporal lobe (RTL) damage **(A)** and right non-temporal lobe (RNTL) damage **(B)**. The heat map displays the number of patients with lesions in that respective area. Coordinates refer to MNI space. **(C)** Token test scores in each group. Token test scores of the RTL group were lower than that of the control group and that of the RNTL group (Left). The values are expressed as mean ± SE. *$P < 0.05$, **$P < 0.01$. Significantly negative correlation was found between lesion volume and Token test scores in the RTL group only (Central), but not the RNTL group (Right). r_s represents the correlation coefficient. L, left; R, right.

pure tones frequently presented at 550 Hz as the standard stimuli and infrequently presented at 350 Hz as the deviant stimuli.

Procedure

Two blocks including the lexical-tone contrast and the pure-tone contrast were separately and randomly presented to the participant in one session with a 5-min break. Each block consisted of 800 trials. The participants were instructed to watch a silent movie and ignore the auditory stimuli they heard. The detection thresholds of lexical tones and pure tones were measured first, and all stimuli were then presented binaurally at 78 dB above the detection thresholds for each listener through headphones (TDH-39; Telephonics, Farmingdale, NY, United States) in an electrically shielded soundproof room. The standard stimuli were presented with a probability of 7/8 and the deviant stimuli were presented with a probability of 1/8. The stimulus order was pseudorandomized while maintaining a restriction that each deviant stimulus was separated by at least two standard stimuli. The inter-stimulus onset interval was 550 ms for the lexical-tone contrast and 500 ms for the pure-tone contrast.

Data collection and analysis

The EEGs were recorded with 17 Ag/AgCl electrodes (Brain Products GmbH, Munich, Germany) placed at the standard electrode sites (F3, Fz, F4, FC1, FC2, FC5, FC6, C3, Cz, C4, P3, Pz, P4, FCz, Fpz, left mastoid, and right mastoid) according to the extended international 10–20 system. Two electrodes were used to measure the vertical and horizontal electrooculograms (EOGs). The reference electrode was attached at FCz and the ground electrode was placed between Fpz and Fz. Current signals (0.1–100 Hz) were continuously recorded by BrainAmp DC amplifier and sampled at 500 Hz. Impedances were maintained at <5 kΩ for all electrodes. The EEG and EOG data were recorded online and digitized using Brain Vision Recorder software (Brain Products, Munich, Germany).

Data from the head recordings were processed offline using Brain Vision Analyzer software (Brain Products, Munich, Germany). The recording was rejected when it was evidently contaminated by the EMG signal. An automatic ocular correction was then performed. Data were re-referenced to the average of the left and right mastoids and filtered (1–30 Hz). Epochs obtained from the continuous data were 600 ms in length, including a 100-ms pre-stimulus baseline, and were rejected when fluctuations in amplitude were >100 μV. The event-related potentials evoked by the standard and the deviant stimuli were calculated by averaging individual trials. MMN was derived from a different wave by subtracting the event-related potential evoked by the standard stimuli from that evoked by the deviant stimuli (Picton et al., 2000; Naatanen et al., 2004). Scalp topographic maps were produced using Brain Electric Source Analysis (MEGIS Software GmbH, Munich, Germany).

Speech comprehension test

We measured the speech comprehension ability of Mandarin Chinese participants by using the shortened Mandarin Chinese version of the Token test (De Renzi and Faglioni, 1978). The materials consisted of tokens of different colors (white, blue, yellow, red, and green), shapes (squares and circles), and sizes (large and small). The examinee followed verbal instructions that increased in complexity from simple commands (e.g., "Touch a circle"; "Touch the red circle") to more challenging commands such as "Before touching the yellow circle, pick up the red square." The Token test scores were adjusted for years of education. Adjusted scores between 25 and 28 were regarded as an indicator of mild comprehension problems, those between 17 and 7 indicated moderate problems, and those below 17 indicated severe or very severe problems. The Token test scores of subject No. 3 in the RNTL group were not collected.

Statistical analysis

Two sets of electrodes on the left (F3, FC1, FC5, C3) and right (F4, FC2, FC6, C4) sides of the scalp were identified as the regions of interest (Doeller et al., 2003; Luo et al., 2006). The amplitudes of MMN recorded at the four electrodes on each side were averaged within a time window from 20 ms before the peak of MMN between 100 and 300 ms recorded from electrode Fz to 20 ms after that peak (as indicated by the gray bars in the left panels of Figures 2A,B) (Wang et al., 2013). The MMN amplitude would be set to zero for all subsequent analyses when the averaged value was positive (Robson et al., 2014). The MMN latency was measured between 100 and 300 ms at electrode Fz. Then, the lateralization index (LI) for each stimulus condition and each participant was calculated by using the MMN amplitudes of the left and right sides. The LI was calculated by the following formula:

$$\begin{aligned}\text{LI} = {} & (\text{left MMN amplitude} - \text{right MMN amplitude})/ \\ & (\text{left MMN amplitude} + \text{right MMN amplitude}).\end{aligned}$$

An index value of -1 indicates a lateralized response entirely in the right hemisphere and an index value of $+1$ indicates a lateralized response entirely in the left (Seghier, 2008). Supplementary Figure 1 shows the relationship between the LI for each stimulus condition (lexical-tone contrast and pure-tone

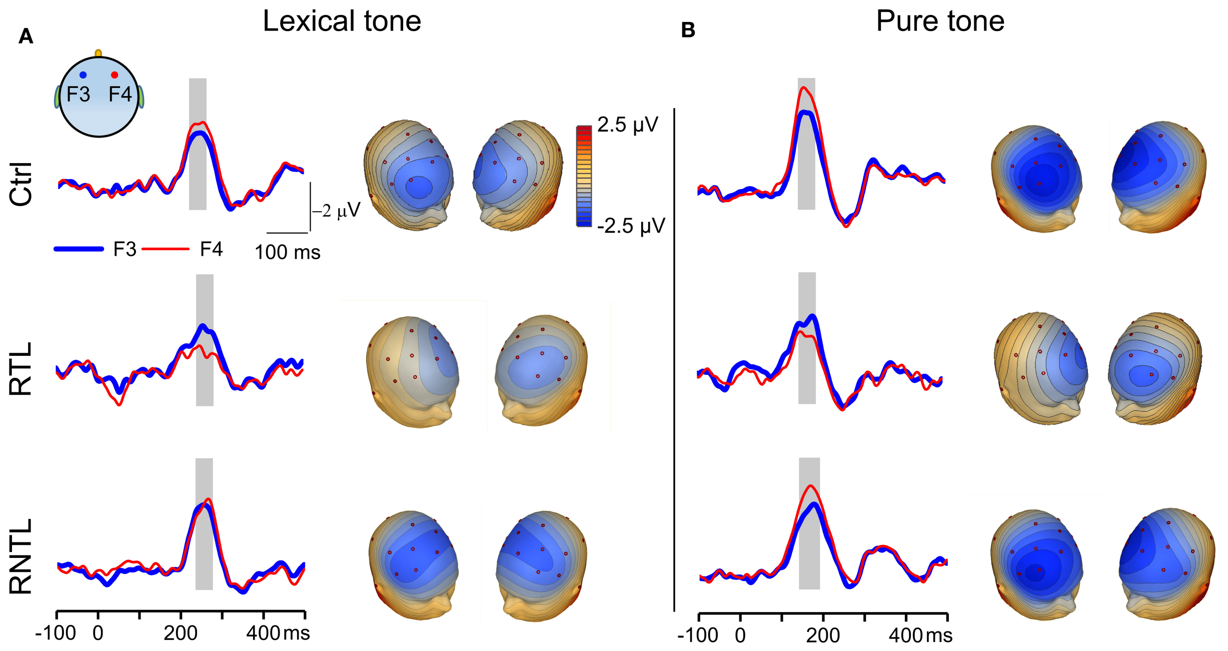


FIGURE 2
MMN responses recorded on the left and right sides of the scalp. **(A)** Grand average traces of MMN evoked by the lexical-tone contrast were recorded from one pair of electrodes on the left (F3, thick blue lines) and right (F4, thin red lines) sides in the control (Upper, $n = 14$), RTL (Central, $n = 11$), and RNTL (Lower, $n = 13$) groups (Left). Gray bars indicate the time window in which MMN amplitude was calculated. Scalp topographic maps constructed from grand average MMN evoked by the lexical-tone contrast are shown at the time point of MMN peak amplitude on electrode Fz (Right). **(B)** Grand average traces recorded from one pair of electrodes on the left and right sides (Left) and grand average scalp topographic maps (Right) of MMN evoked by the pure-tone contrast in the control (Upper, $n = 14$), RTL (Central, $n =11$), and RNTL (Lower, $n = 13$) groups. Gray bars indicate the time window in which MMN amplitude was calculated.

contrast) and the time post-stroke onset in the RTL and RNTL groups. The findings showed no significant relationship.

A Shapiro–Wilk test was performed to check the normality of Token test scores, bilateral MMN amplitudes, and MMN latencies and LI values for each group. The results showed that Token test scores and MMN latencies were normally distributed in all groups, but bilateral MMN amplitudes and LI values were not normally distributed in some groups and stimulus conditions. Welch ANOVA was performed to test the significance of differences in Token test scores among groups, given that the Token test scores among groups (control, RTL, and RNTL) failed the assumption for homogeneity of variance tested using Levene's test. *Post-hoc* pairwise comparisons with the Games–Howell test were performed. For assessing possible lateralization effects, we performed a Wilcoxon signed-rank test between bilateral MMN amplitudes for each stimulus condition and each group. Moreover, the LI for each stimulus condition and each group were assessed by using a one-sample Wilcoxon signed-rank test and compared with the value 0. An LI index value significantly <0 indicates rightward lateralization and a value significantly >0 indicates leftward lateralization. We performed the Kruskal–Wallis *H*-test and used Dunn's test as a *post-hoc* test to determine whether the LI and unilateral MMN amplitudes showed significant differences among groups (control, RTL, RNTL) for each stimulus condition. For exploring the relationships among LIs obtained under different stimulus conditions, lesion volume, and speech-comprehension ability, we obtained the Spearman's correlation coefficients for the following correlations: (i) the correlations between lesion volume and Token test scores in the RTL and RNTL groups; (ii) the correlations between lesion volume and LI for each stimulus condition in the RTL and RNTL groups; (iii) the correlations between the Token test scores and LI for each stimulus condition and each group. The statistical difference was considered significant with the alpha-level set as $P < 0.05$ for all tests. One-way ANOVA was performed to test whether MMN latency showed significant differences among groups, and *post-hoc* pairwise comparisons with the Bonferroni test were performed. All data are expressed as mean ± SE or median ± minimum/maximum value. The correlation coefficients were marked as r_s. SPSS V13 software (IBM, USA) and OriginPro V8 software (OriginLab Corp., USA) were used for statistical analysis and graph plotting.

Results

Speech comprehension ability for each group and its correlation with lesion volume

There was a significant difference in Token test scores among groups as determined by Welch ANOVA [$F_{(2,19.50)} = 6.31$, $P < 0.01$]. *Post-hoc* pairwise comparisons with Games–Howell's test showed that Token test scores in the RTL group were lower than those in the control group (95% confidence interval= -6.17, -0.93; $P < 0.01$) and those in the RNTL group (95% confidence interval= -6.05, -0.66; $P < 0.05$; Figure 1C, Left). Moreover, a significant negative correlation was found between lesion volume and Token test scores in the RTL group ($r_s = -0.84$, $P < 0.01$; Figure 1C, Central); that is, the larger lesion volume, the worse speech comprehension ability of Mandarin Chinese. And no significant correlation was found between lesion volume and Token test scores in the RNTL group ($r_s = 0.27$, $P > 0.05$; Figure 1C, Right).

MMN and LI for the left and right sides of the scalp in lexical and pure tone conditions

MMN waveforms were prominent in both lexical and pure tone conditions as illustrated by sample traces of grand average MMN in response to the lexical-tone contrast and to the pure-tone contrast (Figures 2A,B, left panels). The MMN in the control group under each stimulus condition and that in the RNTL group under the pure-tone contrast were stronger in magnitude when recorded on the right side of the scalp than those recorded on the left side. In contrast, MMN of the RTL group in response to the lexical-tone contrast demonstrated a swapped pattern: it was stronger in magnitude when recorded on the left side of the scalp than that recorded on the right side. More detailed latencies of the MMN between groups under each condition showed no significant difference across groups (Supplementary Figure 2). The right panels of Figures 2A,B show the scalp topographic maps constructed with grand average MMN in response to the lexical-tone contrast and the pure-tone contrast for the control (*Upper*), RTL (*Central*), and RNTL (*Lower*) groups. The MMN topographic maps were obviously lateralized in strength to the right side of the scalp in the control and RNTL groups under each stimulus condition, whereas they were obviously lateralized in strength to the left side of the scalp in the RTL group.

The analysis of MMN amplitudes calculated from four pairs of electrodes on the left (F3, FC1, FC5, C3) and right (F4, FC2, FC6, C4) sides of the scalp in the individual participant (Figure 3A) demonstrates the swapped hemispheric lateralization of the MMN responses to the lexical-tone contrast in the RTL group. Wilcoxon signed-rank test showed that the MMN response was significantly lateralized to the right side of the scalp in the control group for both conditions (lexical tone, $Z = -2.48$, $P < 0.05$; pure tone, $Z = -3.30$, $P < 0.01$) and in the RNTL group for the pure-tone contrast ($Z = -2.20$, $P < 0.05$), whereas the MMN response was significantly lateralized to the left side of the scalp in the RTL group for the lexical-tone contrast ($Z = -2.13$, $P < 0.05$; Figure 3A). No other significant results between bilateral MMNs were shown in the RTL or RNTL group.

The swapped pattern of hemisphere lateralization of MMN responses to the lexical-tone contrast was also revealed by the LI. A one-sample Wilcoxon signed-rank test showed that the LI was significantly less than zero (i.e., lateralized to the right hemisphere) in the control group under each stimulus condition (lexical tone: median value = -0.08, $P < 0.05$; pure tone: median value = -0.12, $P < 0.01$) and in the RNTL patients for the pure-tone contrast (median value = -0.09, $P < 0.05$), whereas the LI was significantly greater than zero (i.e., lateralized to the left hemisphere) in the RTL group under the lexical-tone contrast (median value = 0.24, $P < 0.05$; Figure 3B). The Kruskal–Wallis H test showed significant differences in LI among groups for both conditions (lexical tone: $X^2_{(2)} = 9.67$, $P < 0.01$, with a mean rank LI score of 15.00 for the control group, 28.18 for the RTL group, and 17.00 for the RNTL group; pure tone: $X^2_{(2)} = 9.20$, $P < 0.05$, with a mean rank LI score of 14.07 for the control group, 27.55 for the RTL group, and 18.54 for the RNTL group). *Post-hoc* pairwise comparisons with Dunn's test showed that LIs for both stimulus conditions in the RTL group were larger than those in the control (lexical tone: $P < 0.01$; pure tone: $P < 0.01$) and the RNTL groups (lexical tone: $P < 0.01$; pure tone: $P < 0.05$; Figure 3B). No significant LI difference was observed in either stimulus condition between the control and the RNTL groups.

To explore the causes of the swapped hemisphere dominance, we analyzed the differences in unilateral MMN amplitudes across groups. The Kruskal–Wallis *H*-test showed that the right MMN amplitudes across groups were significantly different for each stimulus condition (lexical tone: $X^2_{(2)} = 9.42$, $P < 0.01$, with a mean rank MMN amplitude score of 17.29 for the control group, 28.00 for the RTL group, and 14.69 for the RNTL group; pure tone: $X^2_{(2)} = 11.76$, $P < 0.01$, with a mean rank MMN amplitude score of 15.71 for the control group, 29.18 for the RTL group, and 15.38 for the RNTL group). *Post-hoc* pairwise comparisons with Dunn's test revealed that the MMN amplitude on the right scalp for each stimulus condition in the RTL group was lower than that in the control (lexical tone: $P < 0.05$; pure tone: $P < 0.01$) and RNTL groups (lexical tone: $P < 0.01$; pure tone: $P < 0.01$). However, MMN amplitudes on the right scalp between the control and RNTL groups were significantly different for neither the lexical-tone contrast ($P > 0.05$) nor the pure-tone contrast ($P > 0.05$). Moreover, MMN amplitudes on the left scalp among groups were significantly

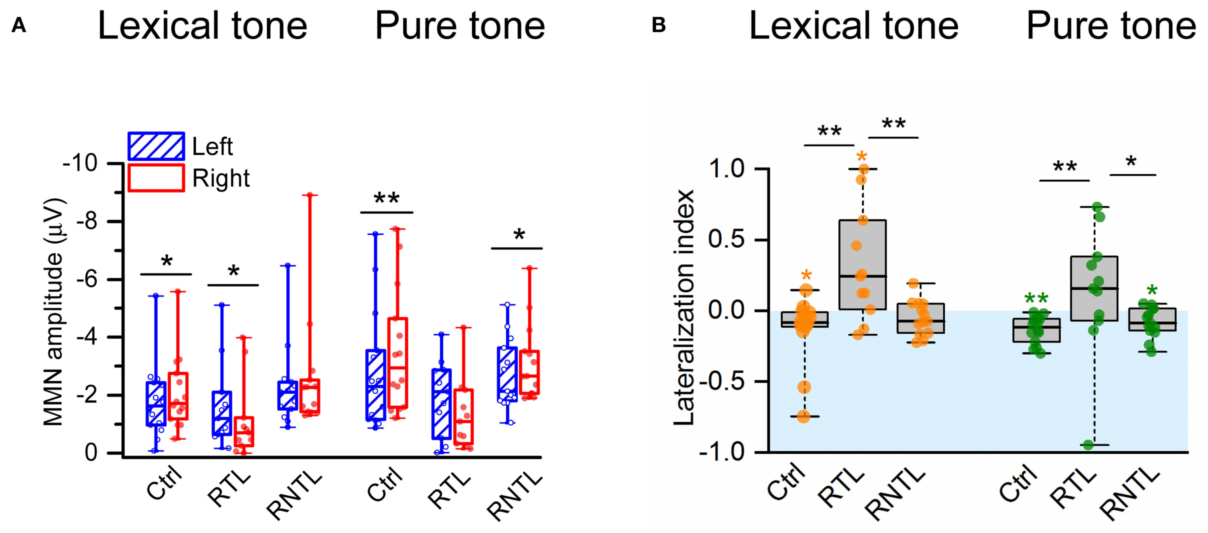


FIGURE 3
MMN amplitudes and lateralization index (LI) were recorded from four electrodes on the left (F3, FC1, FC5, C3) and four electrodes on the right (F4, FC2, FC6, C4) sides of the scalp. **(A)** MMN was significantly larger in amplitude on the right side of the scalp than on the left in the control group for the lexical-tone contrast and the pure-tone contrast and in the RNTL group for the pure-tone contrast but larger in amplitude on the left side of the scalp than on the right in the RTL group for the lexical-tone contrast. **(B)** The comparisons for LI within each group and among groups. The analysis within the group indicates that the LI was significantly less than zero (indicates right hemisphere lateralized response) in the control group for the lexical-tone contrast and the pure-tone contrast and in the RNTL group for the pure-tone contrast but greater than zero (indicates left hemisphere lateralized response) in the RTL group for the lexical-tone contrast. The analysis among groups indicates that the LI in the RTL group was larger than that in the control group and the RNTL group for each stimulus condition. Box plots depict medians with interquartile ranges and whiskers represent the minimum and maximum values. *$P < 0.05$, **$P < 0.01$.

different for neither the lexical-tone contrast ($X^2_{(2)} = 3.20$, $P > 0.05$) nor the pure-tone contrast ($X^2_{(2)} = 1.74$, $P > 0.05$). The reduction in the MMN amplitude in the right hemisphere of the RTL group indicates a swapping of brain dominance.

Correlations between lesion volume and LI and those between LI and token test scores

Our findings showed significant correlations between lesion volume and LI, and between LI and Token test scores in the RTL group. Significant correlation between lesion volume and LI was only found in the RTL group for the lexical-tone contrast ($r_s = 0.72$, $P < 0.05$; Figure 4A, Left), indicating that a larger lesion volume corresponds to less right hemisphere involvement. Significant correlation was also observed between LI and Token test scores in the RTL group for the lexical-tone contrast ($r_s = -0.74$, $P < 0.01$; Figure 4B, Central), suggesting that less right hemisphere involvement corresponds to impaired speech comprehension ability for Mandarin Chinese. No other significant correlations were observed among lesion volume or Token test scores (Figure 4). The detailed results of correlation analysis between the LI for each stimulus condition (lexical-tone contrast and pure-tone contrast) and lesion volume/Token test scores are shown in Supplementary Table 2.

Discussion

In the present study, we investigated how brain damage in stroke patients affects the labor division of the brain hemispheres for auditory processing of linguistic tones. Our study demonstrates that RTL injury results in swapped dominance of brain hemispheres in the preattentive auditory processing of Chinese lexical tones, suggesting that the RTL is a core area for early-stage auditory tonal processing.

Since the early auditory processing of lexical tones was lateralized to the right hemisphere for the control group and the left hemisphere for the RTL group (Figures 2, 3), the findings demonstrate swapped dominance of brain hemispheres in preattentive auditory processing of Chinese lexical tones after RTL stroke. Meta-analyses of lexical-tone processing have suggested significant activations in both temporal lobes in response to lexical tones (Kwok et al., 2017; Liang and Du, 2018), but non-tonal language studies only showed significant activations in the left temporal lobe (Kwok et al., 2017). Lexical-tone studies demonstrate more activations in the RTL than in the left (Liang and Du, 2018). Some researchers suggest that the appearance of language impairments in right-handed patients

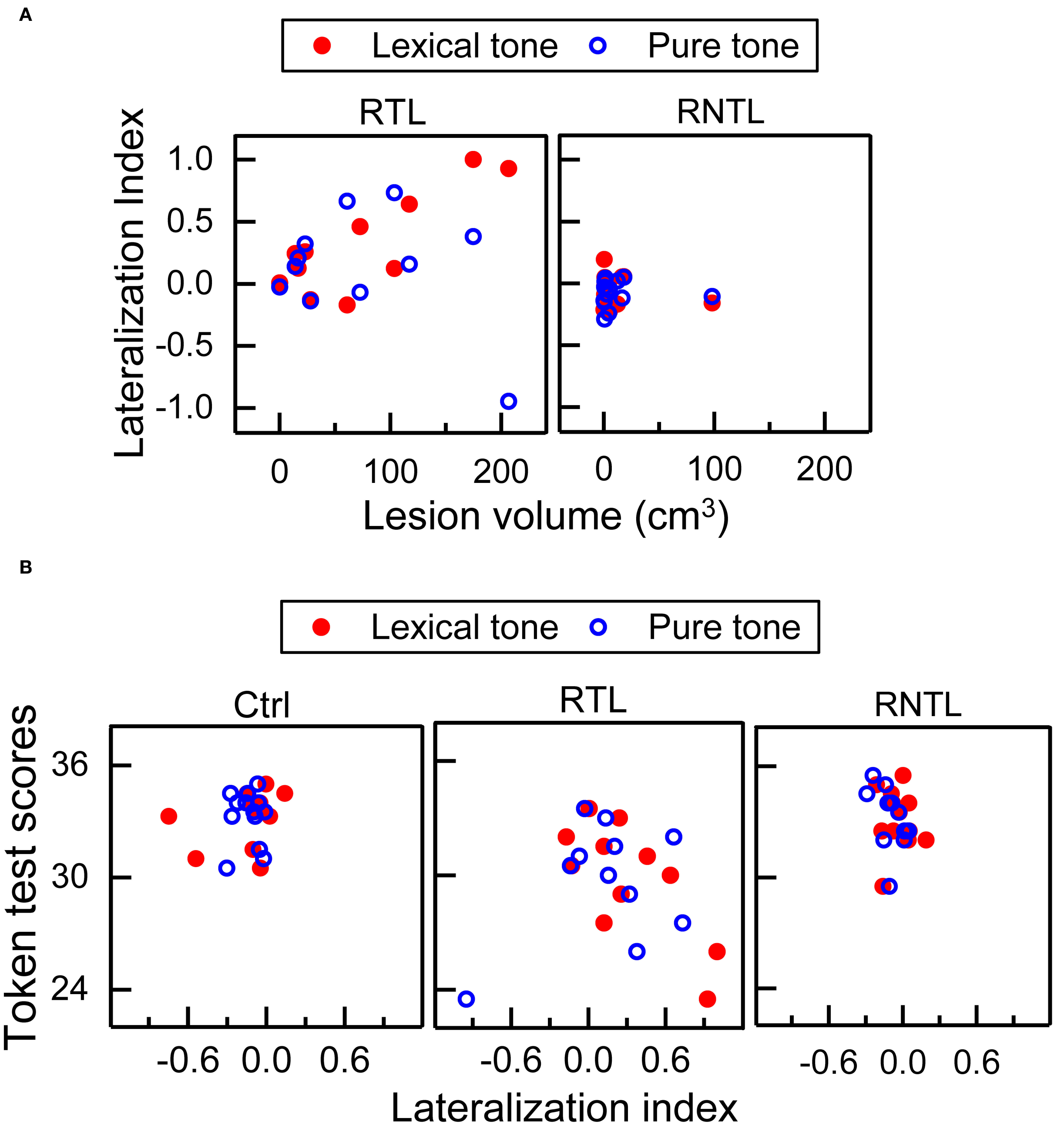


FIGURE 4
Correlations between lesion volume and lateralization index (LI), and those between LI and Token test scores. **(A)** Significant correlation between lesion volume and LI was only found in the RTL group for the lexical-tone contrast—more lesion volume corresponding to less right hemisphere involvement. **(B)** Significant correlation between LI and Token test scores was only found in the RTL group for the lexical-tone contrast—less right hemisphere involvement corresponding to worse speech comprehension ability of Mandarin Chinese.

with right brain injury represents atypical language lateralization before stroke (Gajardo-Vidal et al., 2018). Therefore, the right lateralization in the control group and the impairment of right lateralization in the RTL group would not have been caused by atypical language lateralization. Moreover, the comparisons of LI for the lexical tone and the pure tone within and among groups showed that the RTL group exhibited left lateralization while the control and RNTL groups exhibited right lateralization at a preattentive stage (Figure 3B). Our findings support our hypothesis that RTL injury changes the right hemisphere dominance during the early auditory processing of lexical tones. Notably, both lexical and pure tones reflecting varied spectral information are dominantly processed in the right hemisphere (Zatorre and Belin, 2001; Schonwiesner et al., 2005), which is consistent with the acoustic hypothesis that the hemispheric specialization for processing auditory perception depends on the acoustic structure of the auditory input (Zatorre and Belin, 2001; Schonwiesner et al., 2005). Our results demonstrate that only RTL injury swaps the hemisphere dominance in early auditory processing of lexical tone that carries semantic information. In the RTL group, the reduced right hemisphere involvement for processing lexical tones at a preattentive stage corresponds to the worse speech-comprehension ability for Mandarin Chinese (Figure 4B, Central). Although left hemisphere dominance for processing language has been demonstrated since the 1870s (Broca, 1861; Wernicke, 1874; Tyler et al., 2010, 2011; Teki et al., 2013), the present study suggests that language disorders can occur after injury in the right hemisphere, especially in the RTL.

Although the RTL group exhibited left hemispheric lateralization for processing lexical tones at an early stage, this does not necessarily mean that the left hemisphere compensates for the impaired auditory function of the right hemisphere. The worse speech-comprehension ability for Mandarin Chinese in the RTL group was associated with a reduced right hemisphere-lateralized response for early auditory processing of lexical tones (Figure 4B, Central). This is consistent with a previous study in which the RTL was shown to correlate with speech comprehension (Walenski et al., 2019). Moreover, we found a significant association between lesion volume and LI for the lexical-tone contrast at a preattentive stage in the RTL group: the larger the lesion volume, the lower the right hemisphere-lateralized response (Figure 4A, Left). Similar associations between larger lesion volumes and lower functional improvement have been also found in rats after stroke (Sasaki et al., 2016). Since the RTL group mainly showed lesions in the RTL (Figure 1A), we think that the disappearance of right hemisphere dominance may be caused by the decreased neural activity involved in lexical-tone perception in the right hemisphere. We further suggest that RTL injury impairs the speech comprehension of tonal languages. A previous study demonstrated that the speech comprehension of non-tonal languages can be impaired by right brain damage and that the most frequently impaired language task is auditory sentence-to-picture matching (Gajardo-Vidal et al., 2018).

Notably, in RTL stroke patients, the neural activity revealed by the MMN amplitude on the right scalp was significantly lower than those in the control and RNTL groups, but the neural activity of the left hemisphere in the RTL group was not significantly different. This is in line with the findings of previous studies, which showed that patients with unilateral brain damage may show diminished response on the injured side (Alho et al., 1994; Deouell et al., 2000; Tyler et al., 2011). The swapped hemisphere dominance in early auditory processing of lexical tone in patients with RTL stroke is obviously caused by the decreased MMN amplitude of the right hemisphere. The previous studies indicated that the source generator of MMN originates from the left and right auditory regions (Naatanen et al., 1997; Kujala et al., 2002). Naatanen et al. (1997) demonstrated the source generators of MMN in the left and right auditory cortices for speech and non-speech sounds in their well-known MEG study. If the lesion is in the RTL, then the MMN activity on the right side would be reduced. The more extensive the damage is, the greater the reduction in MMN on the right side. The MMN amplitudes of the left hemisphere (uninjured side) showed no significant difference across groups (Figure 3A). This is inconsistent with the findings of previous studies in which decreased neural activity on the injured side was suggested to result in increased neural activity on the uninjured side in speech perception (Becker and Reinvang, 2007; Tyler et al., 2010, 2011; Teki et al., 2013). Similar findings have been reported in another study showing enhanced interactions between the hemispheres in patients with RTL epilepsy but not in control subjects or patients with RNTL epilepsy (He et al., 2018). Some of the possible reasons to explain why the left MMN amplitude in the RTL group was not increased can be summarized as follows: (i) patients in the RTL group were mostly in the acute stage (post-stroke onset <3 months, Supplementary Table 1), and stable compensation of the left hemisphere for lexical-tone perception may not yet have occurred; (ii) the spectral variation of lexical tones is a basal variation in acoustic patterns, which might be difficult for the left hemisphere to compensate for; and (iii) the lesion volume and the lesion area in RTL stroke patients varies substantially, and it might be difficult to form a stable pattern. Moreover, the decreased MMN amplitude of the right side and unchanged MMN amplitude of the left side in the RTL group may result in the patients not able to discriminate the lexical tones as well as the controls. Future research should add behavioral experiments to explore whether the RTL group has impairments in both MMN amplitude in lexical-tone contrast and the ability to distinguish the lexical tones.

Our present study shows that the RNTL injury also affects the hemisphere dominance in response to the lexical-tone contrast (Figure 3). The RNTL may be the facultative brain regions in early auditory processing of lexical tone, and the RTL may be the obligatory brain region. In addition to the RTL, other brain areas of the right hemisphere, such as the right white matter (Zhao et al., 2016), the right basal ganglia (Chang and Kuo, 2016), and the right visual cortex (Kwok et al., 2015) may also participate in lexical-tone perception. Patients with RNTL injuries mainly showed lesions in the right periventricular white matter, the right basal ganglia, and the right occipital lobe (Supplementary Table 1 and Figure 1B). These results may indicate the importance of the cooperation and connectivity of the multiple brain areas in the early auditory processing of lexical tones, and a special pattern might form when a specific brain region is damaged. Unlike the lexical-tone contrast, the RNTL group still showed right hemisphere dominance in response to the pure-tone contrast (Figures 2, 3). This is consistent with the results of previous studies reporting that the processing of spectral variation for non-speech stimuli mainly occurs in the RTL (Opitz et al., 2002; Schonwiesner et al., 2005). Therefore, early auditory processing of pure tones was not impaired in the RNTL group.

Several limitations should be noted when interpreting our findings. First, the sample size was relatively small. We spent 2 years and recruited 30 stroke patients and 14 healthy control participants. Among these patients, we excluded five stroke patients with severe hearing loss, and one patient withdrew consent. Therefore, the stroke patients were well-characterized. Second, although the difference in age across groups was not significant, participant age showed substantial variation, implying that the sample may not be representative of younger stroke sufferers. Third, since lexical tones and music have the same patterns (Nan and Friederici, 2013; Chen et al., 2018), detailed information about the musical experience of the participants should have been collected. Nevertheless, since the aim of our study was to investigate the effect of right brain injury on the hemispheric dominance of lexical tones, we did not consider the effects of musical training when recruiting subjects. Stroke patients were not always ready for recruitment, and their availability would have been reduced even further if the musical training-related factor had been applied. Fourth, our study lacked exact control between the two stimulus conditions. Because the recorded MMN response in the pure-tone contrast occurred ∼150 ms after the onset of the stimuli under a passive auditory oddball paradigm (Aaltonen et al., 1993), we selected a 200-ms pure tone and a 500-ms inter-stimulus onset interval (ISI) for the pure-tone contrast (Wang et al., 2021). Notably, the MEG data for the neural basis of perceptual processing of lexical tones indicated a left hemispheric dominance for detecting large lexical-tone changes and small deviant contrasts involving less left hemispheric activation in the auditory cortex and greater activation in the right frontal cortex at a later time window (Hsu et al., 2014). The cross-category contrasts also revealed larger MMN responses than within-category contrasts in the left scalp, but not in the right scalp (Xi et al., 2010; Zhang et al., 2011). In addition, an MMN study investigating the effect of allophonic variation on the mental representation and neural processing of lexical tones suggested that activation of the allophonic tonal variants can lead to right-hemisphere-dominant processing of lexical tones, which are otherwise categorically processed via recruitment of both left and right hemispheres (Li and Chen, 2015).

We assessed the speech-comprehension ability for Mandarin Chinese (a tonal language) by the Token test, a language task involving auditory sentence-to-picture matching. The performance in the speech-comprehension task in the RTL group was worse than that in the RNTL and control groups (Figure 1C, Left). In the RTL group, the speech-comprehension ability for Mandarin Chinese negatively correlated with the lesion volume (Figure 1C, Central), indicating a causal role of the RTL in Mandarin speech perception. Considering the growing awareness that aphasia following a stroke can include deficits in other cognitive functions (Schumacher et al., 2019) and the importance of accurately representing lexical-tone information for hearing-impaired Mandarin speakers (Li et al., 2019; Chen et al., 2020), our study highlights the necessity of rehabilitating the language functions of tonal language speakers who suffer from RTL injury and applying formal lexical-tone-related communication tests in clinical assessment and rehabilitation for patients who are speakers of tonal languages and experience brain injury and communication disorders.

To summarize, our findings showed swapped dominance of lateralization from the right to the left hemisphere in patients with RTL injuries but not in those with RNTL injuries, indicating that the RTL is a core area for auditory tonal processing at an early stage or a preattentive stage. These findings indicate the necessity of rehabilitating language functions of tonal language speakers who experience RTL injury.

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 authors.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Committee of the First Affiliated Hospital, University of Science and Technology of China. The patients/participants provided their written informed consent to participate in this study.

Author contributions

LC designed the research. YW and MW performed the research. YW, XL, XG, XW, YQ, and RA analyzed data. LC, XF, HL, BQ, MW, RQ, YW, and XL wrote the paper. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (Grants 81970886, 81570915, and

81870723) and Anhui Provincial Natural Science Foundation (Grant 30973084).

Acknowledgments

Thanks are due to Xiaopeng Han for technical support and to Lihong Huang, Xinyi Lv, and Li Wang for their assistance with recruiting patients.

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/fnins.2022.909796/full#supplementary-material

References

Aaltonen, O., Tuomainen, J., Laine, M., and Niemi, P. (1993). Cortical differences in tonal versus vowel processing as revealed by an ERP component called mismatch negativity (MMN). *Brain Lang.* 44, 139–152. doi: 10.1006/brln.1993.1009

Albouy, P., Benjamin, L., Morillon, B., and Zatorre, R. J. (2020). Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science* 367, 1043. doi: 10.1126/science.aaz3468

Alho, K., Woods, D. L., Algazi, A., Knight, R. T., and Naatanen, R. (1994). Lesions of frontal cortex diminish the auditory mismatch negativity. *Electroencephalogr. Clin. Neurophysiol.* 91, 353–362. doi: 10.1016/0013-4694(94)00173-1

Becker, F., and Reinvang, I. (2007). Mismatch negativity elicited by tones and speech sounds: changed topographical distribution in aphasia. *Brain Lang.* 100, 69–78. doi: 10.1016/j.bandl.2006.09.004

Broca, P. (1861). Sur le siège de la faculté du langage articulé avec deux observations d'aphémie. *Bull. Soc. Anat. Paris* 36, 330–357.

Chang, C. H. C., and Kuo, W. J. (2016). The neural substrates underlying the implementation of phonological rule in lexical tone production: an fMRI study of the tone 3 sandhi phenomenon in Mandarin Chinese. *PLoS ONE* 11. doi: 10.1371/journal.pone.0159835

Chang, H. C., Lee, H. J., Tzeng, O. J., and Kuo, W. J. (2014). Implicit target substitution and sequencing for lexical tone production in Chinese: an FMRI study. *PLoS ONE* 9, e83126. doi: 10.1371/journal.pone.0083126

Chen, A., Peter, V., Wijnen, F., Schnack, H., and Burnham, D. (2018). Are lexical tones musical? Native language's influence on neural response to pitch in different domains. *Brain Lang.* 180–182, 31–41. doi: 10.1016/j.bandl.2018.04.006

Chen, Y., Wong, L. L. N., Qian, J., Kuehnel, V., Christina Voss, S., Chen, F., et al. (2020). The role of lexical tone information in the recognition of mandarin sentences in listeners with hearing aids. *Ear Hear.* 41, 532–538. doi: 10.1097/AUD.0000000000000774

Chobert, J., Francois, C., Velay, J. L., and Besson, M. (2012). Twelve months of active musical training in 8- to 10-year-old children enhances the preattentive processing of syllabic duration and voice onset time. *Cereb Cortex.* 24, 956–967. doi: 10.1093/cercor/bhs377

De Renzi, E., and Faglioni, P. (1978). Normative data and screening power of a shortened version of the Token test. *Cortex* 14, 41–49. doi: 10.1016/S0010-9452(78)80006-9

Deouell, L. Y., Bentin, S., and Soroker, N. (2000). Electrophysiological evidence for an early (pre-attentive) information processing deficit in patients with right hemisphere damage and unilateral neglect. *Brain* 123, 353–365. doi: 10.1093/brain/123.2.353

Doeller, C. F., Opitz, B., Mecklinger, A., Krick, C., Reith, W., Schroger, E., et al. (2003). Prefrontal cortex involvement in preattentive auditory deviance detection: neuroimaging and electrophysiological evidence. *Neuroimage* 20, 1270–1282. doi: 10.1016/S1053-8119(03)00389-6

Dura-Bernal, S., Wennekers, T., and Denham, S. L. (2012). Top-down feedback in an HMAX-like cortical model of object perception based on hierarchical Bayesian networks and belief propagation. *PLoS ONE* 7, e48216. doi: 10.1371/journal.pone.0048216

Gainotti, G., Caltagirone, C., Miceli, G., and Masullo, C. (1981). Selective semantic-lexical impairment of language comprehension in right-brain-damaged patients. *Brain Lang.* 13, 201–211. doi: 10.1016/0093-934X(81)90090-0

Gajardo-Vidal, A., Lorca-Puls, D. L., Hope, T. M. H., Jones, O. P., Seghier, M. L., Prejawa, S., et al. (2018). How right hemisphere damage after stroke can impair speech comprehension. *Brain* 141, 3389–3404. doi: 10.1093/brain/awy270

Gandour, J., Petty, S. H., and Dardarananda, R. (1988). Perception and production of tone in aphasia. *Brain Lang.* 35, 201–240. doi: 10.1016/0093-934X(88)90109-5

Gandour, J., Ponglorpisit, S., Khunadorn, F., Dechongkit, S., Boongird, P., Boonklam, R., et al. (1992). Lexical tones in Thai after unilateral brain damage. *Brain Lang.* 43, 275–307. doi: 10.1016/0093-934X(92)90131-W

Gandour, J., Potisuk, S., Ponglorpisit, S., Dechongkit, S., Khunadorn, F., Boongird, P., et al. (1996). Tonal coarticulation in Thai after unilateral brain damage. *Brain Lang.* 52, 505–535. doi: 10.1006/brln.1996.0027

Ge, J., Peng, G., Lyu, B., Wang, Y., Zhuo, Y., Niu, Z., et al. (2015). Cross-language differences in the brain network subserving intelligible speech. *Proc. Natl. Acad. Sci. USA.* 112, 2972–2977. doi: 10.1073/pnas.1416000112

Hagoort, P., Brown, C. M., and Swaab, T. Y. (1996). Lexical-semantic event-related potential effects in patients with left hemisphere lesions and aphasia, and patients with right hemisphere lesions without aphasia. *Brain* 119, 627–649. doi: 10.1093/brain/119.2.627

He, X., Bassett, D. S., Chaitanya, G., Sperling, M. R., Kozlowski, L., Tracy, J. I., et al. (2018). Disrupted dynamic network reconfiguration of the language system in temporal lobe epilepsy. *Brain* 141, 1375–1389. doi: 10.1093/brain/awy042

Hsu, C. H., Lin, S. K., Hsu, Y. Y., and Lee, C. Y. (2014). The neural generators of the mismatch responses to Mandarin lexical tones: an MEG study. *Brain Res.* 1582, 154–166. doi: 10.1016/j.brainres.2014.07.023

Kadyamusuma, M. R., De Bleser, R., and Mayer, J. (2011). Lexical tone disruption in Shona after brain damage. *Aphasiology* 25, 1239–1260. doi: 10.1080/02687038.2011.590966

Klein, D., Zatorre, R. J., Milner, B., and Zhao, V. (2001). A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers. *Neuroimage* 13, 646–653. doi: 10.1006/nimg.2000.0738

Kujala, A., Alho, K., Valle, S., Sivonen, P., Ilmoniemi, R. J., Alku, P., et al. (2002). Context modulates processing of speech sounds in the right auditory cortex of human subjects. *Neurosci. Lett.* 331, 91–94. doi: 10.1016/S0304-3940(02)00843-1

Kwok, V. P., Wang, T., Chen, S., Yakpo, K., Zhu, L., Fox, P. T., et al. (2015). Neural signatures of lexical tone reading. *Hum. Brain Mapp.* 36, 304–312. doi: 10.1002/hbm.22629

Kwok, V. P. Y., Dan, G., Yakpo, K., Matthews, S., Fox, P. T., Li, P., et al. (2017). A meta-analytic study of the neural systems for auditory processing of lexical tones. *Front. Hum. Neurosci.* 11, 375. doi: 10.3389/fnhum.2017.00375

Lam, N. H., Schoffelen, J. M., Udden, J., Hulten, A., and Hagoort, P. (2016). Neural activity during sentence processing as reflected in theta, alpha, beta, and gamma oscillations. *Neuroimage*. 15, 43–54. doi: 10.1016/j.neuroimage.2016.03.007

Li, N., Wang, S., Wang, X., and Xu, L. (2019). Contributions of lexical tone to Mandarin sentence recognition in hearing-impaired listeners under noisy conditions. *J Acoust Soc Am* 146, EL99. doi: 10.1121/1.5120543

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, B., and Du, Y. (2018). The functional neuroanatomy of lexical tone perception: an activation likelihood estimation meta-analysis. *Front. Neurosci.* 12. doi: 10.3389/fnins.2018.00495

Liberman, A. M., and Whalen, D. H. (2000). On the relation of speech to language. *Trends Cogn. Sci.* 4, 187–196. doi: 10.1016/S1364-6613(00)01471-6

Liu, L., Peng, D., Ding, G., Jin, Z., Zhang, L., Li, K., et al. (2006). Dissociation in the neural basis underlying Chinese tone and vowel production. *Neuroimage* 29, 515–523. doi: 10.1016/j.neuroimage.2005.07.046

Luo, H., Ni, J. T., Li, Z. H., Li, X. O., Zhang, D. R., Zeng, F. G., et al. (2006). Opposite patterns of hemisphere dominance for early auditory processing of lexical tones and consonants. *Proc. Natl. Acad. Sci. USA*. 103, 19558–19563. doi: 10.1073/pnas.0607065104

Mitchell, R. L., and Crow, T. J. (2005). Right hemisphere language functions and schizophrenia: the forgotten hemisphere? *Brain* 128(Pt 5), 963–978. doi: 10.1093/brain/awh466

Naatanen, R., Gaillard, A. W., and Mantysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychol.* 42, 313–329. doi: 10.1016/0001-6918(78)90006-9

Naatanen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., et al. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature* 385, 432–434. doi: 10.1038/385432a0

Naatanen, R., Pakarinen, S., Rinne, T., and Takegata, R. (2004). The mismatch negativity (MMN): towards the optimal paradigm. *Clin. Neurophysiol.* 115, 140–144. doi: 10.1016/j.clinph.2003.04.001

Nan, Y., and Friederici, A. D. (2013). Differential roles of right temporal cortex and broca's area in pitch processing: evidence from music and mandarin. *Hum. Brain Mapp.* 34, 2045–2054. doi: 10.1002/hbm.22046

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113. doi: 10.1016/0028-3932(71)90067-4

Opitz, B., Rinne, T., Mecklinger, A., von Cramon, D. Y., and Schroger, E. (2002). Differential contribution of frontal and temporal cortices to auditory change detection: fMRI and ERP results. *Neuroimage* 15, 167–174. doi: 10.1006/nimg.2001.0970

Picton, T. W., Alain, C., Otten, L., Ritter, W., and Achim, A. (2000). Mismatch negativity: different water in the same river. *Audiol. Neurootol.* 5, 111–139. doi: 10.1159/000013875

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Commun.* 41, 245–255. doi: 10.1016/S0167-6393(02)00107-3

Posner, M. I., and Petersen, S. E. (1989). The attention system of the human brain. *Ann. Neurosci.*503, 686–4939.

Ren, G. Q., Yang, Y., and Li, X. (2009). Early cortical processing of linguistic pitch patterns as revealed by the mismatch negativity. *Neuroscience* 162, 87–95. doi: 10.1016/j.neuroscience.2009.04.021

Rinne, T., Alho, K., Ilmoniemi, R. J., Virtanen, J., and Naatanen, R. (2000). Separate time behaviors of the temporal and frontal mismatch negativity sources. *Neuroimage* 12, 14–19. doi: 10.1006/nimg.2000.0591

Robson, H., Cloutman, L., Keidel, J. L., Sage, K., Drakesmith, M., Welbourne, S., et al. (2014). Mismatch negativity (MMN) reveals inefficient auditory ventral stream function in chronic auditory comprehension impairments. *Cortex* 59, 113–125. doi: 10.1016/j.cortex.2014.07.009

Rorden, C., and Brett, M. (2000). Stereotaxic display of brain lesions. *Behav. Neurol.* 12, 191–200. doi: 10.1155/2000/421719

Sasaki, Y., Sasaki, M., Kataoka-Sasaki, Y., Nakazaki, M., Nagahama, H., Suzuki, J., et al. (2016). Synergic effects of rehabilitation and intravenous infusion of mesenchymal stem cells after stroke in rats. *Phys Ther*. 96, 1791–1798. doi: 10.2522/ptj.20150504

Schonwiesner, M., Rubsamen, R., and von Cramon, D. Y. (2005). Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. *Eur. J. Neurosci.* 22, 1521–1528. doi: 10.1111/j.1460-9568.2005.04315.x

Schumacher, R., Halai, A. D., and Lambon Ralph, M. A. (2019). Assessing and mapping language, attention and executive multidimensional deficits in stroke aphasia. *Brain* 142, 3202–3216. doi: 10.1093/brain/awz258

Seghier, M. L. (2008). Laterality index in functional MRI: methodological issues. *Magn. Reson. Imaging* 26, 594–601. doi: 10.1016/j.mri.2007.10.010

Shankweiler, D., and Studdert-Kennedy, M. (1967). Identification of consonants and vowels presented to left and right ears. *Q. J. Exp. Psychol.* 19, 59–63. doi: 10.1080/14640746708400069

Shankweiler, D., and Studdert-Kennedy, M. (1975). A continuum of lateralization for speech perception? *Brain Lang.* 2, 212–225. doi: 10.1016/S0093-934X(75)80065-4

Shtyrov, Y., Kujala, T., Palva, S., Ilmoniemi, R. J., and Naatanen, R. (2000). Discrimination of speech and of complex nonspeech sounds of different temporal structure in the left and right cerebral hemispheres. *Neuroimage* 12, 657–663. doi: 10.1006/nimg.2000.0646

Si, X., Zhou, W., and Hong, B. (2017). Cooperative cortical network for categorical processing of Chinese lexical tone. *Proc. Natl. Acad. Sci. USA*. 114, 12303–12308. doi: 10.1073/pnas.1710752114

Song, F., Zhan, Y., Ford, J. C., Cai, D. C., Fellows, A. M., Shan, F., et al. (2020). Increased right frontal brain activity during the mandarin hearing-in-noise test. *Front. Neurosci.* 14, 614012. doi: 10.3389/fnins.2020.614012

Teki, S., Barnes, G. R., Penny, W. D., Iverson, P., Woodhead, Z. V., Griffiths, T. D., et al. (2013). The right hemisphere supports but does not replace left hemisphere auditory function in patients with persisting aphasia. *Brain* 136(Pt 6), 1901–1912. doi: 10.1093/brain/awt087

Tyler, L. K., Marslen-Wilson, W. D., Randall, B., Wright, P., Devereux, B. J., Zhuang, J., et al. (2011). Left inferior frontal cortex and syntax: function, structure and behaviour in patients with left hemisphere damage. *Brain* 134(Pt 2), 415–431. doi: 10.1093/brain/awq369

Tyler, L. K., Wright, P., Randall, B., Marslen-Wilson, W. D., and Stamatakis, E. A. (2010). Reorganization of syntactic processing following left-hemisphere brain damage: does right-hemisphere activity preserve function? *Brain* 133, 3396–3408. doi: 10.1093/brain/awq262

Walenski, M., Europa, E., Caplan, D., and Thompson, C. K. (2019). Neural networks for sentence comprehension and production: an ALE-based meta-analysis of neuroimaging studies. *Hum. Brain Mapp.* 40, 2275–2304. doi: 10.1002/hbm.24523

Wang, X. D., Wang, M., and Chen, L. (2013). Hemispheric lateralization for early auditory processing of lexical tones: dependence on pitch level and pitch contour. *Neuropsychologia* 51, 2238–2244. doi: 10.1016/j.neuropsychologia.2013.07.015

Wang, X. D., Xu, H., Yuan, Z., Luo, H., Wang, M., Li, H. W., et al. (2021). Brain hemispheres swap dominance for processing semantically meaningful pitch. *Front. Hum. Neurosci.* 15, 621677. doi: 10.3389/fnhum.2021.621677

Wernicke, C. (1874). *Der aphasische Symptomen complex: Enie Psychologische Studie auf Anatomischer basis*. Breslau: Kohn und Weigert.

Whalen, D. H., and Liberman, A. M. (1987). Speech perception takes precedence over nonspeech perception. *Science* 237, 169–171. doi: 10.1126/science.3603014

Wong, P. C., Parsons, L. M., Martinez, M., and Diehl, R. L. (2004). The role of the insular cortex in pitch pattern perception: the effect of linguistic contexts. *J. Neurosci.* 24, 9153–9160. doi: 10.1523/JNEUROSCI.2225-04.2004

Xi, J., Zhang, L., Shu, H., Zhang, Y., and Li, P. (2010). Categorical perception of lexical tones in Chinese revealed by mismatch negativity. *Neuroscience* 170, 223–231. doi: 10.1016/j.neuroscience.2010.06.077

Zatorre, R. J., and Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex* 11, 946–953. doi: 10.1093/cercor/11.10.946

Zatorre, R. J., Belin, P., and Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends Cogn. Sci.* 6, 37–46. doi: 10.1016/S1364-6613(00)01816-7

Zhang, L., Xi, J., Xu, G., Shu, H., Wang, X., Li, P., et al. (2011). Cortical dynamics of acoustic and phonological processing in speech perception. *PLoS ONE*6, e20963. doi: 10.1371/journal.pone.0020963

Zhao, Y., Chen, X., Zhong, S., Cui, Z., Gong, G., Dong, Q., et al. (2016). Abnormal topological organization of the white matter network in Mandarin speakers with congenital amusia. *Sci. Rep.* 6, 26505. doi: 10.1038/srep26505

Zhou, W. J., Wang, Z. Y., Wang, S. W., and Zhao, L. (2021). Processing neutral tone under the non-attentional condition: a mismatch negativity study. *J. Integr. Neurosci.* 20, 131–136. doi: 10.31083/j.jin.2021.01.301

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 30 September 2022
DOI 10.3389/fpsyg.2022.884102

OPEN ACCESS

EDITED BY
Fang Liu,
University of Reading,
United Kingdom

REVIEWED BY
Albert Lee,
The Education University of Hong Kong,
Hong Kong SAR, China
Jiahong Yuan,
Baidu, United States

*CORRESPONDENCE
Frank Kügler
kuegler@em.uni-frankfurt.de

SPECIALTY SECTION
This article was submitted to
Auditory Cognitive Neuroscience,
a section of the journal
Frontiers in Psychology

RECEIVED 25 February 2022
ACCEPTED 01 August 2022
PUBLISHED 30 September 2022

CITATION
Wang B, Kügler F and Genzel S (2022) The interaction of focus and phrasing with downstep and post-low-bouncing in Mandarin Chinese.
Front. Psychol. 13:884102.
doi: 10.3389/fpsyg.2022.884102

The interaction of focus and phrasing with downstep and post-low-bouncing in Mandarin Chinese

Bei Wang[1], Frank Kügler[2]* and Susanne Genzel[3]

[1]Key Laboratory of Language, Cognition and Computation, School of Foreign Languages, Beijing Institute of Technology, Beijing, China, [2]Department of Linguistics, Goethe University Frankfurt, Frankfurt, Germany, [3]i2x GmbH, Berlin, Germany

L(ow) tone in Mandarin Chinese causes both downstep and post-low-bouncing. Downstep refers to the lowering of a H(igh) tone after a L tone, which is usually measured by comparing the H tones in a "H…HLH…H" sentence with a "H…HHH…H" sentence (*cross-comparison*), investigating whether downstep sets a new pitch register for the scaling of subsequent tones. Post-low-bouncing refers to the raising of a H tone after a focused L tone. The current study investigates how downstep and post-low-bouncing interact with focus and phrasing in Mandarin Chinese. In the experiment, we systematically manipulated (a) the tonal environment by embedding two syllables with either LH or HH tone (syllable X and Y) sentence-medially in the same carrier sentences containing only H tones; (b) boundary strength between X and Y by introducing either a syllable boundary or a phonological phrase boundary; and (c) information structure by either placing a contrastive focus in the HL/HH word (XF), syllable Y (YF), or the sentence-final word (ZF). A wide-focus condition served as the baseline. With systematic control of focus and boundary strength around the L tone, the current study shows that the downstep effect in Mandarin is quite robust, lasting for 3–5 H tones after the L tone, but eventually levelling back again to the register reference line of a H tone. The way how focus and phrasing interact with the downstep effect is unexpected. Firstly, sentence-final focus has no anticipatory effect on shortening the downstep effect; instead, it makes the downstep effect lasts longer as compared to the wide focus condition. Secondly, the downstep effect still shows when the H tone after the L tone is on-focus (YF), in a weaker manner than the wide focus condition, and is overridden by the post-focus-compression. Thirdly, the downstep effect gets greater when the boundary after the L tone is stronger, because the L tone is longer and more likely to be creaky. We further analyzed downstep by measuring the F0 drop between the two H tones surrounding the L tone (*sequential-comparison*). Comparing it with F0 drop in all-H sentences (i.e., declination), it showed that the downstep effect was much greater and more robust than declination. However, creaky voice in the L tone was not the direct cause of downstep. At last, when the L tone was under focus (XF), it caused a post-low-bouncing effect, which is weakened by a phonological phrase boundary. Altogether, the results showed that although intonation is largely controlled by informative functions, the physical-articulatory controls are relatively persistent, varying within the pitch range of 2.5 semitones. Downstep and post-low-bouncing in Mandarin Chinese thus seem to be mainly due to physical-articulatory movement on varying pitch, with the gradual tonal F0 change meeting the requirement of smooth transition across syllables, and avoiding confusion in informative F0 control.

KEYWORDS

downstep, post-low-bouncing, phrasing, focus, intonation, Mandarin Chinese

Introduction

Intonation carries communicative functions, such as focus and phrasing, but much of intonation variation also comes from tonal interactions. To better understand the interaction of tone and intonation, it is important to take into account of both informative and articulatory effects (Xu et al., 2012). In Mandarin, for instance, L tone causes *pre-low-raising, post-low-bouncing,* and *downstep* in the surrounding H tones. Pre-low-raising refers to the pitch raising in the H tone preceding the L tone (Lee et al., 2021). Post-low-bouncing is the phenomenon that F0 of the post-low syllables suddenly goes up first, then drops back gradually, in the condition that the following syllables carry neutral tones or the L tone is under focus (Shen, 1994; Chen and Xu, 2006; Gu and Lee, 2009; Prom-on et al., 2012). Downstep refers to the downtrend of F0 caused by L tones, in the way that the H tones after a L tone is with lower F0 than previous H tones (Xu, 1997, 1999; Shih, 2000; Laniran and Clements, 2003; Connell, 2011). The first two tonal effects have been extensively studied and well explained with articulatory movement of pitch control (Prom-on et al., 2012; Lee et al., 2021).

In this paper, our main goal is to study *downstep* in Mandarin, and its interaction with focus and phrasing. To be more specific, we aim to study how on-focus raising and post-focus compression (PFC) in F0 interact with downstep, and if a phrase boundary terminates downstep. Moreover, considering the presence of global F0 declination in all-H sentences (Shih, 2000; Yuan and Liberman, 2014), we investigate whether downstep and declination share the same pitch lowering mechanism. To go further, a L tone in Mandarin is commonly accompanied by creaky voice (Kuang, 2017, 2018), we thus investigate whether creaky voice may cause downstep. Thirdly, we aim to study the interaction of boundary with post-L-bouncing, which happens when a L tone is focused. Our investigation tackles the question whether a phonological phrase boundary cancels post-L-bouncing or not.

The next section starts with a review of downstep and declination followed by a review on post-low-bouncing, and then focus and phrasing. To better understand pitch from both linguistic and articulatory perspectives, we also briefly introduce a review on laryngeal movement of varying pitch. In the end of section “Background,” the research questions are summarized.

Background

Downstep and declination

There is a global downtrend or declination in a sentence (e.g., Gussenhoven, 2004). Articulatorily, declination is arguably caused by a decrease in subglottal pressure over time (Lieberman, 1967; Collier, 1975; Pierrehumbert 1980; Gelfer et al., 1983). Beside declination, lexical tones and tonal interactions also cause downtrend in F0 contours, e.g., downstep lowers the following H tones. Downstep has long been discussed in African languages (Yoruba (Niger-Congo): Ward, 1952; *cf.* Courtenay, 1971; Luo (Nilotic): Tucker and Creider, 1975; Twi (Akan): Stewart, 1965; Genzel and Kügler, 2011; Kügler 2017; Tswana (Southern Bantu): Zerbian and Kügler, 2015, 2021; among many others). In the African linguistic tradition, downstep is distinguished from downdrift (Stewart, 1965; Hombert, 1974; see Hyman and Leben, 2017 for an overview), terrace (Courtenay, 1971), or automatic and non-automatic downstep (see detailed discussion in Connell, 2001; Rialland and Somé, 2011; Leben, 2014). Strictly speaking, downstep refers to a new register or ceiling established for subsequent H tones after a L tone (Snider, 1990; Snider and van der Hulst, 1993; Connell, 2017; *cf.* Akumbu, 2019). The differentiation between downstep and downdrift, or automatic and non-automatic downstep concerns the fact that in several African languages, both an overtly realized L tone and a floating L tone functions as the trigger of the lowering process. A floating L tone triggers non-automatic downstep, whereas a phonetically realized L tone triggers automatic downstep (Hyman and Leben, 2017). Phonetically, no difference is found between these two types of downstep (e.g., Genzel and Kügler, 2011; Kügler 2017 for Akan). Since there is no floating L tone in Mandarin Chinese, we do not need to differentiate them. We here take the broad definition of downstep as the lowering of F0 after a L tone, following Shih (1988); Xu (1999); Laniran and Clements (2003), and Genzel (2013) among many others, see (1).

1. Downstep: In a HLH tone sequence, the second H is realized with lowered F0 compared to the first H, due to the L tone (*sequential-comparison*). The size of downstep can be paradigmatically calculated as the difference of F0-maximum in the H tones after a L tone and in the corresponding H tones of an all-H tone phrase (*cross-comparison*).

In some West-African languages, downstep initiates a new pitch register to which subsequent tones are scaled, phonologically termed as register tones (e.g., Snider, 1998) or register features (e.g., Akumbu, 2019). In Mandarin, there is no study directly concerning the effect of downstep as setting up a new register tone or register line. We here introduce three studies, which suggest that downstep in Mandarin does not seem to set up a new register tone. First, it showed that several H tones after a L are lowered in F0 as compared to all H-tone sentences, then the pitch gradually reaches the target in the all-H sentence toward the end of the sentence (see **Figure 4**, pp. 66 in Xu, 1999). Second, Gu and Lee (2009) found that the lowering effect in the H tones is greater when the preceding L tone is lower. Third, Wang and Xu (2011) used sentence with HLHL…HL and LHLH…LH tone sequences, the sentence-medial H tones reach roughly the same height as the corresponding H tones in an all-H sentence, which is explained as the balance between pre-Low raising and downstep. Thus, downstep seems to be a tonal feature with gradual change in pitch. A terracing pattern of H tones in the LH sequence—as found in the West-African pattern—does not seem to exist in Mandarin Chinese.

What lacks in previous studies is that how downstep interacts with other informative functions, e.g., prosodic boundary and focus. The first question relates to the domain of downstep. The domain of downstp appears to vary across languages. In Kishamba, morpheme boundaries act as a trigger of downstep (Odden, 1986). In Tswana (Southern Bantu), downstep occurs between prosodic words within a phonological phrase, whereas phonological phrase boundaries block downstep (Zerbian and Kügler, 2015, 2021). In Yoruba, downstep applies across all boundaries within a breath group, which could roughly be interpreted as an intonation phrase (Courtenay, 1971). In Japanese, only an accented word (H*L) within a Major Phrase (MaP) triggers downstep (Pierrehumbert and Beckman, 1988; Selkirk and Tateishi, 1991). The downstep effect in Mandarin as reported in Xu (1999) showed that a phrase boundary does not seem to block downstep, though no systematic data on this issue was provided.

As for the interaction of downstep and focus, we here introduce two studies. Ishihara (2007) studied downstep systematically with sentences in the structure as N1 + N2 + N3 + VP (N and VP are abbreviations of noun and verb phrase respectively). It showed that downstep between N2 and N3 is only partially reset, when N3 is focused and when the syntactic boundary is stronger between N2 and N3. It indicated that downstep is weakened by a strong phrase boundary, and a focused H tone after the L tone. Xu (1999) has shown similar results in Mandarin that the size of downstep seems to be reduced when the H tone after the L tone is focused.

It has been also found that downstep can be canceled in yes/no questions in Hausa (Lindau, 1986), meaning that final F0 raising may counter-balance the downstep effect. In Mandarin, however, it does not seem to be the case as shown in Xu (1999). It requires more systematic analysis on whether sentence final F0 raising interferes with the downstep effect.

As mentioned above, another term easy to be confused with downstep is declination, which refers to the F0 downtrend from the beginning through the end of an utterance. We can see that the crucial difference between declination and downstep is its scope. While declination is a gradual lowering of F0 within an intonation phrase, downstep is a local lowering of F0. Declination has been found in both non-tonal languages ('t Hart & Cohen, 1973; Maeda, 1976; Cooper and Sorensen, 1977; Pierrehumbert, 1979; Sorensen and Cooper, 1980; Cohen et al., 1982; Umeda, 1982; Ladd 1988) and tonal languages (Cantonese: Zhang, 2017; Ge and Li, 2018; Chinese: Xu, 1999; Shih, 2000; Shih and Lu, 2010). Some researchers argue that declination is a fundamental effect in human speech due to a drop in subglottal air pressure (Lieberman, 1967; Collier, 1975; Pierrehumbert, 1979; Gelfer et al., 1983; Gussenhoven, 2004). However, other researchers stated that declination is a combined effect from different functions, e.g., sentence stress and terminal fall (Lieberman and Tseng, 1980; Xu, 1999; Liu and Xu, 2005), topic initial F0 raising (Umeda, 1982; Wang and Xu, 2011) and discourse structure (Hirschberg and Pierrehumbert, 1986; Nakajima and Allen, 1993; Sluijter and Terken, 1993). Downstep and pre-low bouncing, as introduced earlier, also contribute to the overall declination (Liberman and Pierrehumbert, 1984; Pierrehumbert and Beckman, 1988; Shih, 1988; Xu, 1999). Shih (2000) used sentences with the tone sequence of LRH…HN (L, R, H and N stands for low, rising, high and neutral tone respectively), and found that the H tones show declination in the way that the lowering slope is steeper in shorter sentences, after taking apart focus and final lowering. In Shih and Lu (2010) the intonation of an all H tone digital string (338–811-3783) drops from 300 Hz to almost 100 Hz. Similarly, Yuan and Liberman (2014) found that shorter utterances have steeper declination in both the top line and the baseline, after excluding the initial rising and final lowering effects. They are in favor of the idea that declination is linguistically controlled, but not just a by-product of the physics and physiology of talking. It is possible that the declination in the previous three studies still involves some other unknown effects which are hidden by the regression model. In the current study, we calculated declination syllable-by-syllable, as the F0 drop between two adjacent H tones.

Post-low-bouncing

A L tone could also cause F0 raising after it, especially when the following syllables carry the neutral tone, termed as post-low-bouncing (Mandarin and Cantonese: Chao, 1968; Lin and Yan, 1980; Shih, 1988; Chen and Xu, 2006; Gu and Lee, 2009; *cf.* Prom-on et al., 2012). As discussed in Prom-on et al. (2012), post-low-bouncing has been considered mostly as an articulatory phenomenon, limited to the first neutral tone after the low tone. They emphasized that post-low F0 bouncing is different from a carryover effect, although it occurs between tones. The carryover effect shows in the way that the initial F0 of a syllable is heavily assimilated to the final F0 of the preceding tone, but over the course of the current syllable, F0 gradually approaches its own tonal target. To account for such assimilatory effect, Xu and Wang (2001) proposed the Target Approximation model, which represents the production of successive tones as a process of asymptotically approaching each tonal target within the time interval of the respective syllable, starting from the offset F0 of the preceding syllable. Post-low-bouncing, instead, is the process that pitch increases first then drops back to the underlying target. They discussed the possible physical mechanism behind the low-bouncing effect and suggested a *balance-perturbation hypothesis*. In simple words, after producing a very low F0, the extrinsic laryngeal muscles, especially the sternohyoids (Ohala, 1972; Atkinson, 1978), stop contracting and thus temporarily tip the balance between the two antagonistic forces maintained by the intrinsic laryngeal muscles, resulting in a sudden increase of the vocal fold tension (Prom-on et al., 2012, pp. 422). It still requires articulatory studies to verify the *balance-perturbation hypothesis*. From pitch analysis, one way to test it is to vary syllable duration in the L tone. A longer L tone may reduce post-low-bouncing as it gives more time for the muscles releasing the force. We hence predict that post-low-bouncing is weakened if the L tone is at a phrase boundary, as the L tone is with final lengthening.

Focus and phrasing

There has been extensive research on how focus is realized prosodically in many languages (for an overview see Kügler and Calhoun, 2020). In Mandarin, focus is realized by increasing the pitch range, intensity, duration and articulatory fullness of the focused word, and reducing the F_0 and intensity of the following words (post-focus-compression, PFC), while leaving the pre-focus words largely unchanged (Xu, 1999; Chen and Gussenhoven, 2008; Wang and Xu, 2011). Although Mandarin is tonal, its prosodic focus pattern is very similar to English (Cooper et al., 1985; de Jong, 1995; Xu and Xu, 2005), German (Féry and Kügler, 2008) and many other Indo-European languages (Xu et al., 2012). A recent study found that PFC can go across a relative strong prosodic boundary in Mandarin (e.g., a boundary between to clauses), indicating that phrasing does not interfere with post-focus constituents (Wang et al., 2018b). In other words, focus and phrasing are largely encoded in parallel in intonation, though focus may cause prosodic boundaries in some languages (e.g., Kügler and Calhoun, 2020).

Prosodic boundaries are generally indicated by different phonetic cues such as pre-boundary lengthening, silent pause, F0 reset, phonological boundary tones and changes in voice quality (for detailed discussion, see Wang et al., 2018b). In Mandarin Chinese, boundary strength is realized with gradient means rather than categorical ones, differentiated mainly in pre-boundary lengthening and optional silent pause, but not F0 (Xu and Wang, 2009; Wang et al., 2018b). Although pitch reset has been found at a strong boundary (Dutch: de Pijper and Sandeman, 1994; Swerts, 1997; English: Ladd, 1988), F0 plays a limited role to distinguish boundary strength in Mandarin Chinese when tones and focus are carefully controlled (Xu and Wang, 2009; Wang et al., 2018b). Minimum F0 is lowered at a strong boundary with a silent pause for about 200 ms, but not at a phrase boundary within a sentence (Wang et al., 2018b).

It has been found in many languages that pre-boundary syllables are longer than non-final syllables (e.g., English: Byrd, 2000; Finnish: Nakai et al., 2009; Dutch: Swerts and Geluykens, 1994), and articulatory gestures have slower velocity (Krivokapic and Byrd 2012). The prolonged syllable might give rise to fully realized phonetic targets (Lindblom, 1990; DiCanio et al., 2021), e.g., tones in Mandarin (see Figure 4 in Wang et al., 2018b, p. 36). Phrase-final tones carry both lexical tone and post-lexical tone, e.g., a pitch accent and a boundary tone (Arvaniti and Fletcher, 2020). On the other hand, phrase-final position may also be the locus of glottalization (Huffman, 2005), devoicing (Wagner, 2002), and a gradual decay in intensity and F0 (Gussenhoven, 2004; Ladd, 2008; *cf.* DiCanio et al., 2021).

Relating to the current study, we aim to find out whether a phonological phrase boundary reduces or even blocks downstep and post-low-bouncing effect, assuming that the pre-boundary syllable carrying a L tone is longer and hence the tone is fully realized with pitch raising toward the end of the syllable, since fall-rise is the citation form of the L tone in Mandarin. Another possibility is that pitch goes lower when the L tone is longer, thus makes a greater downstep effect.

Laryngeal movement of varying pitch

After introducing the studies on the linguistic meaning of tone and intonation, we here would like to go back to articulatory studies on pitch control. It will help us to understand downstep and post-low-bouncing, since down to the bottom of the questions raised above, it is all about how the muscles, bones, vocal folds and brain cooperate to realize the pitch targets. The observed pitch contours reflect both linguistic meanings and articulatory constrains.

Yuan and Liberman (2014) discussed articulatory studies on how F0 is controlled. We here just briefly cite some most relevant studies. F0 is determined by the stiffness and effective mass of the vocal folds and the subglottal air pressure (Murry, 1971; Hollien, 1974, 1983; Baer, 1979; Titze, 1988; Stevens, 2000; Zhang, 2016). Intrinsic laryngeal muscles, especially the cricothyroid muscle (CT), are the main contributor to the adjustment of the stiffness and effective mass of the vocal folds. The contraction of CT raises F0; the relaxation of CT, along with the activity of other laryngeal muscles, lowers F0 (Collier, 1975; Atkinson, 1978). Extrinsic laryngeal muscles, which suspend and support the larynx, can also change the states of the vocal folds through vertical larynx movement (Ohala 1972; Honda 1995; Hirose 1997), and F0 falls as the larynx moves down.

F0 lowering is not only accompanied by larynx lowering, relating to extrinsic laryngeal muscles (Honda, 1995; Hirose,

1997) but also involves the joint supraglottal action (Lindqvist-Gauffin, 1969, 1972; *cf.* Lindblom, 2009). In Mandarin, the basic role of larynx height in the execution of tone is complicated by the relationship of larynx height to the state of the larynx: constriction of the supra-glottal laryngeal structures is facilitated by raising the larynx (Edmondson and Esling, 2006) and inhibited by lowering the larynx (Moisik et al., 2014; Moisik and Esling, 2014). Moisik et al. (2014) shows that a L tone target can be reached either by lowering the larynx, or by combining the raise of larynx height and laryngeal constriction, which may lead to creakiness in the low tone. They show that producing the H tone requires any tone involving lowering in pitch is easily becoming creaky, especially the L tone (Kuang, 2017, 2018).

Ladefoged (1973, p. 75) suggested that the creaky voice phonation mechanism is that "because the arytenoid cartilages move forward as they come together; the vocal cords tend to be less stretched in creaky voiced sounds; they are therefore likely to vibrate at a lower frequency. But the coming together of the arytenoids and the movements of the thyroid cartilage that stretch the vocal cords are independent laryngeal gestures, so that it is quite possible for creaky voiced sounds to occur on any pitch." Creaky voice in Mandarin L tone exhibits various laryngealization properties in acoustic waveforms, including aperiodicity, period doubling, or low-frequency pulse-like vibratory patterns (Gerratt and Kreiman, 2001; Keating et al., 2015). In Mandarin, creaky voice relates to the low target in pitch that the L tones are less creaky when the pitch range is raised, but creakier when the pitch range is lowered (Kuang, 2017, 2018). In previous studies on downstep and post-low-bouncing, creaky voice is usually not taken into account. A consequence of this discussion leads to the question whether creakiness causes downstep or not.

Research questions and hypotheses

The main goal of the current study is to understand the property of downstep in Mandarin. The second goal is to provide some analysis on how post-low-bouncing interacts with boundary strength. These will lead us to better understand how intonation is shaped by both informative functions and articulatory constrains. The research questions and hypotheses are summarized as the following.

1. How do focus and boundary interact with downstep? We divide this question into 6 sub-questions.

Q1: Does downstep set up a new register tone?

According to Xu (1999), we predict that downstep effect lasts for several syllables and approach the all-H reference line gradually in wide focus condition.

Q2: Does a sentence-final focus terminates downstep?

We predict that the answer is no because downstep is presumably local, and pitch target is realized syllable-by-syllable as stated in PENTA model (Xu et al., 2022).

Q3: Is downstep eliminated by on-focus F0 raising and post-focus-compression?

We predict that informative functions of intonation may override an articulatory effect.

Q4: How does a phonological phrase boundary interact with downstep?

Given that pre-boundary L is lengthened, the tonal target is expected to be fully realized, and in turn, that may lead to greater downstep effect, since the L tone is lower or even being creaky.

Q5: Do declination and downstep share the same mechanism?

The answer to this question actually depends on how to measure declination and downstep. It also remains controversial whether there is any separate articulatory mechanism controlling declination. Our prediction is that downstep and declination may come from different articulatory control, since downstep is local whereas declination is global.

Q6: Is creaky voice the cause of downstep?

Downstep is caused by a L tone, which is usually creaky in Mandarin (Kuang, 2017). It is possible that creaky voice is the main cause of downstep.

2. When a L tone is under focus, post-low-bouncing is expected. Does a phrase boundary block post-low-bouncing (Q7)?

According to *balance-perturbation hypothesis* (Prom-on et al., 2012), we predict that post-low-bouncing is weakened if the L tone is at a phrase boundary.

Materials and methods

The experiment aimed to study the size and scope of downstep and post-low-bouncing in Mandarin Chinese, concerning its interaction with focus and phrasing. The size of downstep and post-low-bouncing effects was measured by comparing sentences with all H tones and a comparable sentence with a L tone inserted at the target position, while keeping the rest of the two sentences exactly the same. In this way, we can test whether downstep sets up a new pitch register, as taken the all-H sentence for reference. We named it as *cross-comparison* to answer Q1-Q4. Besides, we also calculated the

F0 difference between the two H tones surrounds the L tone, and compared it with the F0 lowering in all-H sentences. We named it as *sequential-comparison* to answer Q5. Thus, the property of downstep and declination can be compared. Moreover, downstep effect caused by creaky and normal L tones were compared, to answer Q6. Post-low-bouncing only occured in the condition of the L tone being focused, thus focus condition is fixed. Only the boundary after the L tone was varied to test whether a strong boundary ends post-low-bouncing (Q7).

Reading materials

The carrier sentences contained only H tones, except for a neutral tone at sentence-final position. Two target words were embedded in the middle of the carrier sentence, one consisted of a LH word (named as syllable X and Y) triggering downstep and post-low-bouncing, and the other one consisted of a HH word, serving as the reference. The two sentences of each item were read in varied contexts eliciting 4 different focus, and 2 boundary conditions.

Three variables were independently manipulated in this experiment, that is, tone of syllable X (either H or L tone), boundary strength between syllable X and Y (syllable boundary or phrase boundary) and focus type (wide focus (WF), focus on syllable X (XF), on syllable Y (YF) and in sentence final position (ZF)). One set of the sentences in the condition of syllable and phrase boundary were provided in (1a) and (1b). Each sentence was with the syntactic structure as S-V1-O1-V2-O2, and the target words (syllables X and Y) were put in the O1 and V2 position, respectively. Here, by comparing the F0 of syllable Y and that of the following H tones between the two sentences (LH and HH), we can calculate the effect size and scope of downstep. The statistical analysis will then test for how many syllables after the L tone the downstep effect lasts, with consideration of boundary and focus conditions.

For the two boundary conditions, a monosyllabic homophone of the target syllable Y was used to construct sentences with different syntactic boundaries. Based on the assumption of the syntax-phonology interface, prosodic boundaries, in particular in this experimental setting, are the result of matching syntactic constituents onto prosodic constituents (Selkirk, 2011). Thus, in the syllable boundary condition (1a), the HL_XH_Y was one word, whereas in the phrase boundary condition (1b), the HL_X was a word, and the following H_Y was an adverb, phrased together with the following words as a verb phrase (VP). Thus, prosodic boundary in condition (1a) was weaker than that in (1b), named as a syllable boundary (SylB) and a phrase boundary (PhrB) respectively. In example (1a), the HHH sequence (yin1ou1dou1樱欧兜, Ying1ou1 bag1) meant a bag printed with ying1ou1 (a make-up word for an exotic plant), whereas in (1b), 'dou1' in the HHH sequence (ying1ou1.dou1樱欧都, Ying1ou1 all1) was an adverb, meant "all" to modify the following verb "lingchu (take-out)." In this way, the two boundary conditions were clearly distinguished by using two different characters (兜 vs. 都, bag vs. all). It was the same construction for the HLH sequence, in which the HL tone word is ying1ou3 (樱藕Ying1ou3), which was also a make-up word for an exotic plant. Here, the contrast of syllable X, either being L or H toned, was straightforward by using the two different characters (藕vs 欧, ou3 vs. ou1). In this way, no specific explanation of the material was necessary for the speakers. They were easily able to read the sentences with different tones and phrasing conditions in a natural way.

Focus was elicited by varying a preceding background sentence, which required a correction of the corresponding word in the target sentence. Taken the HL_XH_Y sentence in the syllable boundary condition (see 1a), the four focus conditions are presented in (2). Here, the H tone (syllable Y) is critical to test the

(1a) *Syllable* boundary (SylB) between syllable X (ou, L 藕 or H 欧) and Y(dou 兜 H):

汪英 清出 樱**藕/欧 兜** 拎 出 公司 车 间 了
(Wang1ying1 qing1chu1 (ying1**ou1/3 dou1)**ω)φ ling1chu1 gong1si1 che1jian1 .le0.
1 2 3 4 5 6 7 8 9 10 11 12 13 14
H H H H H H/L H H H H H H H N
X Y Z
wangying find (yingou bag) take-out company workshop ASP
'Wangying found an **Yingou bag** and took it out of the workshop of the company'.

(1b) *Phrase* boundary (PhrB) between syllable X (ou, L 藕 or H 欧) and Y(dou 都 H)

汪英 清出 樱**藕/欧** **都** 拎 出 公司 车 间 了
(Wang1ying1 qing1chu1(ying1**ou1/3)**ω)φ (**dou1**ling1chu1)ω gong1si1 che1jian1 le0.
1 2 3 4 5 6 7 8 9 10 11 12 13 14
H H H H H H/L H H H H H H H N
X Y Z
Wangying find yingou all take-out company factory ASP
'Wangying has found **Yingou** and took **all** of them out of the factory of the company'.

effects of downstep and post-low-bouncing, and their interaction with focus. Thus, syllable Y was manipulated as either post-focus (focus on syllable X), on-focus (focus on syllable Y) or pre-focus (focus on syllable Z). A wide focus condition served as the baseline. Similar contexts were constructed for the other sentences, see Appendix I for the whole sentence sets.

The background sentences of the four focus conditions for the sentence (1a) are as follows.

Wide focus: "ni3 ting1shuo1 le0 ma0?" (Have you heard about it?)

X-focus: "bu2shi4ying1an1" (It is not "Yingan.")

Y-focus: "bu2shi4bao1" (It is not the tote.)

Z-focus: "bu2shi4lou2dao4" (It is not the corridor.)

We constructed two sets of items. In total, 2 (tone of syllable X) × 2 (boundary between X and Y) × 4 (focus) × 2 (sets) × 3 (repetitions) × 8 (speaker) = 768 sentences were analyzed.

Speakers

Eight native Mandarin speakers participated in the experiment at Minzu University of China (5 female and 3 male speakers), from the age of 20 to 28. They were born and brought up in Beijing, spoke no other Chinese dialects and reported no hearing or speaking impairments. They were paid with small amount of money for taking part in the experiment.

Recording procedure

The subjects were recorded individually in the speech lab at Minzu University of China. They were asked to read aloud both the context and the target sentences at a normal speed and in a natural way. They sat before a computer monitor, on which the test sentences were displayed, using AudiRec, a custom-written recording program. To make the reading task a little easier for the speakers, the focused words were highlighted with color. A Shure 58 Microphone was placed about 10 cm in front of the speaker. All sentences were digitized directly into a Thinkpad computer and saved as WAV files. The sampling rate was 48 KHz and the sampling format was one channel 6-bit linear. Each speaker repeated the whole set of sentences 3 times in different random order, with about 5 minutes break between sessions. Before the formal recording, they read the sentences silently to get familiar with them, and to make sure that they understood the meaning. The total recording time was about an hour.

Acoustic measurements and statistical methods

The target sentences were extracted and saved as separate WAV files. ProsodyPro (Xu, 2013) running under Praat (Boersma and Weenink, 2013–2022), was used to take F0 and duration of each syllable measurements from the target sentences, which were all segmented into syllables manually, and at the same time hand-checked vocal cycles markings generated for errors, such as double-marking and period skipping. ProsodyPro then generated syllable-by-syllable F0 contours that were either time-normalized or in the original time scale. At the same time, the script extracted various measurements, including maximum F0, minimum F0 and duration of each syllable. We could measure F0 at the offset of a syllable, however maximum F0 is toward the very end of the syllable (see Figure 5), it is highly probable that the two values are with very little difference. Maximum F0 is much more widely applied in previous studies (e.g., Xu, 1999; Genzel and Kügler, 2011; Prom-on et al., 2012). Thus, we choose maximum F0 to measure downstep effect.

The statistic tests were carried out in the R environment (R Core Team, 2016) by using lme4 package Version 1.1–18 (Bates, et al., 2016) to estimate the effect of the fixed factors (tone, boundary and focus) and the random factors (speaker and sentence set) on the acoustic parameters, e.g., maximum F0 and duration. Regression coefficients (bs), standard errors (SEs) and *t*-values (t = b/SE) are reported, taken t > 2.0 as reaching the significant level at $p < 0.05$ (Gelman and Hill, 2006). In the results, we reported the best-fit model according to the model comparisons with the lowest AIC and BIC. For the fixed factors, we took the model with interaction only when there was significant interaction.

Creaky L tone was visually identified by checking the spectrum and the WAV files. Zhang (2016) distinguished four types of creaky voice (see Figure 3 in that paper). We grouped all these types as creaky voice. Since F0 is the main concern in this paper, we here labeled the part with aperiodic pulses as creaky, see Figure 1. In this way, the part of the regular pulses was used to get the F0 values of the syllable. Most of the creaky L tone was similar to what Figure 1 shows.

Results

In this section, the graphic analysis firstly shows how focus and phrasing are realized in intonation (Figures 2, 3), followed by quantitative analysis of F0 and duration (Figure 4 and Table 1).

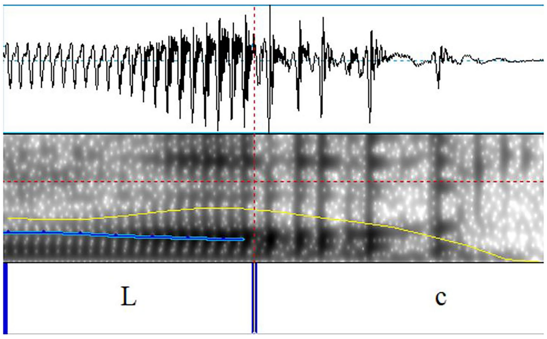


FIGURE 1
An example of the syllable with creaky L tone. Here, L and c stand for the part with periodic and aperiodic pulses.

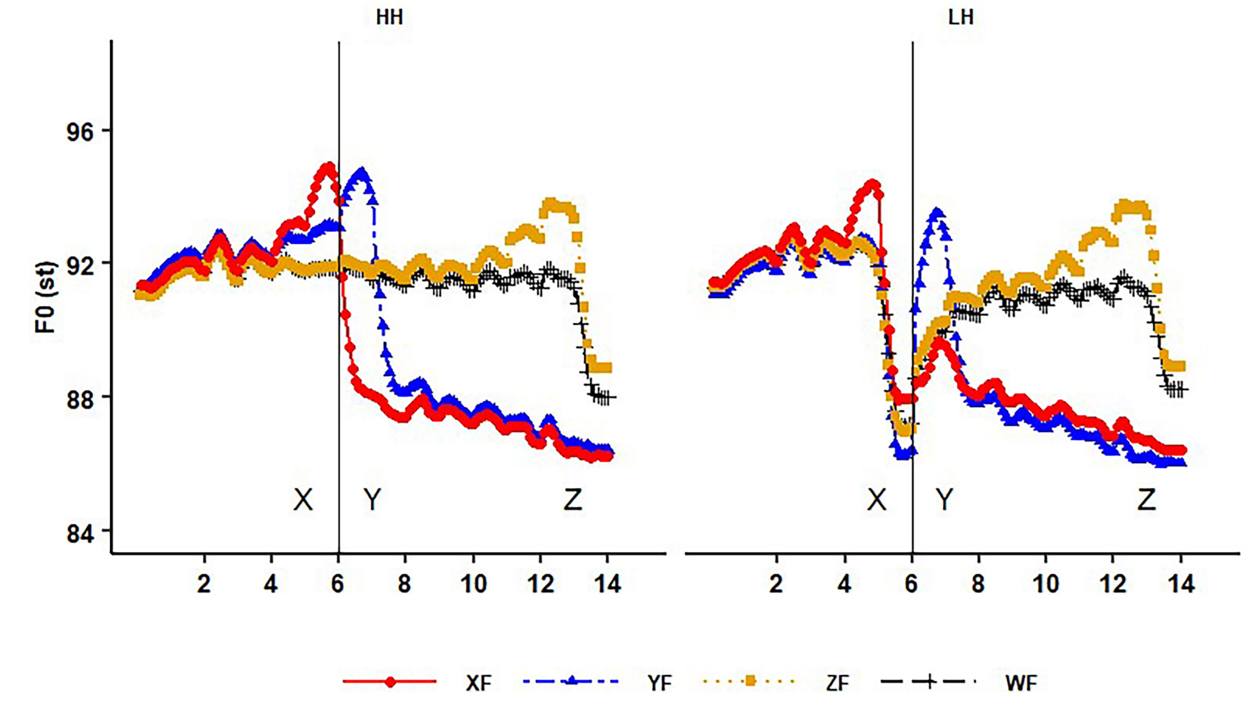


FIGURE 2
Time-normalized intonation contours of the HH (left) and LH (right) sentences in the conditions of the *syllable* boundary (between syllable X and Y), with the four focus conditions overlaid in one figure. Here XF, YF, ZF and WF stand for focus in word X, Y, Z and the wide focus condition. The *x*-axis are the syllable numbers. The vertical line indicates the critical boundary between X and Y.

These two sections serve to confirm that our results are largely consistent with previous studies on focus and phrasing, so that we are confident to further analyze their interaction with the tonal manipulation on intonation. To get an overview of the results, downstep and post-low-bouncing are firstly visually analyzed with intonation contours (Figure 5). Downstep is then quantitatively analyzed with two different methods, i.e., (a) the *cross- comparison* between LH and HH sentences to verify how many syllables it takes for the H tones after the L tone reaching the all-H sentences to answer Q1-Q4 (Figure 6 and Table 2); (b) the *sequential-comparison* between the H tones surrounding the L tone. By comparing the decrease of the H tones in the LH and HH sentence, we can tear apart the declination and the downstep effect to answer Q5 (Figures 7, 8 and Table 3). Thirdly, we noticed that L tones are mostly creaky, especially in the phrase-boundary condition. Therefore, we aim at answering the question whether the change of phonation type to creaky voice is a cause on downstep. We then analyzed the pitch height in the H tone after the L tone as compared between the creaky and normal L tones to answer Q6 (Figure 9). We can show that the change of phonation type is not the direct cause of downstep. For post-low-bouncing, it only happens when the L tone is focused (XF). In line with the findings in Prom-on et al. (2012), we here provide further analysis on its interaction with boundary strength to answer Q7 (Figure 10). With this analysis, we can justify that the *balance-perturbation hypothesis* holds, which predicts weaker post-low-bouncing when the L tone is longer.

Graphic analysis on focus and phrasing

First, we present intonation contours to show how focus is encoded in intonation. Figure 2 presents the HH and LH sentences in the condition of syllable boundary, with the 4 focus conditions overlaid in one figure. In each sentence, 10 time-normalized F0 points for each syllable were averaged across 48 observations (8 speakers × 2 sets × 3 repetitions).

We can see in Figure 2 that focus is realized as the tri-zone pattern as defined in Xu (1999) and repetitively found in many other studies (e.g., Wang et al., 2018b). Looking at the HH tone sentences, we can clearly see that the on-focus syllables show raised F0 and expanded pitch range; the post-focus words exhibit lowered and compressed pitch; while the pre-focus words are similar to the wide focus condition. It holds in the LH sentences as well, except that when the L tone word (e.g., ying1ou3) is focused (XF), the pre-low H is raised. And in the YF condition, on-focus F0 raising still applies in the H tone after the L tone. Thus, downstep does not override (or cancel) on-focus F0 raising. The sentences in the phrase boundary show a very similar pattern, which is not presented here for the interest of space. A phrase boundary does not block post-focus F0 compression (PFC), as likewise reported in Wang et al. (2018b). In general, it confirms that tonal interactions and phrasing do not change how focus is realized, though the amount of focal raising appears to differ between tone conditions.

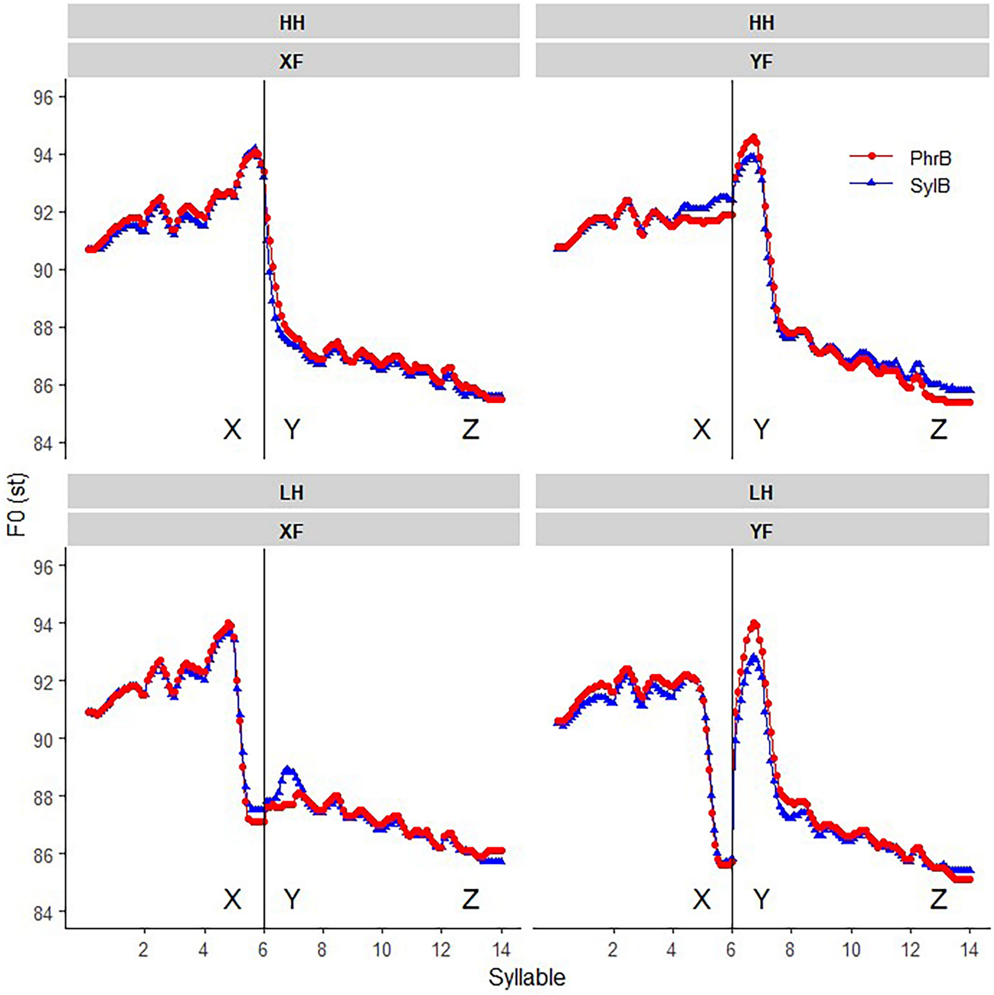


FIGURE 3
The time-normalized intonation contours of the two boundary conditions in the HH and the LH sentences under the XF and YF conditions. Here SylB and PhrB stand for syllable and phrase boundary between syllable X and Y. The *x*-axis are the syllable numbers.

Secondly, Figure 3 presents how boundary strength is encoded in intonation in the XF and YF conditions. No clear difference in F0 between the two boundary conditions can be seen here, in both the HH and LH (lower row) sentences. In WF and ZF conditions, the two boundary conditions do not show clear difference either, which is not presented here for the interest of space. It is in consistence with Wang et al. (2018b) that F0 plays a limited role on phrasing, especially on boundaries within a sentence. Importantly, when pre- and post-boundary syllables are under focus (the X and Y focus condition), there is still no clear sign of using F0 to mark boundary strength. Thus, focus in Mandarin does not seem to invulnerably insert a prosodic boundary.

The above graphic observations show that F0 variation is mainly triggered by focus and tone, but not by prosodic boundaries. We further analyzed the nature of the boundary and whether speakers distinguished the two boundary conditions phonetically. The following analysis of syllable duration (see section "Acoustic analysis on the interaction of focus and boundary," Figure 4) confirms that boundary strength was encoded mainly in pre-boundary lengthening, but not F0.

Acoustic analysis on the interaction of focus and boundary

From the graphic analysis (see Figures 2, 3), we can see that the intonation patterns of focus and phrasing are consistent with previous studies, e.g., Xu (1999) and Wang et al. (2018b). Since focus and boundary effects have already been extensively studied, statistical analysis on all the syllables is not presented here. Statistic test on syllable X is of particular interest as it interacts with downstep and post-low-bouncing. To better understand the interaction between boundary and focus on syllable X, we present the boxplot of maximum F0 and duration of syllable X in Figure 4.

Linear-mixed-models on maximum F0 and duration in syllable X were carried out in HH and LH sentences separately,

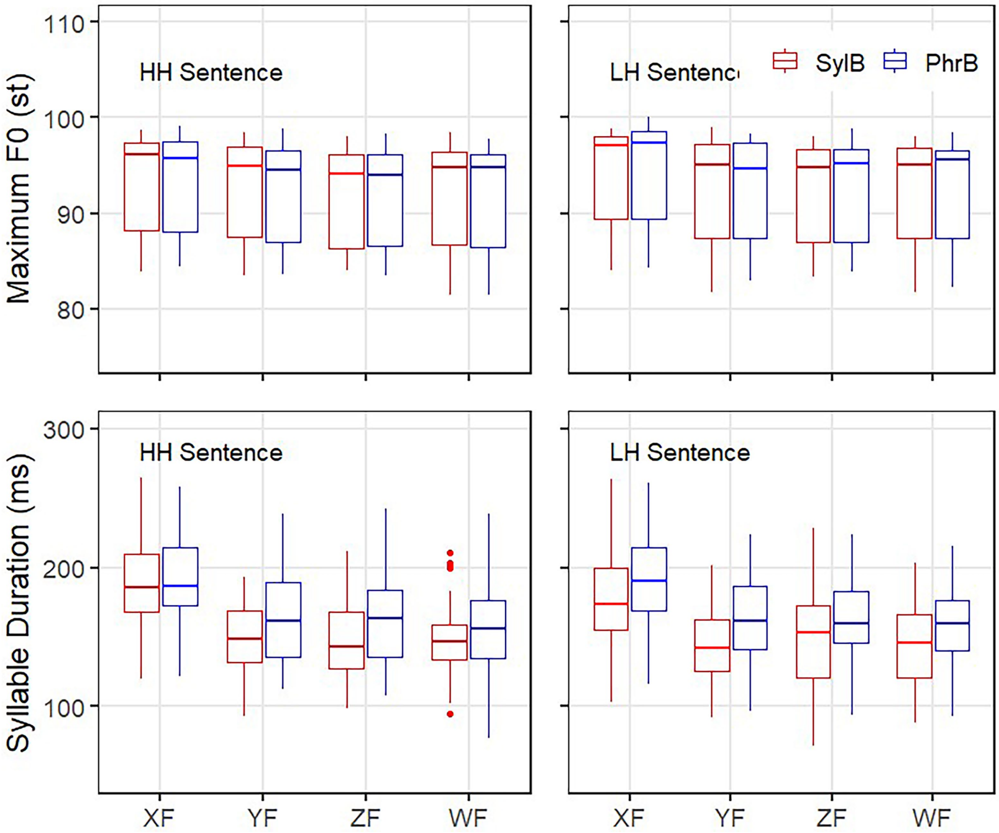


FIGURE 4
Maximum F0 and duration of syllable X in the H_XH_Y(left) and L_XH_Y(right) sentences in the two boundary conditions (SylB and PhrB) and the four focus conditions.

with focus and boundary as two non-interactive fixed factors, while speaker and sentence set as random factors (see Table 1). Wide-focus and syllable-boundary were set as the baseline conditions. The LMM model was chosen to meet the criteria that (1) the model with presumed interaction did not show significant interactions, thus we took this model without interaction; and (2) it was with the lowest AIC and BIC while we tried different ways of setting the random effects.

As for focus effect on syllable X, together with the observations in Figure 4, the statistical analysis in Table 1 shows that focus significantly increases both maximum F0 (about 2.8 st) and duration (about 66 ms) of syllable X (see the line of XF in Table 1). In the Y focus condition, the 3 syllables HXY (H means the high tone before syllable X, e.g., ying1ou1dou1 'Yingou bag') is possibly grouped as one prosodic word, thus syllable X is also with increased maximum F0 (about 1.4 st) and duration (about 24 ms; see the line of YF) in Table 1, which is in consistent with the findings in Chen (2006) on the durational domain of focus.

As for boundary effect on syllable X, the data in Table 1 (also see Figures 3, 4) show that boundary does not have any effect in maximum F0 (92.7 st vs. 92.6 st), but only in duration of syllable X (192 ms vs. 212 ms). No interaction was found between focus and boundary in the duration of syllable X, meaning that the pre-boundary lengthening applies to roughly the same degree in all the focus conditions (see Figure 4), which is in consistent with Wang et al. (2018b). The above results hold for both HH and LH sentences. It leads us to conclude that focus and tone do not interfere with pre-boundary lengthening. Thus, durational adjustment due to focus, boundary and tone is also largely encoded in parallel. We can then further test whether the lengthened L tone decreases or increases the level of downstep and post-low-bouncing in the following sections.

From Figure 3, we can see that maximum F0 in the L tone is actually the end point of the preceding H tone, which does not show any difference between the two boundary conditions (see Table 1 and Figure 4). Does the minimum F0 in the L tone differ between the boundary conditions? With similar LMM tests in the LH and HH sentences separately, taken boundary and focus as two fixed factors with interaction, and speaker as the random factor, the minimum F0 of syllable X in the LH sentence showed no difference in the two boundary conditions either (86.4 st on average in both conditions) (Estimate = −0.369, SE = 0.426, df = 352, t = −0.866, p = 0.387). However, there was an interaction between focus and phrasing, i.e., when the L tone is focused (XF), the minimum F0 in the phrase boundary condition is significantly lower than in the syllable boundary condition (Estimate = −1.402, SE = 0.598, df = 352, t = −2.344, p = 0.0196). In the other three focus

TABLE 1 LMM analysis on maximum F0 and duration in syllable X, with HH and LH sentences separately tested taking focus and boundary as non-interactive fixed factors, whereas speaker and set as random factors in the equation as lmer(dv~focus+boundary+(1|speaker)+(1|repetition)+(1|set), data=DT), here dv stands for dependent variable, which is MaxF0 and duration.

		HH				LH			
Random effects:		Num of Observations 429							
MaxF0		Var	SD			Var	SD		
	Speaker	24.89	4.98			29.15	5.40		
	Rep	0.06	0.25			0.04	0.21		
	Set	0.01	0.13			0.24	0.49		
	Res	1.10	1.05			1.22	1.10		
Duration									
	Speaker	165.57	12.87			181.46	13.47		
	Rep	2.14	1.46			6.07	2.46		
	Set	553.76	23.53			387.53	19.93		
	Res	1049.2	32.39			1305.3	36.12		
Fixed effects:									
		Est	SE	*df*	*t*	Est	SE	*df*	*t*
MaxF0	Inter	92.0	1.67	8.22	54.91*	91.57	1.84	8.63	49.75
	XF	2.80	0.14	413	19.58*	2.05	0.15	413	13.54*
	YF	1.44	0.14	413	10.08*	0.35	0.15	413	2.29*
	ZF	0.10	0.14	413	0.70	0.07	0.15	413	0.45
	PhrB	0.01	0.10	413	0.01	0.08	0.11	413	0.75
Dur	Inter	165.92	17.55	1.22	9.46*	168.55	15.30	1.36	10.97
	XF	66.08	4.41	416	14.98*	61.93	4.92	416	12.60*
	YF	24.41	4.41	416	5.53*	30.09	4.92	416	6.12*
	ZF	−7.19	4.41	416	−1.63	6.62	4.92	416	1.35
	PhrB	20.06	3.12	416	6.44*	23.58	3.48	416	6.78*

Note: * stands for $p < 0.05$.

conditions, no difference in minimum F0 was found between the two boundary conditions.

When we labeled the speech data, we noticed that most of the L tones were creaky, that was 84.1% and 74.6% in the phrase and syllable boundary, conditions respectively. It is possible that creakiness is an additional feature of a stronger boundary, when minimum F0 cannot go any lower at a phrase boundary (Kuang, 2017).

To summarize, (1) focus is reliably realized in a tri-zone pattern, i.e., pre-focus F0 is largely intact, on-focus F0 is raised and post-focus F0 is lowered and compressed; in addition, focus increases duration of the focused syllable; (2) boundary strength has very little effect on maximum or minimum F0, but mainly realized by pre-boundary lengthening, which is independent from focus and tone; (3) The L tone is more likely to be creaky when it is before a phrase boundary than a syllable boundary.

Graphic analysis on downstep and post-low-bouncing

The analysis on focus and boundary in section "Graphic analysis on focus and phrasing" and section "Acoustic analysis on the interaction of focus and boundary" shows that the current experiment is in agreement with previous findings on these two effects (Xu, 1999; Wang et al., 2018b). It validates the following analysis on the interaction of these two functional variations with the tonal effects, i.e., downstep and post-low-bouncing. As introduced in the beginning of the results section, we here firstly report the *cross-comparison* on assessing the downstep effect adopted from Xu (1999) and Shih (2000) among many others, by comparing crossly between the HH and LH sentences (see Figure 5).

In the wide- and Z focus sentences, we can see in Figure 5 that F0 raises greatly in syllable Y in the LH sentence, which is the procedure of target approximation from a low starting point to the H target. As expected, F0 in syllable Y does not reach the height as the HH tone sentences in several H tones after the L tone, showing a clear downstep effect. We can also see that the downstep effect becomes weaker when the H tones are in a longer distance from the L tone. Five new findings are as below.

1. The downstep effect also holds when the focused word is sentence final (ZF), indicating that on-focus F0 raising in word Z does not seem to have any anticipatory effect on downstep.
2. The above observations hold in both the syllable and phrase boundary conditions. Thus, a stronger phrase boundary does not block the downstep effect. Despite a longer duration in the L tone before a phrase boundary (see Figure 4), downstep still applies. The following analysis shows that this is because the L tone is with lower F0 and even becomes creaky at a phrase boundary.
3. When the H tone right after the L tone is focused (YF), the downstep effect still shows in syllable Y but not in the following H tones. Surprisingly, even on-focus F0 raising does not cancel the downstep effect. In other words, we can say that downstep does not cancel on-focus F0 raising. It further confirms that the downstep effect is relatively robust. However, post-focus-compression (PFC) seems to override the downstep effect since there is no clear difference in the H tones after syllable Y between the HH and LH sentences, which is statistically confirmed below in Figure 6.
4. Comparing the two boundary conditions it seems that downstep is greater in the phrase boundary condition, however, in the YF condition the downstep effect is weaker in the phrase boundary condition.
5. When the L tone is under focus (XF), instead of downstep, the post-low-bouncing effect shows in the adjacent H tones. Here, the H tone after the L tone goes up first, then drops gradually, as compared to the all-H sequence, as reported in Prom-on et al. (2012). Note that the H tones in the baseline condition are realized lower, i.e., in a compressed pitch register (post-focal compression, Prom-on et al., 2012; Xu et al., 2012). The new finding is that the post-low-bouncing effect seems to be weaker in the

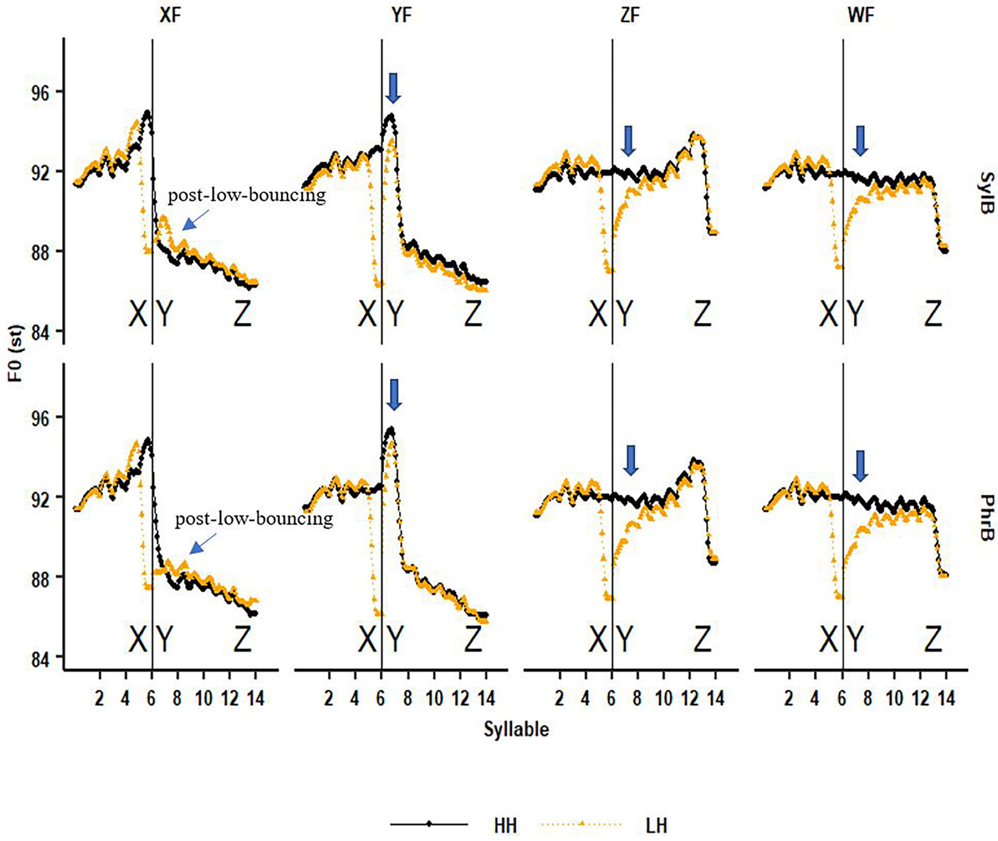


FIGURE 5
The comparison between the HH (black line) and LH (yellow line) sentences in four focus conditions (from left to right are the conditions of focus in syllable X, Y, Z and in wide-focus) under the condition that the boundary between X and Y is a syllable (SylB; upper row) or a phrase boundary (PhrB; lower row), as indicated by the vertical line. The downward arrow indicates where the downstep effect can be seen.

phrase boundary condition than in the syllable boundary condition, which supports the *balance-perturbation hypothesis* as proposed in Prom-on et al. (2012).

In summary, the graphic analysis of Figure 5 shows that: (1) The downstep effect is relatively robust in varied focus and boundary conditions. More specifically, downstep is not blocked by a phrase boundary, neither is it overridden by on-focus F0 raising or phrase boundary. (2) When the L tone is under focus, post-low bouncing is found in the following H tones, and seems to be weakened by a phrase boundary.

The *cross-comparison* of downstep effect

The main questions to be quantitatively analyzed are the size and the domain of downstep and post-low-bouncing effect, and their interactions with focus and phrase boundary.

Downstep is firstly analyzed by comparing the LH and the corresponding HH sentences in the WF, ZF and YF conditions. In the *cross-comparison,* the size of the downstep effect is calculated by the difference in maximum F0 between the H tones in the LH and HH sentence in syllable Y (syllable 7) and the following syllables (syllable 8 to 14). The post-low-bouncing effect is calculated in the X-focus condition in a similar way (see section "F0 analysis on post-low-bouncing effect").

Figure 6 presents the size of downstep effect in the three focus and two boundary conditions. The mean values show how much F0 maximum is lowered in the LH sentence as compared to the HH sentence in the corresponding syllable. Paired-sample T tests were applied in each syllable to test whether the difference reached statistical significance at the level of $p < 0.05$, which is marked by a * in Figure 6.

To get an overall statistical analysis of the factors on the downstep effect, a LMM was applied, setting focus, boundary, and syllable as fixed factors with interactions presumed (WF, syllable boundary, and the 7th syllable are set as the base-line condition), while speaker is the random factor (see Table 2). Putting it together with the t-test in Figure 6, the following findings are statistically supported: (1) The downstep effect in Y-focus condition is significantly smaller than that in wide-focus condition, while no difference is found between Z-focus and wide-focus condition. (2) The downstep effect decreases as the H tones are in longer distance from the L tone. (3) Unexpectedly, the downstep effect is greater in the phrase boundary condition than the syllable boundary

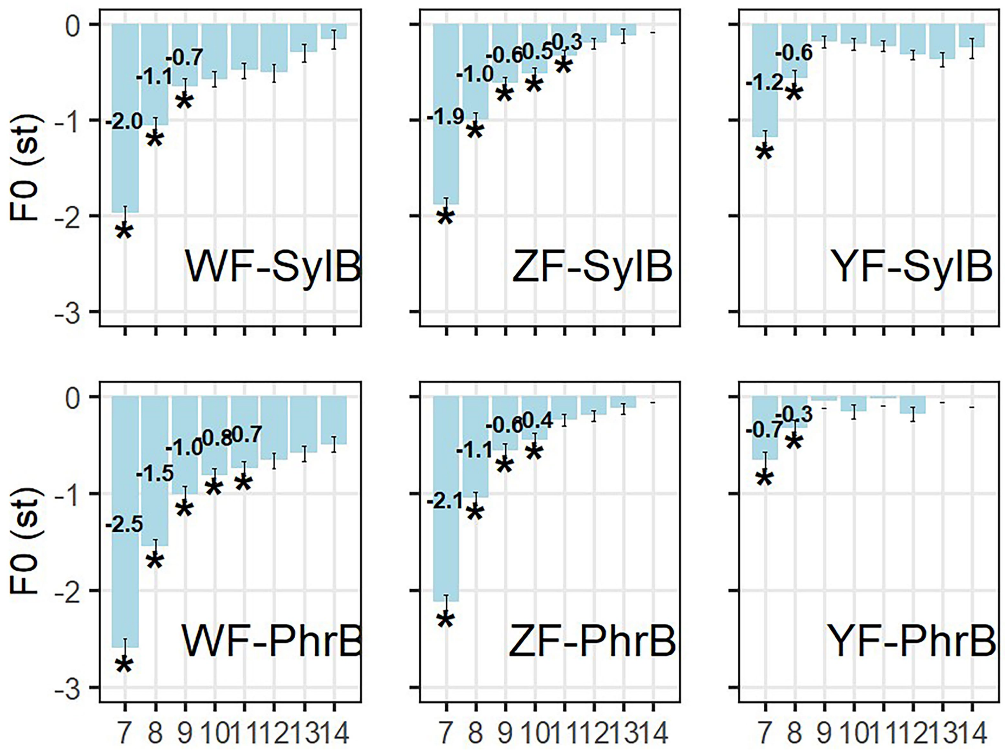


FIGURE 6

The downstep size in the Wide-focus (WF), Z-focus(ZF), and Y-focus (YF) conditions, divided by syllable (SylB) and phrase boundary (PhrB) conditions. The significant downstep effects are marked with * indicating that $p<0.05$. The x-axis shows syllable numbers, in which the 7th is syllable Y, the H tone right after the L tone.

TABLE 2 LMM analysis on downstep size (difference of maximum F0 between LH and HH sentences in the H tones) with the equation as lmer(downstepsize~focus * syllable * boundary+(1 | speaker), data=DT2).

Number of observations: 2544				
Random effects:		**Variance**	**SD**	
Speaker	(Intercept)	0.078	0.279	
Residual		1.218	1.104	
Fixed effects:	Estimate	SE	*df*	*t*
(Intercept)	2.84	0.27	435	10.47*
YF	−1.53	0.36	2,522	−4.30*
ZF	0.11	0.36	2,522	0.30
syllable	−0.20	0.02	2,522	−8.57*
PhrB	0.79	0.36	2,522	2.21*
YF:syllable	0.12	0.03	2,522	3.56*
ZF:syllable	−0.02	0.03	2,522	−0.69
YF:PhrB	−1.16	0.50	2,522	−2.31*
ZF:PhrB	−0.55	0.50	2,522	−1.08
syllable:PhrB	−0.04	0.03	2,522	−1.28
YF:syllable:PhrB	0.06	0.05	2,522	1.18
ZF:syllable:PhrB	0.02	0.05	2,522	0.45

Note: * stands for $p < 0.05$.

condition, especially in the wide-focus conditions. It is probably because the L tone is with lower minimum F0 and with more creaky voice (see section "Acoustic analysis on the interaction of focus and boundary"), thus the following H tone is with a larger difference from the all-H reference, as compared to the syllable boundary condition. (4) In the Y-focus condition, the downstep effect interacts with focus and boundary, in the way that the downstep effect in the adjacent syllable of the L tone is greater in the syllable boundary condition than in the phrase boundary condition.

The *sequential-comparison* on downstep and declination

Another way to analyze downstep is the degree of F0 lowering after a L tone. In this way, downstep effect can be compared with declination, which was analyzed by calculating the difference of the maximum F0 in the adjacent H tones in all-H sentences. Firstly, we just analyzed the two H tones surrounding the L tone, that is the difference of maximum F0 between syllable 7 and 5 (Difsy7sy5), presented in Figure 7 as boxplots divided by focus conditions, with HH and LH sentences compared directly. Since the results already show that boundary has no effect on maximum F0 in syllable X (Figure 4), we here averaged the two boundary conditions in Figure 7.

To evaluate whether there is declination in all H tone sentence, we compared "Difsy7sy5" in the wide focus condition of the HH

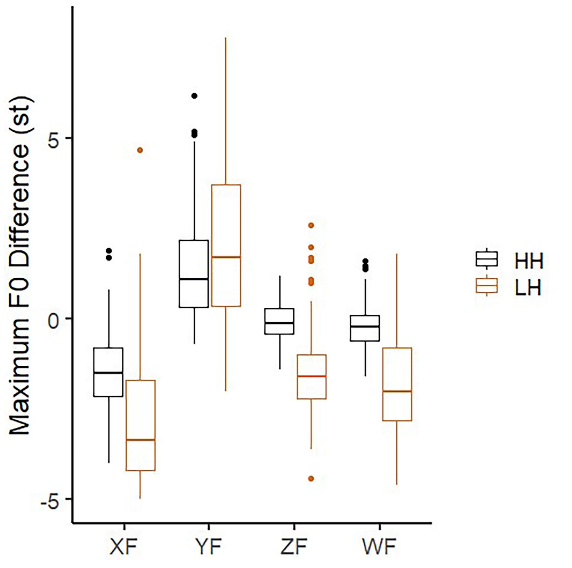


FIGURE 7
Maximum F0 difference between the two H tones surrounding syllable X (either L or H) in different focus conditions, while the two boundary conditions are averaged.

sentences with 0 in a one-sample *t*-test (*t* = −3.359, *df* = 107, *p* = 0.001). The 95% confidence interval is −0.3 to −0.07. With the same analysis, however, declination is not found in the Z-focus condition (*t* = −1.46, *df* = 105, *n.s.*). Thus, declination is to a much less degree and vulnerable to be cancelled by a final focus.

The LMM model on "Difsy7sy5," with focus, boundary and tone as fixed factors and speaker as random factor, showed a main effect in tone and focus, but not in boundary (see Table 3). It further confirms that F0 plays a limited role on differentiating boundary degrees.

In general, by comparing HH with LH in Figure 7, we can see that the difference on the degree of F0 drop in the all-H tone sentence is significantly less than the downstep effect in XF, ZF and WF conditions (*p* < 0.05). The interaction between focus and tone is not found in XF condition. We can see that Difsy7sy5 is greater in LH than in the HH sentence. Besides, Difsy7sy5 in the HH sentence is much smaller in the XF than in the WF condition, which reflects post-focus-compression (PFC) in F0. The new finding here is that downstep still shows aside from PFC. It means that the downstep effect is not just the general downtrend of F0. Declination and downstep are presumably not from the same articulatory mechanism.

When focus is on the H tone after the L tone (YF), the pitch difference between the two H tones (syl5 and syl7) is greater in the LH than the HH sentences. This comes from pre-low-raising (Lee et al., 2021). Here, it also shows the pre-low-raising is independent of on-focus F0 raising.

Then, we further tested whether declination holds all along the sentence by comparing maximum F0 of each adjacent H tones, see Figure 8. We here only consider the wide-focus condition. The *** in the figure indicates that the F0 raise or drop between two adjacent H tones in all-H sentence is greater than 0 by on-sample *t*-test with *p* < 0.001, otherwise there is no difference between the two H tones. To put it in a simple way, the *** means that there is either F0 raising or declination in the current syllable. We can see that in the HH sentences, F0 goes up in the beginning of the sentence (increased 0.34 st), then drops gradually for about 3 syllables (decreased 0.25 st). However, between syllable 7 and 6 and between syllable 9 and 8, no significant difference is found in maximum F0. These two positions are phrase boundaries. It is possible that a phrase boundary cancels declination. Toward the end of the sentence, declination is absent as well. The last syllable is a neutral tone, which causes a sharp drop in F0. Thus, declination is with a very small pitch drop between two H tones, and can be easily cancelled due to topic, boundary, tone and other reasons.

TABLE 3 LMM analysis on the difference of maximum F0 in the H tones before and after syllable X, with focus, boundary and tone as fixed factors, whereas speaker as a random factor in the formula as: difs7s5~tone * focus + boundary + (1 | speaker).

Number of observations: 858				
Random effects:		**Variance**	**SD**	
Speaker	(Intercept)	0.3804	0.6167	
Residual		1.7171	1.3104	
Fixed effects:	Estimate	SE	*df*	*t*
(Intercept)	−0.138	0.245	14.755	−0.56
toneLH	−1.655	0.179	841.013	−9.214*
XF	−1.264	0.178	840.997	−7.088*
YF	1.724	0.178	841.000	9.649*
ZF	0.101	0.179	841.011	0.566
boundaryPhrB	−0.103	0.089	841.001	−1.149
toneLH:XF	−0.084	0.253	841.005	−0.332
toneLH:YF	2.326	0.253	841.013	9.181*
toneLH:ZF	0.298	0.253	841.009	1.178

Note: * stands for *p* < 0.05.

If we look at the LH sentences, we can see that the H tones around the L tone causes much greater F0 change than the all-H sentence. Toward the end of the sentence, the adjacent H tones do not differ in F0, which is similar to the all-H sentence. It is in agreement with the *cross-comparison* of the downstep effect, that downstep gets weaker as the H tones are further from the L tone.

All-together, both the *cross-* and *sequential-comparison* show that downstep effect is in a much greater degree than declination. Downstep effect is robust, lasting for about 2–3 syllables. Downstep effect is not cancelled by focus or boundary, whereas declination can be cancelled by these two informative functions.

Creaky L tone

The very last question concerning downstep is whether it is caused by creaky voice, or whether creaky L tones cause greater downstep effect. The number of creaky L tone in different conditions is presented in Table 4. In line with previous studies, we also see that

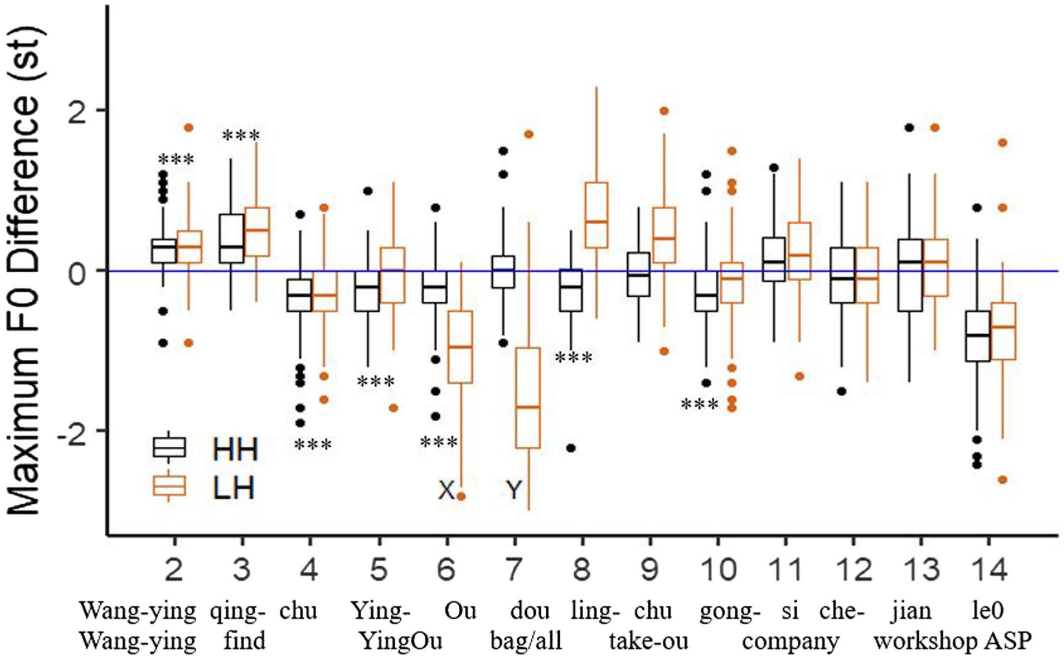


FIGURE 8
Boxplot of the maximum F0 difference between two adjacent H tones, as compared between HH and LH wide-focus sentences. The number in the *x*-axis (2–14) means that this is the maximum F0 of the current syllable minus the preceding syllable. *** here indicates significant difference between 0 by one-sample *t*-test in the HH sentences.

TABLE 4 The percentage of creaky L tone in different focus and boundary conditions (%).

	Syllable boundary	Phrase boundary
XF	95.8	91.6
YF	60.4	72.9
ZF	70.8	85.4
WF	64.5	83.3

when the L tone is under focus and at a phrase boundary, it is more likely to be creaky. We do not go into detailed analysis on the acoustic parameters of the creaky L tone. Instead, we simply calculate the amount of creakiness in L tones to answer the question whether a creaky low tone causes greater downstep effect. In Figure 9, the maximum F0 of syllable Y is plotted against the duration of the creaky part in syllable X, with four focus conditions divided in different plots. When the creaky duration is 0, it means this is a normal L tone. Here we do not see any clear trend of a creaky L tone causes lower F0 in the following H tone, which is supported by the LMM model analysis with creaky, focus and gender as fixed factors and speaker as a random factor (lmer(maxF0syl7 ~ Creakylablel*focus*Gender+ (1|speaker), data = creaky)). The LMM shows significant effect in focus and gender, whereas creaky does not show any effect (Estimate = −0.2455, SE = 0.416, *df* = 347, *t* = −0.59, *n.s.*). Thus, creaky L tone is not the direct cause of downstep, but strengthens downstep. It then explains why downstep effect is greater after a phrase boundary (Figure 6).

F0 analysis on post-low-bouncing effect

The post-low-bouncing effect was calculated as the difference of maximum F0 in the H tones between the LH and HH sentences in the XF condition (the L tone is on-focus). As can be seen in Figures 7, 10, F0 maximum is lower in the syllable right after the L tone (syllable 7), that is because the F0 maximum of syllable 7 in the HH sentence is the offset of the previous H tone (the maximum F0 hence appears at the onset of the syllable 7 representing the transition from the focused H tone to a post-focally H tone). Post-low-bouncing shows at the end of syllable 7, which can be observed in the maximum F0 of syllable 8 and 9, then the pitch gradually drops back. The linear-mixed-model analysis was carried out with syllable and boundary as two fixed factors (with interaction), while speaker and set are random factors. In this statistical test, we only considered syllable 8 to 10, since no difference is shown between HH and LH sentences after the 10th syllable (see Figure 10). The LMM analysis shows a significant effect in syllable (SE = 0.093, *t* = −5.087*), boundary (SE = 1.193, *t* = −3.018*) and the interaction (SE = 0.132, *t* = 2.899*). Thus, the following observations in Figures 5, 10 are statistically supported: (1) The post-low-bouncing effect gradually decreases in the syllables after the L tone; (2) A phrase boundary weakens post-low-bouncing effect, especially in the second H tone after the L tone.

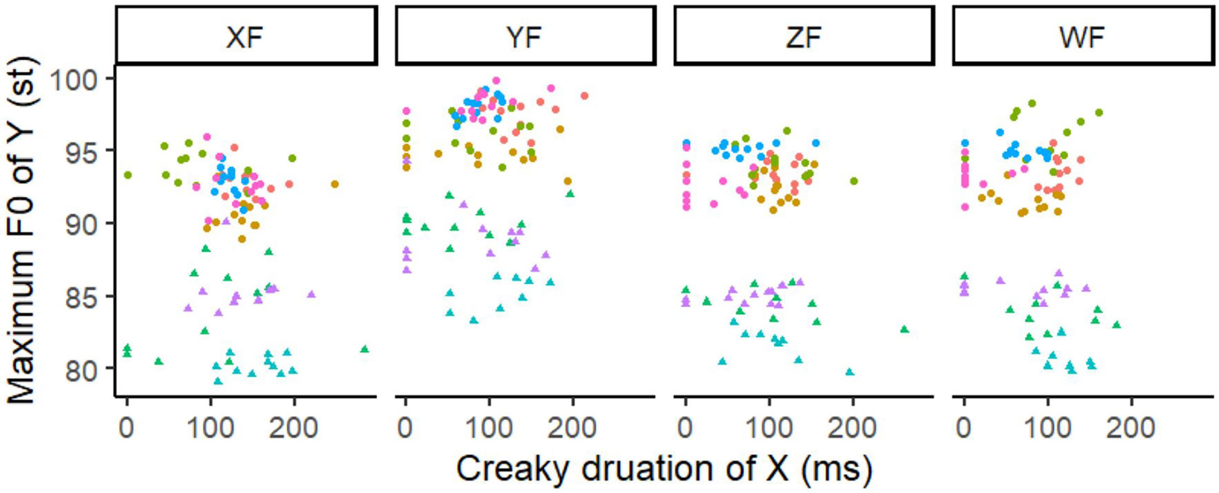


FIGURE 9
Scatter plot of the maximum F0 in syllable Y as functioned by duration of the creaky part in syllable X, with the color and shape differentiating speakers. The focus conditions are divided in each plot. The points with creaky duration of X being 0 means that this is a normal L tone.

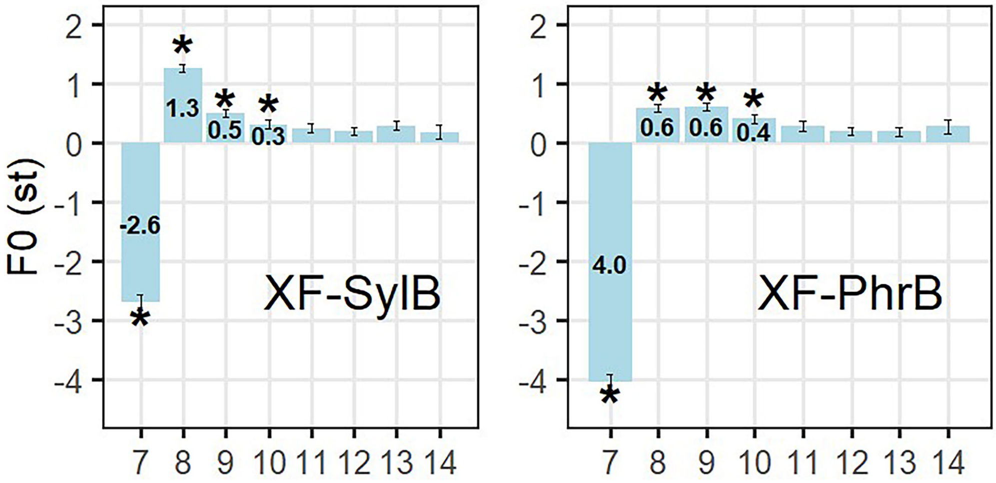


FIGURE 10
Post-low-bouncing effect in the X-focus condition when the boundary between syllable X and Y is either a syllable (SylB) or a phrase (PhrB) boundary. The *x*-axis shows syllable numbers, in which the 7th is syllable Y, the H tone right after the L tone.

General discussion

The new contribution of the current study is on how pragmatic functions interacts with downstep and post-low-bouncing. With the control of focus, post-low-bouncing was brought in, which was mainly analyzed for neutral tones in previous studies (Chen and Xu, 2006; Prom-on et al., 2012). In the current study, it happened in the following H tones when the L tone is focused (XF in Figure 5). Although, a large part of intonation variation is informative, we want to emphasize that articulatory constrains on pitch change could not be neglected, since post-low-bouncing and downstep last for several syllables with decreasing in size from of 2.5 to 0.5 st, interacting actively with phrasing and focus. Our findings support the *additive division hypothesis* of pitch range, proposed in Liu et al. (2021). They found that pitch range of 5–12 st above the baseline signals both focus and surprise, suggesting an overlap between different layers of meanings within this pitch range. In their study, to perceive focus, F0 needs to be raised about 3 st. We here show the downstep and post-low-bouncing are in a pitch range of less than 2.5 st (Figures 6, 10), whereas on-focus F0 raising in syllable X is 2.8 st on average (Table 1). In a rough sense, it explains why on-focus F0 raising needs to be about 3st, beneath which F0 variation reflects tone and articulatory constrains. It is possible that any sudden and great pitch raising may bring-in informative meaning, thus it takes several syllables for downstep and post-low-bouncing to go back to the reference line. The process of target approximation as proposed in PENTA model (Xu et al., 2022) probably reflects both articulatory and perceptual constrains.

Relating to the tonal variation due to the L tone, the pre-low-raising was systematically studied in Lee et al. (2021) in Thai and Cantonese, that is, the H tone is raised in pitch before a L tone. They discussed three possible explanations: (a) a velocity account, (b) a perceptual account, and (c) an anatomical account. More specifically, (a) the raising pitch in the preceding syllable may increase the distance of the downward movement toward the low tone; (b) pre-low-bouncing may enhance tonal contrasts to aid comprehension; (c) if pre-low-raising is not actively planned, it may be the direct result of intrinsic laryngeal muscle movement. Their analysis does not support (b), the perceptual account. Putting it together with Prom-on et al. (2012) and the current study, we can conclude that pitch movements caused by a L tone (pre-L-raising, downstep and post-low-bouncing) are largely the outcome of intrinsic and extrinsic laryngeal muscle movement. Below we will provide detailed discussion on the research questions.

How do focus and boundary interact with downstep (Q1-Q6)?

This is actually a very complicated question, since focus in Mandarin involves both on-focus raising and post-focus-compression in F0 (Shih, 1988; Xu, 1999; Chen and Gussenhoven, 2008; Wang and Xu, 2011; Wang et al., 2018b), see Figure 2 in the current study. Besides, downstep refers to the relevant pitch height in the H tones, either as compared to all-H reference line (*cross-comparison* answering Q1-Q4), or as the F0 drop between the H tones before and after the L tone (*sequential-comparison* answering Q5). To make the question even more complicated, L tones usually become creaky (Q6). No previous study has considered the influence of creakiness on downstep (Q6). Thus, the first question is split to the following 6 sub-questions, aiming to fully understand the property of downstep, and to take apart declination and downstep. The results are interpretable and coherent to each other if we take the idea that downstep is mostly constrained by articulatory movement, instead of conveying linguistic meaning.

Q1: Does downstep set up a new register tone?

The answer is No. The original motivation of this study was whether downstep in Mandarin can be modelled as a phonetic or as a phonological tonal interaction. On the one hand, downstep was observed in West-African tone languages (Welmers, 1959). The downstepped H tone defines a new ceiling for subsequent tones which was interpreted as a systematic, phonological effect, and downstep was phonologically modelled in terms of register tones (Snider, 1998) or register features (Akumbu, 2019). On the other hand, if downstep were a phonetic effect, the expectation is that the locally lowered F0 raises gradually back to its original register line. The present study suggests that downstep in Mandarin is indeed a phonetic tonal interaction. We observed that after a L tone, F0 does not raise back to the height of the all-H-tone sentences and lasts for several H tones decreasing in size, as has been repeatedly found in previous studies in Mandarin (Shih, 1988; Xu, 1999). The locally induced tonal interaction smoothly levels out such that the original reference line for a high tone in Mandarin is reached again (Figure 6). Thus, downstep in Mandarin is different from those in African languages. Moreover, our data showed that the effect size and the domain of the effect vary as a function of focus and prosodic boundary in Mandarin.

Q2: Does a sentence-final focus ends downstep?

We predicted that the answer is no because downstep is presumably local, and pitch target of each tone is realized syllable-by-syllable as stated in PENTA model (Xu et al., 2022). Indeed, we found that a late focus does not end downstep. Unexpectedly, downstep effect lasts longer in the Z-focus condition than in the wide focus condition (Figure 6). This is probably different from Hausa (Lindau, 1986), in which downstep can be canceled in yes/no questions. It is possible that speakers try not to cause confusion, otherwise any pitch raising before the final word may increase the prominence level in that word, given that sentence-final focus is quite similar to wide focus intonation (Xu, 1999; Liu and Xu 2007; Xu et al., 2012). Since the study on Hausa concerns question intonation, whereas ours is on final-focus, a controlled study of downstep in yes-no-questions in Mandarin would shed more light on this case.

Q3: Is downstep eliminated by on-focus F0 raising and post-focus-compression?

We predicted that informative functions of intonation may override an articulatory effect. However, the results show that downstep is only weakened by on-focus F0 raising and post-focus-compression but not fully cancelled. This result is new. It indicates that downstep, as an articulatory pitch movement, is pretty robust. According to Xu and Sun (2002), the time of pitch rise can be estimated by $t = 89.6 + 8.7\ d$ (here d stands for the change of pitch in semitone). Using this algorithm, we calculated the estimated time from the minimum F0 of the L tone to the maximum F0 of the following H tone. The exact duration of the H tone is actually longer than the estimated time (mean = 33 ms, sd = 35.6). It means that the observed downstep is not because of time pressure. When the H tone is focused (YF), the exact H tone duration is 60 ms (sd = 47.8) longer than the estimated time, however downstep effect still shows (Figures 5, 6). It further confirms that even in the condition of a longer H tone, downstep still applies. Thus, we draw the conclusion that informative intonation functions do not override downstep. The interaction between focus and downstep is gradual.

Q4: How does a phonological phrase boundary interact with downstep?

As predicted, the pre-boundary L is lengthened at a phrase boundary (Figure 4), the tonal target is fully realized with higher frequency of being creaky (Table 4), and in turn, it leads to greater downstep effect (see Figure 6). In wide focus condition, the L tone is lengthened about 14 ms in the phrase boundary, with no difference in minimum F0 between the two boundary conditions (86.7 vs. 86.1 st). Instead, creaky L tone occurs more frequently in the phrase boundary condition than the syllable condition (84% vs. 67%). That might be the reason why the H tone is a little lower in the phrase boundary condition than in the syllable boundary condition (90.1 st vs. 90.4 st), showing as a greater downstep effect under the phrase-boundary condition. However, creakiness *per se* does not seem to cause downstep (see below, Q6).

Q5: Do declination and downstep share the same mechanism?

The answer to this question actually depends on how to measure declination and downstep. It also remains controversial whether there is any separate articulatory mechanism of declination. We here take the *sequential-comparison* by calculating the difference of adjacent H tones (Figures 7, 8). As predicted, we can see that downstep and declination come from different articulatory control. However, it is not because downstep is local whereas declination is global, rather downstep lasts for several syllables as well. It is because the downstep effect shows in a larger scale and in a more robust manner than declination. It is possible that there is some underlying articulatory control on declination, however, it is pretty weak and vulnerable to be overridden by varied reasons. We are in agreement with other studies (Xu, 1999; Shih, 2000; Yuan and Liberman, 2014), showing that the general global downtrend, as modelled with a top and bottom regression line of intonation, is a combined effect from different functions. We further suggest not to just take the global downtrend in an abstract way, but to analyze it with full consideration of local tonal interactions.

Q6: Is creaky voice the cause of downstep?

Downstep is caused by a L tone, which is usally creaky in Mandarin (Kuang, 2017). Is it possible that creaky voice is the main cause of downstep? In our study we found that the L tone is more likely to be creaky when it is under focus and before a phrase boundary (Table 4). It confirms the claim by Kuang (2017) that creaky voice correlates with low pitch target. As discussed in Q4, more creaky L tones at a phrase boundary causes greater downstep effect. However, normal L tone causes roughly the same degree of downstep, as showed in the LMM that creakiness does not have any effect on the maximum F0 of the following H tone. No correlation is found between the duration of the creaky part in L tones and the pitch height in the following H tones (Figure 9). It indicates that creaky voice is probably not the direct cause of downstep. A normal L tone also causes downstep. However, a creaky L tone leads to a greater downstep effect.

Does a phrase boundary block post-low-bouncing (Q7)?

According to the *balance-perturbation hypothesis* (Prom-on et al., 2012), we predicted that post-low-bouncing is weakened if the L tone is at a phrase boundary. It is indeed the case, as shown in Figure 7. They hypothesized that after producing a very low F0, the extrinsic laryngeal muscles (e.g., sternohyoids) stop contracting to maintain the balance between the two antagonistic forces in the intrinsic laryngeal muscles. When the L tone is focused, the extra force may cause a sudden increase of the vocal fold tension, resulting in the raise in F0 in the following H tone. We here see that when the L tone is before a prosodic phrase boundary, it then probably gives a little more time to release the tension between the extrinsic and intrinsic laryngeal muscles. This would explain the difference in size of post-low bouncing found in our data. In line with Prom-on et al. (2012), we also found that post-low-bouncing occurs in H tones when the L tone is under focus. In their study, neutral tones after a L tone show post-low-bouncing. The reason might lie in the fact that post-focal words are weakened in intensity and compressed in F0. The weakened H tones at post-focal position might share some similar mechanism with weak articulatory movement in the neutral tones.

At last, we here briefly introduce some preliminary findings in the current study, relating to Moisik et al. (2014). They have found that low F0 tone targets in Mandarin can not only be reached by lowering the larynx, but also by combining the raise of larynx height and laryngeal constriction, which may lead to creakiness in the low tone. In the L tone, the amount of F0 lowering correlates with larynx lowering in male speakers (r = 0.73 and 0.86), while the female speaker uses larynx raising ($r = 0.13$; Figures 11-13, pp. 39 in their study). In our study, the minimum F0 of the low tone (X) is positively correlated to the maximum F0 of the following H tone (Y) in the male speakers (wide focus: $y = -2 + 0.99x$, $r^2 = 0.66$; X-focus condition: $y = 9.4 + 0.87x$, $r^2 = 0.736$), but not in the female speakers (wide focus: $y = 65 + 0.27x$, $r^2 = 0.073$; X-focus condition: $y = 79 + 0.12x$, $r^2 = 0.01$). To fully understand the anatomical process in downstep, articulatory studies considering gender difference are required.

Conclusion

To answer all the research questions concerning the interaction of focus/boundary with downstep/post-low-bouncing, we can draw the following conclusions.

In the wide focus condition, the downstep effect lasted for 3 syllables and gradually reached back to the all-H tone reference line. Downstep thus does not set up a new reference line in Mandarin (Q1). A sentence-final focus makes the downstep effect last for 5 syllables (Q2). When the H tone right after the low tone was focused (YF), on-focus F0 raising and post-focus-compression (PFC) weakened downstep (Q3). A phrase boundary strengthened downstep (Q4). We further analyzed downstep by measuring the F0 drop between the two H tones surrounding the L tone (*sequential-comparison*). Comparing it with F0 drop in all-H sentences, it showed that the downstep effect was much greater and more robust than declination (Q5). However, creaky voice in the L tone was not the direct cause of downstep (Q6). At last, when the L tone was under focus (XF), it caused a post-low-bouncing effect on the following H tones and lasted for about 3 syllables with F0 dropping back gradually. Moreover, post-low-bouncing is weakened by a phonological phrase boundary (Q7).

In general, this study showed that downstep and post-low-bouncing, as articulatory controls and local tonal interaction effects, interact with the execution of sentence-level pragmatic functions like focus and prosodic boundary. Pragmatic effects do not cancel or override articulatory effects, but affect the size and domain of the tonal interactions.

Data availability statement

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

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

FK initiated the general research question. BW, FK, and SG designed the experiment together. BW checked the labeling of the wav files, wrote the paper, and finalized the data analysis, closely working together with FK. SG did a preliminary graphic and statistical analysis. All authors contributed to the article and approved the submitted version.

Funding

The experiment was supported by SFB632 "Information Structure" at Potsdam University. The writing procedure was further supported by Social Science Foundation of China to BW (18BYY079). Goethe University Frankfurt provided financial support for open access support, and a DFG grant KU 2323-4/1 to FK supported further assistance.

Acknowledgments

We thank Xiaxia Zhang and Qian Wu for running the experiment and labeling the speech data. We also thank Yi Xu for constructive discussions and suggestions. Part of the preliminary results were presented at the TAL conference in 2018 (Wang et al., 2018a), we thank the audience.

Conflict of interest

SG is employed by i2x GmbH.

The remaining 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. The research was initiated while SG was affiliated with Potsdam University, and hence the research was conducted without 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.2022.884102/full#supplementary-material

References

Akumbu, P. W. (2019). "A featural analysis of mid and downstepped high tone in Babanki," in *Theory and Description in African Linguistics: Selected Papers From the 47th Annual Conference on African Linguistics*. eds. E. Clem, P. Jenks and H. Sande (Berlin: Language Science Press), 3–20.

Arvaniti, A., and Fletcher, J. (2020). The Autosegmental-Metrical theory of intonational phonology. *Oxford Handbooks in linguistics. The Oxford handbook of language prosody*, 77–95.

Atkinson, J. E. (1978). Correlation analysis of the physiological factors controlling fundamental voice frequency. *J. Acoust. Soc. Am.* 63, 211–222. doi: 10.1121/1.381716

Baer, T. (1979). Reflex activation of laryngeal muscles by sudden induced subglottal pressure changes. *The Journal of the Acoustical Society of America* 65, 1271–1275.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2016). *lme4: Linear Mixed-Effects Models Using Eigen and S4 (R Package Version 1.112)*. Vienna: R Foundation for Statistical Computing.

Boersma, P., and Weenink, D. (2013–2022). Available at: http://www.fon.hum.uva.nl/praat/

Byrd, D. (2000). Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica* 57, 3–16. doi: 10.1159/000028456

Chao, Y. R. (1968). *A Grammar of Spoken Chinese*. Berkeley, CA: University of California Press.

Chen, Y.-Y. (2006). Durational adjustment under corrective focus in Standard Chinese. *Journal of Phonetics* 34, 176–201. doi: 10.1016/j.wocn.2005.05.002

Chen, Y.-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

Chen, Y.-Y., and Xu, Y. (2006). Production of weak elements in speech -- evidence from f0 patterns of neutral tone in standard Chinese. *Phonetica* 63, 47–75. doi: 10.1159/000091406

Cohen, A., Collier, R., and 't Hart, J. (1982). Declination: Construct or Intrinsic Feature of Speech Pitch? *Phonetica* 39, 254–273.

Collier, R. (1975). Physiological correlates of intonation patterns. *The Journal of the Acoustical Society of America* 58, 249–255.

Connell, B. (2001). *Downdrift, downstep, and declination*. Paper presented at the Typology of African Prosodic Systems Workshop, Bielefeld University, Germany.

Connell, B. (2011). "Downstep" in *Companion to Phonology*. eds. M. V. Oostendorp, C. J. Ewen, E. Hume and K. Rice (Oxford: Blackwell Publishing), 824–847.

Connell, B. (2017). "Tone and intonation in Mambila," in *Intonation in African Tone Languages*. eds. L. J. Downing and A. Rialland (Berlin: De Gruyter), 132–166.

Cooper, W. E., Eady, S. J., and Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *J. Acoust. Soc. Am.* 77, 2142–2156. doi: 10.1121/1.392372

Cooper, W. E., and Sorensen, J. M. (1977). Fundamental frequency contours at syntactic boundaries. *The Journal of the Acoustical Society of America* 62, 683–692.

Courtenay, K. (1971). Yoruba: A'terraced-level'language with three tonemes. *Studies in African Linguistics* 2, 239–255.

de Jong, K. J. (1995). The supraglottal articulation of prominence in English: linguistic stress as localized hyperarticulation. *J. Acoust. Soc. Am.* 97, 491–504. doi: 10.1121/1.412275

de Pijper, J. R., and Sandeman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *J. Acoust. Soc. Am.* 96, 2037–2047. doi: 10.1121/1.410145

DiCanio, C., Benn, J., and Castillo García, R. (2021). Disentangling the effects of position and utterance-level declination on the production of complex tones in Yoloxóchitl Mixtec. *Lang. Speech* 64, 515–557. doi: 10.1177/0023830920939132

Edmondson, J. A., and Esling, J. H. (2006). The valves of the throat and their functioning in tone, vocal register and stress: laryngoscopic case studies. *Phonology* 23, 157–191. doi: 10.1017/S095267570600087X

Féry, C., and Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents in German. *J. Phon.* 36, 680–703. doi: 10.1016/j.wocn.2008.05.001

Ge, C., and Li, A. (2018). "Declination and boundary effect in Cantonese declarative sentence" in *Paper presented at the 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*

Gelfer, C. E., Harris, K. S., Collier, R., and Baer, T. (1983). "The Denver Center for the Performing Atrs, Inc." in *Is declination actively controlled Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control* (Denver, Colorado), 113–126.

Gelman, A., and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*: Cambridge university press.

Genzel, S. (2013). *Lexical and post-lexical tones in Akan*. (PhD Thesis.), Universität Potsdam, Potsdam.

Genzel, S., and Kügler, F. (2011). "Phonetic Realization of Automatic (Downdrift) and non-automatic Downstep in Akan." in *Paper presented at the Proceedings of the XVII ICPhS*, Hong Kong.

Gerratt, B. R., and Kreiman, J. (2001). Toward a taxonomy of nonmodal phonation. *Journal of Phonetics* 29, 365–381.

Gu, W., and Lee, T. (2009). Effects of tone and emphatic focus on F0 contours of Cantonese speech: A comparison with standard Chinese. *Chin. J. Phon.* 2, 133–147.

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

Hirose, H. (1997). "Investigating the physiology of laryngeal structures," in *The handbook of phonetic sciences*. eds. W. J. Hardcastle and J. Laver (Oxford: Blackwell), 116–136.

Hirschberg, J., and Pierrehumbert, J. B. (1986). "The intonational structuring of discourse" in *Paper presented at the 24th Annual Meeting of the Association of Computational Linguistics*

Hollien, H. (1974). On vocal registers. *Journal of Phonetics* 2, 125–143.

Hollien, H. (1983). "In search of vocal frequency control mechanisms" in *Vocal Fold Physiology: Comtemporary Research and Clinical Issues*. eds. D. M. Bless and J. H. Abbs (College-Hill Press), 361–367.

Hombert, J.-M. (1974). Universals of downdrift: their phonetic basis and significance for a theory of tone. *Studies in African Linguistics* 5, 169–183.

Honda, K. (1995). "Laryngeal and extra-laryngeal mechanisms of F0 control," in *Producing Speech: Contemporary Issues*. eds. F. Bell-Berti and L. J. Raphael (New York, NY: American Institute of Physics), 215–232.

Huffman, M. K. (2005). Segmental and prosodic effects on coda glottalization. *Journal of Phonetics* 33, 335–362.

Hyman, L. M., and Leben, W. R. (2017). Word prosody II: tone systems. *UC Berkeley PhonLab Annu. Rep.* 13, 178–209. doi: 10.5070/P7131040752

Ishihara, S. (2007). Major phrase, focus intonation, multiple spell-out (MaP, FI, MSO). *Linguistic Rev.* 24, 137–167. doi: 10.1515/TLR.2007.006

Keating, P. A., Garellek, M., and Kreiman, J. (2015). "Acoustic properties of different kinds of creaky voice" in *Paper presented at the ICPhS*

Krivokapic, J., and Byrd, D. (2012). Prosodic boundary strength: an articulatory and perceptual study. *J. Phon.* 40, 430–442. doi: 10.1016/j.wocn.2012.02.011

Kuang, J. (2017). Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *The Journal of the Acoustical Society of America* 142, 1693–1706.

Kuang, J. (2018). The influence of tonal categories and prosodic boundaries on the creakiness in Mandarin. *The Journal of the Acoustical Society of America* 143:EL509-EL515.

Kügler, F. (2017). "Tone and intonation in Akan," in *Intonation in African Tone Languages*. eds. L. Downing and A. Rialland (Berlin: Mouton de Gruyter), 89–129.

Kügler, F., and Calhoun, S. (2020). "Prosodic encoding of information structure: a typological perspective," in *The Oxford Handbook of Language Prosody*. eds. C. Gussenhoven and A. Chen (Oxford: Oxford University Press).

Ladd, D. R. (1988). Declination "reset" and the hierarchical organization of utterances. *J. Acoust. Soc. Am.* 84, 530–544. doi: 10.1121/1.396830

Ladd, D. R. (2008). *Intonational phonology*. Cambridge: Cambridge University Press.

Ladefoged, P. (1973). The features of the larynx. *Journal of Phonetics* 1, 73–83.

Laniran, Y. O., and Clements, G. N. (2003). Downstep and hihg raising: interacting factors in Yoruba toe production. *J. Phon.* 31, 203–250. doi: 10.1016/S0095-4470(02)00098-0

Leben, W. R. (2014). The Nature(s) of Downstep. Paper Presented at the SLAO/1er Colloque International, Humboldt Kolleg Abidjan.

Lee, A., Prom-on, S., and Xu, Y. (2021). Pre-low raising in Cantonese and Thai: effects of speech rate and vowel quantity. *J. Acoust. Soc. Am.* 149, 179–190. doi: 10.1121/10.0002976

Liberman, M. Y., and Pierrehumbert, J. (1984). "Intonational invariance under changes in pich range and length" in *Language sound structure*. eds. M. Aronoff and R. O (Cambridge, MA: MIT), 157–233.

Lieberman, P. (1967). *Intonation, perception, and language: Cambridge*, MA: MIT Press.

Lieberman, P., and Tseng, C. (1980). On the fall of the declination theory: breath-group versus "declination" as the base form for intonation. *The Journal of the Acoustical Society of America* 67:S63.

Lin, M., and Yan, J. (1980). Beijinghua qingsheng de shengxue xingzhi (The acoustic nature of the mandarin neutral tone). *Fangyan (Dialect)* 3, 166–178.

Lindau, M. (1986). Testing a model of intonation in a tone language. *J. Acoust. Soc. Am.* 80, 757–764. doi: 10.1121/1.393950

Lindblom, B. (1990). *Explaining phonetic variation: A sketch of the H&H theory speech production and speech modelling*. Berlin: Springer, 403–439.

Lindblom, B. (2009). F0 lowering, creaky voice, and glottal stop: Jan Gauf- fin's account of how the larynx works in speech. Paper presented at the Fonetik 2009, Stockholm.

Lindqvist-Gauffin, J. (1969). *Laryngeal Mechanisms in speech. Quarterly Progress and Status Report, Speech Transmission Laboratory*. Stockholm: Royal Institute of Technology, 26–31.

Lindqvist-Gauffin, J. (1972). *A Descriptive Model of Laryngeal Articulation in Speech. Quarterly Progress and Status Report, Speech Transmission Laboratory*. Stockholm: Royal Institute of Technology, 1–9.

Liu, F., and Xu, Y. (2005). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica* 62, 70–87.

Liu, F., and Xu, Y. (2007). "Question intonation as affected by word stress and focus in English." in *Paper presented at the 16th International Congress of Phonetic Sciences*, Saarbrucken.

Liu, X., Xu, Y., Zhang, W., and Tian, X. (2021). Multiple prosodic meanings are conveyed through separate pitch ranges: evidence from perception of focus and surprise in mandarin Chinese. *Cogn. Affect. Behav. Neurosci.* 21, 1164–1175. doi: 10.3758/s13415-021-00930-9

Maeda, S. (1976). *A characterization of American English Intonation*. (Doctoral Dissertation) MIT.

Moisik, S. R., and Esling, J. H. (2014). Modeling the biomechanical influence of epilaryngeal stricture on the vocal folds: A low-dimensional model of vocal–ventricular fold coupling. *J. Speech Lang. Hear. Res.* 57, S687–S704. doi: 10.1044/2014_JSLHR-S-12-0279

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

Murry, T. (1971). Subglottal pressure and airflow measures during vocal fry phonation. *Journal of Speech and Hearing Research* 14, 544–551.

Nakai, S., Kunnari, S., Turk, A., Suomi, K., and Ylitalo, R. (2009). Utterance-final lengthening and quantity in northern Finnish. *J. Phon.* 37, 29–45. doi: 10.1016/j.wocn.2008.08.002

Nakajima, S., and Allen, J. F. (1993). A study on prosody and discourse structure in cooperative dialogues. *Phonetica* 50, 197–210.

Odden, D. (1986). On the role of the obligatory contour principle in phonological theory. *Language* 62, 353–383. doi: 10.2307/414677

Ohala, J. J. (1972). How is pitch lowered? *J. Acoust. Soc. Am.* 52:124. doi: 10.1121/1.1981808

Pierrehumbert, J. (1979). The perception of fundamental frequency declination. *Journal of the Acoustics Society of America* 66, 363–379.

Pierrehumbert, J. (1980). *The phonology and phonetics of english intonation*. (Ph. D. doctoral thesis), Massachusetts Institute of Technology, Cambridge.

Pierrehumbert, J. B., and Beckman, M. E. (1988). *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Prom-on, S., Liu, F., and Xu, Y. (2012). Post-low bouncing in mandarin Chinese: acoustic analysis and computational modeling. *J. Acoust. Soc. Am.* 132, 421–432. doi: 10.1121/1.4725762

Rialland, A., and Somé, A.-P. (2011). "Downstep and linguistic scaling in Dagara-Wulé" in *Tones and features: Phonetic and Phonological Perspectives*. eds. J. A. Goldsmith, W. E. Hume and L. Wetzels (Berlin & New York: Mounton De Gruyter), 108–134.

R Core Team. (2016). *R: A language and environment for statistical computing.* Vienna: R Foundation forStatisticalComputing.

Selkirk, E. (2011). The syntax-phonology interface. In J. A. Goldsmith, J. Riggle and A. C. L. Yu (Eds.), *The handbook of phonological theory (2nd ed.)* (Vol. *2*, pp. 435–484). Oxford: Wiley-Blackwell.

Selkirk, E., and Tateishi, K. (1991). "Syntax and downstep in Japanese," in *Interdisciplinary Approaches to Language: Essays in Honor of S.-Y. Kuroda*. eds. C. Georgopoulos and R. Ishihara (Dordrecht: Kluwer Academic Publishers), 519–543.

Shen, J. (1994). Hanyu yudiao gouzao he yudiao leixing (intonation structures and patterns in mandarin) (in Chinese). *Fangyan (Dialect)* 3, 221–228.

Shih, C. L. (1988). Tone and intonation in mandarin. Working papers, Cornell phonetics. *Laboratory* 3, 83–109.

Shih, C. (2000). "A declination model of Mandarin Chinese" in *Intonation: Analysis, Modelling and Technology*. ed. A. Botinis (Kluwer Academic Publishers), 243–268.

Shih, C., and Lu, H.-Y. D. (2010). "Prosody transfer and suppression: Stages of tone acquisition" in *Paper presented at the Speech Prosody 2010-Fifth International Conference*

Sluijter, A., and Terken, J. (1993). Beyond sentence prosody: paragraph intonation in Dutch. *Phonetica* 50, 180–188.

Snider, K. L. (1990). Tonal upstep in Krachi: evidence for a register tier. *Language* 66, 453–474. doi: 10.2307/414608

Snider, K. (1998). "Tone and utterance length in Chumburung: an instrumental study." in *Paper presented at the The 28th Colloquium on African Languages and Linguistics*.

Snider, K., and van der Hulst, H. (1993). *Issues in the Representation of Tonal Register* Berlin: Mouton de Gruyter, 1–27.

Sorensen, J. M., and Cooper, W. E. (1980). "Syntactic coding of fundamental frequency in speech production" in *Perception and production of fluent speech*. ed. R. A. Cole (Hillsdale, NJ: Erlbaum), 399–440.

Stevens, K. N. (2000). *Acoustic phonetics*. New York: MIT press.

Stewart, J. M. (1965). The typology of the Twi tone system. *Bull. Inst. African Stud.* 1, 1–27.

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *J. Acoust. Soc. Am.* 101, 514–521. doi: 10.1121/1.418114

Swerts, M., and Geluykens, R. (1994). Prosody as a marker of information flow in spoken discourse. *Lang. Speech* 37, 21–43. doi: 10.1177/002383099403700102

Titze, I. R. (1988). The physics of small-amplitude oscillation of the vocal folds. *The Journal of the Acoustical Society of America* 83, 1536–1552.

't Hart, J., and Cohen, A. (1973). Intonation by rule: a perceptual quest. *Journal of Phonetics* 1, 309–327.

Tucker, A. N., and Creider, C. A. (1975). Downdrift and downstep in Luo. In R. K. Herbert (Ed.), *Proceedings of the Sixth Conference on African Linguistics, OSU Working Papers in Linguistics* no. (pp. 125–134).

Umeda, N. (1982). "F0 declination" is situation dependent. *Journal of Phonetics* 10, 279–290.

Wagner, M. (2002). The role of prosody in laryngeal neutralization. MIT Working Papers. *Linguistics* 42, 373–392.

Wang, B., Kügler, F., and Genzel, S. (2018a). Downstep effect and the interaction with focus and prosodic boundary in Mandarin Chinese Paper presented at the Tonal Aspects of Languages (TAL). Berlin, Germany.

Wang, B., and Xu, Y. (2011). Differential prosodic encoding of topic and focus in sentence-initial position in mandarin Chinese. *J. Phon.* 39, 595–611. doi: 10.1016/j.wocn.2011.03.006

Wang, B., Xu, Y., and Ding, Q. (2018b). Interactive prosodic marking of focus, boundary and newness in mandarin. *Phonetica* 75, 24–56. doi: 10.1159/000453082

Ward, I. C. (1952). *Introduction to the Yoruba Language*. Cambridge: W. Heffer & Sons Ltd.

Welmers, W. E. (1959). Tonemics, morphotonemics, and tonal morphemes. *General Linguist.* 4, 1–9.

Xu, Y. (1997). Contextual tonal variations in mandarin. *J. Phon.* 25, 61–83. doi: 10.1006/jpho.1996.0034

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. *J. Phon.* 27, 55–105. doi: 10.1006/jpho.1999.0086

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

Xu, Y., Chen, S.-w., and Wang, B. (2012). Prosodic focus with and without post-focus compression (PFC): A typological divide within the same language family? *The Linguist. Rev.* 29, 131–147. doi: 10.1515/tlr-2012-0006

Xu, Y., Prom-on, S., and Liu, F. (2022). "The PENTA model: Concepts, use and implications" in *Prosodic Theory and Practice*. eds. S. Shattuck-Hufnagel and J. Barnes (Cambridge: The MIT Press)

Xu, Y., and Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *The Journal of the Acoustical Society of America* 111, 1399–1413.

Xu, Y., and Wang, Q. E. (2001). Pitch targets and their realization: evidence from mandarin Chinese. *Speech Comm.* 33, 319–337. doi: 10.1016/S0167-6393(00)00063-7

Xu, Y., and Wang, M. L. (2009). Organizing syllables into groups—evidence from F0 and duration patterns in mandarin. *J. Phon.* 37, 502–520. doi: 10.1016/j.wocn.2009.08.003

Xu, Y., and Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *J. Phon.* 33, 159–197. doi: 10.1016/j.wocn.2004.11.001

Yuan, J., and Liberman, M. (2014). F0 declination in English and Mandarin Broadcast News Speech. *Speech Communication* 65, 67–74.

Zhang, L. (2017). Cantonese lexical tone and declination [in Chinese]. *Language sciences Yuyan Kexue* 2, 182–191.

Zhang, Z. (2016). Mechanics of human voice production and control. *The Journal of the Acoustical Society of America* 140, 2614–2635.

Zerbian, S., and Kügler, F. (2015). "Downstep in Tswana (Southern Bantu)." in *Paper Presented at the ICPhS*, Glasgow.

Zerbian, S., and Kügler, F. (2021). Sequences of high tones across word boundaries in Tswana. *J. Int. Phon. Assoc.* 1–22. doi: 10.1017/S0025100321000141

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 17 January 2023
DOI 10.3389/fpsyg.2022.911349

Internal structure of intonational categories: The (dis)appearance of a perceptual magnet effect

Joe Rodd[1] and Aoju Chen[2]*

[1]Office of Education, Ministry of Education, Culture and Science, The Hague, Netherlands, [2]Institute for Language Sciences, Utrecht University, Utrecht, Netherlands

The question of whether intonation events are speech categories like phonemes and lexical tones has long been a puzzle in prosodic research. In past work, researchers have studied categoricality of pitch accents and boundary tones by examining perceptual phenomena stemming from research on phoneme categories (i.e., intonation boundary effects—peaks in discrimination sensitivity at category boundaries, perceptual magnet effects—sensitivity minima near the best exemplar or prototype of a category). Both lines of research have yielded mixed results. However, boundary effects are not necessarily related to categoricality of speech. Using improved methodology, the present study examines whether pitch accents have domain-general internal structure of categories by testing the perceptual magnet effect. Perceived goodness and discriminability of re-synthesized productions of Dutch rising pitch accent (L*H) were evaluated by native speakers of Dutch in three experiments. The variation between these stimuli was quantified using a polynomial-parametric modeling approach. A perceptual magnet effect was detected: (1) rated "goodness" decreased as acoustic-perceptual distance relative to the prototype increased (Experiment 1), and (2) equally spaced items far from the prototype were more frequently discriminated than equally spaced items in the neighborhood of the prototype (Experiment 2). These results provide first evidence for internal structure of pitch accents, similar to that found in color and phoneme categories. However, the discrimination accuracy gathered here was lower than that reported for phonemes. The discrimination advantage in the neighborhood far from the prototype disappeared when participants were tested on a very large number of stimuli (Experiment 3), similar to findings on phonemes and different from findings for lexical tones in neutral network simulations of distributional learning. These results suggest a more transient nature of the perceptual magnet effect in the perception of pitch accents and arguably weaker categoricality of pitch accents, compared to that of phonemes and in particular of lexical tones.

KEYWORDS

intonational phonology, Autosegmental-metrical theory, pitch accent, categorical perception, internal structure, perceptual magnet effect, goodness rating, differential discrimination

OPEN ACCESS

EDITED BY
Annie Tremblay,
University of Kansas,
United States

REVIEWED BY
Matthew Lehet,
Michigan State University,
United States
Anja Kuschmann,
University of Strathclyde,
United Kingdom

*CORRESPONDENCE
Aoju Chen
aoju.chen@uu.nl

SPECIALTY SECTION
This article was submitted to
Auditory Cognitive Neuroscience,
a section of the journal
Frontiers in Psychology

RECEIVED 02 April 2022
ACCEPTED 07 November 2022
PUBLISHED 17 January 2023

CITATION
Rodd J and Chen A (2023) Internal structure of intonational categories: The (dis)appearance of a perceptual magnet effect.
Front. Psychol. 13:911349.
doi: 10.3389/fpsyg.2022.911349

Introduction

Intonational phonology concerns the mapping of phonetic-level variation in fundamental frequency (F0, also known as pitch in speech perception) to abstract units, which are then in turn mapped to meanings. The most widely accepted theory of intonational phonology, the autosegmental-metrical theory (hereafter AM theory), characterizes F0 or pitch (hereafter pitch) movement in terms of a series of high and low tones, organized sequentially (Pierrehumbert, 1980; Beckman and Pierrehumbert, 1986; Gussenhoven, 2004; Ladd, 2008; Arvaniti and Fletcher, 2020). These tones can either stand on their own as single tonal targets, or be combined into bi-tonal and tri-tonal targets. Such tonal targets can be of a lexical nature such as lexical tones in languages like Thai and Mandarin, and lexical pitch accents in languages like Tokyo Japanese and Stockholm Swedish, or of a post-lexical nature such as pitch accents in intonation languages like English, Dutch, and Italian. These tonal targets are aligned onto the segmental stream, and are organized into a phrasal structure of intermediate phrases within intonational phrases. Each type of phrase additionally potentially carries a boundary tone marking the right edge of that phrase. Just as in segmental phonology, phonetic realization rules govern the transformation of this abstract representation of the melody into a realizable pitch contour and the temporal alignment of tones to the segmental stream. However, phonetic implementation is underspecified in intonation (Arvaniti, 2011), leaving room for phonetic variation.

A critical assumption of the AM theory is that pitch accents are discrete or phonological categories, similar to lexical tones and lexical pitch accents. However, this assumption has been the subject of continuous debate in the field of prosody, because pitch accents and lexical tones are different in several aspects. First, lexical tones are far more densely distributed than pitch accents because each syllable can be specified for a lexical tone, one syllable per word is specified for a lexical pitch accent, but only some words are realized with a post-lexical pitch accent (hereafter pitch accent) in an utterance (Arvaniti and Ladd, 2009, *cf.* Xu et al., 2015). Second, pitch accents are more difficult to establish than lexical tones, because a meaning difference suffices to tell two lexical tones apart but is no sufficient to determine whether two pitch contours are from the same category or two distinct categories (Arvaniti and Fletcher, 2020). Third, the functional difference between pitch accents and lexical tones has led to the claim that lexical tones are stored in the lexicon and may thus be more consistently and precisely represented in the prosodic system than pitch accents are (Hallé et al., 2004; Francis et al., 2008). Finally and probably most importantly, there is still no consensus on what should be taken as empirical evidence for or against the categoricality of pitch accents (Gussenhoven, 1999; Prieto, 2012).

The present study aims to contribute to a clearer understanding of categoricality of pitch accents from an understudied perspective. Because this line of research is deeply rooted in the methodology used in research on categoricality of phoneme categories, we will first briefly review two perceptual phenomena that are tested to support the categoricality of phonemes ("Categoricality of phoneme categories"), then offer a brief critical review of past research on categoricality of pitch accents following the methodology stemming from research on phoneme categories ("Past work on categoricality of pitch accents"), and finally outline our approach to categoricality of pitch accents and present our hypotheses and predictions (Section The current study).

Categoricality of phoneme categories

Categoricality of phoneme categories has been experimentally studied by testing two perceptual phenomena, i.e., discrimination sensitivity peaks at phonemic boundaries and poor discrimination sensitivity within phonemic boundaries, reaching minima near the best exemplars or prototype of a category. The former is known as categorical perception, typically established through the so-called categorical perception (CP) paradigm consisting of an identification task and a discrimination task (Liberman et al., 1957). However, this term is also strongly associated with a class of hypothesized mechanisms, which assume that phonemes are perceived in terms of phonemic categorization or phonemic labelling (see Iverson and Kuhl, 2000 for a brief review). To separate the perceptual phenomenon and its hypothesized mechanisms, we will use the term *phoneme boundary effect* (Wood, 1976) in this paper, following Iverson and Kuhl (2000). The phenomenon of minimum discrimination near the prototype of a phoneme category is known as the *perceptual magnet effect* (Kuhl, 1991; Davis and Kuhl 1994), established through perceptual goodness rating and discrimination tasks. These two perceptual effects appear to stem from different processes. More specifically, Iverson and Kuhl (2000) tested the perception of English /i/ and /e/ vowels in conditions differing in the range of stimuli presented in each block of stimuli (or context variance, following Macmillan et al., 1988). They found that in the condition with reduced context variance, the sensitivity peaks near vowel boundaries disappeared whereas the sensitivity minima remained. This finding was interpreted to mean that the phoneme boundary effect may arise from cognitive encoding strategies such as perceptual anchoring (Macmillan et al., 1988), but the perceptual magnet effect from auditory processing (Iverson and Kuhl, 2000).

Animal studies (e.g., Kuhl and Miller, 1975; Kuhl et al.,1978) and non-speech studies with humans (e.g., Kluender et al., 1988) have shown that the phoneme boundary effect is present in animals with no access to phonemic labels and in human listeners listening to non-speech stimuli. However, research on rhesus monkeys' perception of vowels has yielded no evidence for a perceptual magnet effect (Kuhl, 1991). Together with Iverson and Kuhl (2000), these findings suggest a lack of a direct link between the phoneme boundary effect and the presence of phoneme categories in listeners' mental representation. It is thus highly questionable to take evidence for a phoneme boundary effect as evidence for categoricality of phonemes.

In contrast, the presence of a perceptual magnet effect has been argued to reflect domain-general internal structure of categories (Kuhl, 1991; Lacerda, 1995). Every category has presumably an indefinite number of members or exemplars. Crucially, not all members are perceived to be good or representative exemplars of a category by listeners; members closer to best exemplars are harder to discriminate than members further away. Furthermore, the magnitude of the perceptual magnet effect in phonemic perception can be affected by individual differences in phonemic categorization and ability to label synthetic stimuli as good exemplars of phoneme categories (Iverson and Kuhl, 2000). For example, Aaltonen et al. (1997) found that listeners exhibited a perceptual magnet effect on the mismatch negativity measure only if they could consistently label their stimuli as /i/ or /y/ in Finnish. Iverson and Kuhl (2000) found that the perceptual magnet effect decreased for listeners who were less clear on which stimuli they perceived to be good exemplars of /r/ in English. Similarly, in Lively and Pisoni's (1997) study on the perception of /i/, their listeners showed considerably more variability in goodness ratings than has been reported in other studies of the same phoneme and exhibited no perceptual magnet effect.

Past work on categoricality of pitch accents

Over the past decades, researchers have primarily studied categoricality of pitch accents and boundary tones by examining an intonation boundary effect, the equivalent of the phoneme boundary effect, using a range of methods, such as the CP paradigm, a reaction time (RT) paradigm, and semantic identification (see Gussenhoven, 1999; Prieto, 2012; Gussenhoven and van de Ven, 2020 for reviews). Evidence for an intonation boundary effect has been at best inconsistent. For example, using the CP paradigm, Ladd and Morton (1997) examined the difference between a "normal" high and "emphatic" high pitch accent in English and found an identification boundary but no discrimination peak. When RT was measured during the identification task on a comparable stimuli set, slower reactions were found at the identification boundary, suggesting a categorical interpretation of peak height in English intonation (Chen, 2003). In Bari Italian, counter-expectational questions, narrow-and contrastive statements are all realized on L*H + L%, with varying peak heights. Savino and Grice (2011) combined the CP paradigm with the RT measurement and found that differences between the "question" meaning and either "statement" interpretation were perceived categorically, but the two "statement" meanings were not in Bari Italian. Similar findings were also reported for utterance-initial pitch peaks between (lower) statements and (higher) non-statements in Catalan (Prieto, 2004). Regarding peak alignment, Pierrehumbert and Steele (1989) used a repetition task to test for categoricality between English L* + H and L + H*. Their participants were asked to repeat stimuli from a continuum that varied in peak alignment in 20 ms steps. The repetitions fell into two categories, leading the authors to conclude that the peak alignment dimension was represented in a binary manner. However, using a CP-with-RT approach, Chen (2003) found no evidence of categorical perception on a similar stimulus continuum in British English.

However, for at least two reasons it is problematic to equate evidence for a boundary effect with evidence for the categoricality of pitch accents and conversely, to interpret a lack of evidence for a boundary effect as evidence against the categoricality of pitch accents. First, as discussed in "Categoricality of phoneme categories", a boundary effect or categorical perception is not necessarily related to categoricality of speech categories. Second, in experiments on intonation boundary effects, meaning attributes are used as the labels to access whether two intonational events are two distinct categories in the identification task. This adaption of the CP paradigm itself raises questions on whether intonational events are stored in the mental representation as speech categories independent of meaning attributes, like phonemes and lexical tones.

Compared to research on intonation boundary effects in the perception of pitch accents and boundary tones, there is much less research on perceptual magnet effects in the perception of postulated intonational categories. But existent work has similarly yielded mixed findings. For example, Schneider and Möbius (2005) found a perceptual magnet effect for the low boundary tone (L%) but not for the high boundary tone (H%) in German when stimuli were presented as isolated sentences, but in both boundary tone categories when each stimulus was preceded by a felicitous context (Schneider et al., 2009). In both studies, the boundary tones were varied on a one-dimensional continuum of pitch height at the end of the sentence. Moreover, the prototype and non-prototype of the boundary tones were determined on a semantic basis by asking listeners to rate each stimulus on how well it represented a statement or a question, different from the approach taken in studies on phoneme categories. Adopting Schneider and co-workers' methodology, Fivela (2012) tested for a perceptual magnet effect in H* + L and H* followed by a low phrase accent in Pisa Italian, where these accents serve a function of marking 'continuation/reintroduction' and 'correction/opposition' respectively within contrastive focus. Tokens of H* + L and H* followed by a low phrase accent were varied along a two-dimensional continuum (peak alignment, peak height). A perceptual magnet effect was found for H* + L but not for H*. In fact, for H*, discrimination was better in pairs in the vicinity of the prototype of H* than in pairs further away from it.

The perceptual magnet effect is reliant on a concept of acoustic-perceptual distance to define how far an exemplar is from the prototype of the category, and to define the spacing between pairs of items. This distance metric should be derived from quantification of the acoustic variation that causes change in category identity. In the segmental domain, this is relatively straightforward: formant frequencies characterize vowels, for instance, whilst voice onset time conveys the voicing distinction in stops. Intonation, as changes in pitch in time anchored to the segmental stream, is by definition multi-dimensional: changes in pitch scaling, peak-and valley alignment and accent duration all conceivably contribute to category identity. Furthermore, if intonational categories are like phoneme categories and have internal structure of categories, the goodness of a member as the prototype of a postulated intonational category should arguably be independent of meaning attributes. Small variation in pitch accents and boundary tones can convey subtle shades of meaning. A representative exemplar of a pitch accent to convey a certain meaning attribute may not be equally representative of that pitch accent as an abstract category in the acoustic sense. Hence, the question arises as to whether the multi-dimensional nature of intonation and the semantically-driven choice of prototypes may be the cause for absence of the perceptual magnet effect in Pisa Italian H* (Fivela, 2012) and the reliance of the perceptual magnet effect on the presence of felicitous contextual information in German boundary tones (Schneider et al., 2009).

The current study

In the current study, circumventing methodological limitations in previous research on the perceptual magnet effect in intonational categories, we aim to find out whether pitch accents can be considered speech categories by examining whether they have domain-general internal structure of categories. To this end, we adopted parametric modeling of intonation (Reichel, 2011; Walsh et al., 2013) to quantify variation in pitch accents along five dimensions (more on this in "General methodological issues"), and tested for the presence of a perceptual magnet effect in the L*H pitch accent on the Dutch one-word utterance Mi in three experiments ("Experiment 1: Goodness rating of resynthesized stimuli", "Experiment 2: Discrimination", and "Experiment 3: Discrimination in a within-subject design"). The L*H pitch accent was selected because it was one of two pitch accents of which productions were systematically collected and analyzed by Chen et al. (2014). Testing L*H before the other pitch accent (i.e., H*L) is desirable because of the increased variability in shape in H*L compared to alignment in L*H found by Caspers and van Heuven (1993); if there is more variability in the shape, that implies that the perceptual space defined by CoPaSul parameters will similarly be larger, making a perceptual magnet effect easier to detect.

We hypothesize that pitch accents have internal structure, in the same way that phonemes do. If this hypothesis is true, we predict first that stimuli closer to the prototype of L*H will receive higher goodness ratings than those further away from it, and second that discrimination accuracy will be worse in stimuli closer to the prototype than in stimuli further away from it. We refer to these two predictions as the 'gradient goodness' symptom and the 'differential discriminability' symptom, respectively.

General methodological issues

Modeling approach

We adapted the CoPaSul (contour, parametric, and superpositional) intonation model (Reichel, 2011) to quantify pitch accent variation. CoPaSul models a linear global declination contour in the domain of the intonational phrase, then uses a series of parametrically defined third-order polynomial functions to stylize the residual movement in the domain of the accent group. We adapted CoPaSul in two ways. First, we removed the global contour, which models declination in connected speech, and was not relevant in our isolated stimuli. Second, we substituted CoPaSul's natural polynomials for orthogonal polynomials. Natural polynomials are mathematically straightforward in computation, but the parameters are by definition correlated with each other. This is undesirable for the purposes of this study, because pairs of parameters then have a highly correlated distribution, which complicates the generation of stimuli sets that vary predictably and evenly. Using Legendre orthogonal polynomials (Grabe et al., 2007) instead of natural polynomials solves this problem without worsening the quality of stylization. This adjustment resulted in a round rather than ovoid exemplar cloud, making the calculation of acoustic-perceptual distance between exemplars and the placement of referents to test more straightforward. For convenience, we refer to the resulting model as Simplified Orthogonal CoPaSul (SOCoPaSul).

SOCoPaSul characterizes different shapes of intonation contours in terms of four parameters: a parameter controlling the local pitch level (INTERCEPT), two inter-related parameters that control the rising or falling direction of the intonation contour and the peak alignment (CO1 and CO3), a parameter controlling peak shape, from convex to concave (CO2), as shown in Figure 1. The interactions between parameter values create more complex shapes. To complete our characterization of the prosodic properties of each pitch accent exemplar, we also added its duration as a fifth metric, to capture the interaction of duration with the other parameters. The exemplars of each pitch accent were thus modeled in a five-dimensional space in SOCoPaSul. The acoustic-perceptual distance between two exemplars was the Euclidean distance in this five-dimensional space.

The stimuli

The stimuli were generated in five steps.

Step 1: Selecting the prototype

We selected the prototype of L*H in Dutch using recordings from Chen et al. (2014). Adopting Caspers's (2000) elicitation method, Chen et al. (2014) studied the realisation of L*H and H*L in Dutch and their equivalents in Mandarin Chinese, i.e., the rising and falling tones (Tone 2 and Tone 4).[1] Caspers (2000)

[1] Chen et al. (2014) was conducted as part of a larger project on the subcortical processing of pitch in Dutch and Mandarin Chinese. In this type of research, tokens of /mi/ are frequently used as stimuli (e.g., Wong et al., 2007), because it contains only sonorant segments and it is long enough to realize different pitch patterns, including Mandarin lexical tones, but it is not too long for using EEG to track the frequency-following response of the brainstem. For this reason, we used Mi spoken as a proper name with L*H H% and its resynthesized renditions as stimuli in this study.

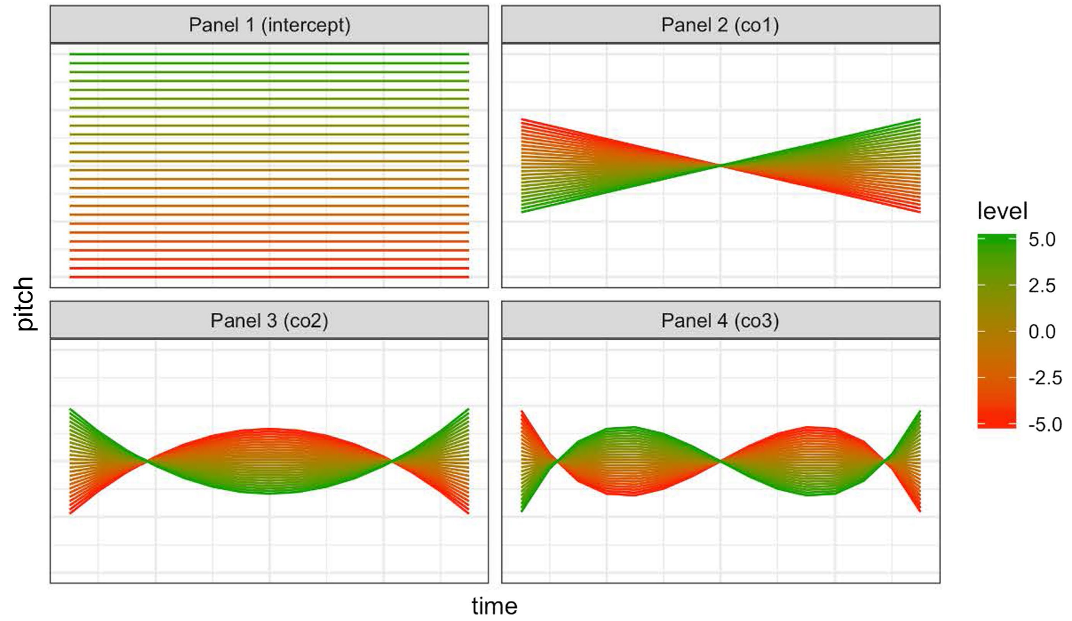


FIGURE 1
Parameters in SOCoPaSul. This figure shows, in each of the four panels, the effect on the contour shape of changing each parameter from a high value (green) to a low value (red), whilst holding the values of all the other parameters constant. Panel one depicts variation of the intercept (pitch level), panel 2 and panel 4 depicts variation in the rising or falling direction of the intonation contour and the peak alignment (CO1 and CO3), panel 3 controls peak shape, from convex to concave (CO2).

designed 24 'situational contexts' to elicit renditions of four signal-accent intonation patterns (i.e., an accent-lending rise, an accent-lending fall, an accent-lending rise and fall on one syllable, and an accent-lending rise and a half fall on one syllable) realised on proper names in three contexts from both the 'default' and 'vocative' perspectives (referring to the referent vs. addressing the referent directly). Her accent-lending rise was analysed as L*H H% or low rise according to Gussenhoven (1984, 2002, 2004). The accent L*H is associated with the meaning 'testing' in Gussenhoven's model. Chen et al. (2014) used three of Casper's situational contexts for the meaning 'testing' to elicit the proper name *Mi* in L*H H% from the default perspective. Their pilot experiment with three native speakers of Dutch confirmed that these contexts could indeed consistently elicit this low-rise contour.

For the current purpose, we tested the prototypicality of instances of L*H (followed by H%) realised on Mi with or without the original context in a perception experiment. In this experiment, native speakers of Dutch ($n = 5$, 5 females, mean age: 24;8, SD = 3;0) listened to 210 instances of Mi spoken with L*H in three situational contexts by seven speakers, half of them in isolation and the other half in the original situational context, intersected by instances of Mi spoken with H*L (followed by L%), and rated how good the production of the rising pattern was. Prior to the experiment, they were told that the researchers would like to find out what a typical Dutch rising pattern should sound like. They conducted the rating on a 7-point equal-appearing interval scale from 'bad production of a rising intonation' to 'good production of a rising intonation' using a computer program which allowed them to listen to each recording up to three times. The participants were moderately consistent in their ratings (Cronbach's $\alpha = 0.59$), but they all rated the instance of L*H that was the overall favourite highly (≥ 6). The instances of L*H were on the average slightly but statistically significantly higher rated when presented with the context than without the context (mean = 5.23, SD = 1.733 in the context condition; mean = 4.84, SD = 1.886 in the no-context condition; $t = 4.419$, $df = 524$, two-tailed: $p < 0.001$). However, the instance that was rated the highest (mean = 6.44) in the isolation condition was also rated the highest in the context condition (mean = 6.44). This instance of L*H was then selected as the prototype of L*H (Figure 2). It was produced in Caspers's (2000) 'default testing' context D1B:

Je neemt deel aan een docentenvergadering. Er moet een leerling worden benoemd in het schoolbestuur. Een aantal kandidaten wordt geopperd door je collega's en je hebt zelf iemand in gedachten waarvan je absoluut niet weet hoe die persoon zal vallen bij de rest; je doet een voorzichtige suggestie: Mi.

[You are attending a staff meeting. A pupil has to be appointed to the school administration. A number of candidates are put forward by your colleagues and you yourself have someone in mind of whom you are absolutely unsure whether that person will be acceptable to the others; you offer a tentative suggestion:]

Note that according to Gussenhoven (2004, p. 299) '… it (is) hard to discern any meaning difference between the high rise (H* H%) and the low rise'. This suggests that the pitch accent of the rising patterns elicited in Caspers's (2000) D1A context could also be H* based on the meaning it was supposed to convey. However, Gussenhoven's (2004, p. 288) description of how the high rise and the low rise should be realised in monosyllabic words and our close inspection of examples of the low rise in the online course on ToDI (Gussenhoven, 2005) suggest that the shape of the rising pattern in our selected instance (Figure 2; and the resynthesized instances of the prototype, see the gold-colored patterns in Figure 3) is comparable to that of L*H followed by H%, not comparable to that of H* followed by H% or any other rising nuclear contours (e.g., L*H %, H* %).

Step 2: Extraction of contours and quantifying natural productions using SOCoPaSul

The ProsodyPro Praat script (Xu, 2013) was used to extract the time-normalized pitch contour of each of the tokens in Chen et al.'s (2014) dataset. The SOCoPaSul polynomial model was then fitted to each normalized curve. This gave a cloud of values from naturally produced tokens of L*H for all five SOCoPaSul dimensions. The deviation around the prototype was calculated for each of the dimensions of these tokens.

Step 3: Selecting non-prototype referents

From near the edge of the cloud of natural productions of L*H, we selected two points in space to serve as potential non-prototype referents (i.e., the inside-limit non-prototype referents). They were placed at different points in the co1-co2 plane, but at the same distance from the prototype. The other parameters were kept constant at 0. These two inside-limit non-prototype referents are shown as the blue and purple contours in Figure 3.

We additionally created two further referent points out of the range of the natural productions of L*H (i.e., the outside-limit non-prototype referents, see the green and pink contours in Figure 3) by alternating the intercept to give the pink referent the same pitch register as the blue set, and the green referent the same register as the purple referent. The green and pink referents therefore differed from the prototype by the same amount as the blue and purple referents in co1-co2-co3 space, but were further in co1–co2–co3-intercept space. This allowed the testing of the impact of the inclusion of the intercept, and the testing of the impact of different pitch registers whilst holding the size of the excursion and the valley alignment constant. The pitch register was defined as the mean pitch of the first three "time-points" of the contour. The time points were 15 equally spaced points in the temporal dimension, which defined the pitch for the purposes of the manipulation. So in an item with a longer duration, the first three time-points were slightly longer than in an item with a shorter duration. Defining the absolute register in this way was a deliberate choice, because it meant that items that were identical other than their duration scaling received the same value for pitch register. The excursion size was defined as the difference between the highest pitch in the contour and the pitch register. The valley alignment was defined as the number of time points in which the curve that remained within 15 Hz of the pitch at the first time-point.

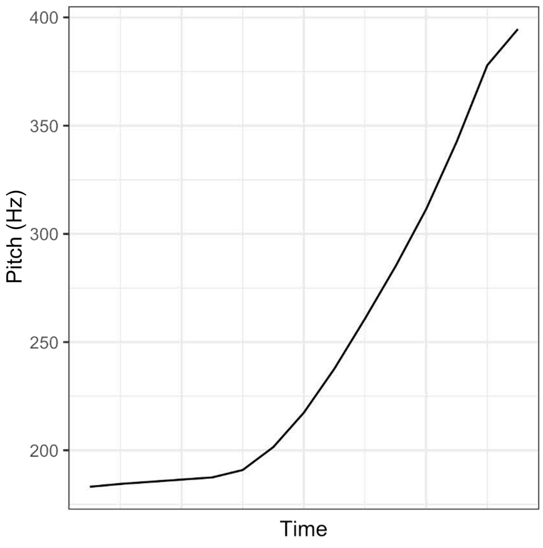


FIGURE 2
The prototype of L*H in the Dutch one-word utterance Mi.

Step 4: Creating neighboring exemplars for each referent

Around each referent, we created a pattern of "neighboring" points (shown as the blurred contours centering each contour in the left panel of Figure 3), arranged in a star-burst pattern, so that there were neighbors that were close to the referent, and neighbors that were further from the referent. Two differently-sized star-burst patterns were defined. The values of each parameter in each star-burst pattern were defined in z-scores. The origin was at point (0,0,0,0,0). In the large start-burst pattern, the first orbit had a radius of 0.3 standard deviations from the referent, the second orbit a radius of 0.6, the third 0.9, the fourth 1.2, the fifth 1.5 and the outermost orbit had a radius of 1.8 standard deviations. The smaller star-burst pattern consisted of the two inner-most orbits of the larger star-burst pattern, those with radii of 0.3 and 0.6 standard deviations. Each point in the five-dimensional space represented a stimulus. Its coordinates represented the parameters that describe it: (intercept, co1, co2, co3, duration). We used the smaller star-burst pattern to create two orbits of neighbors around the prototype referent (used in Experiments 2 and 3) and the larger star-burst pattern to create six orbits of neighbors around the prototype referent (used in Experiment 1) and around each of the non-prototype referent (used in Experiments 1, 2 and 3).

Step 5: Resynthesizing the prototype

Target pitch levels for each time-point for each stimulus were calculated using R Statistical Software (R Core Team, 2015) by applying the SOCoPaSul parameters (the coefficients and the

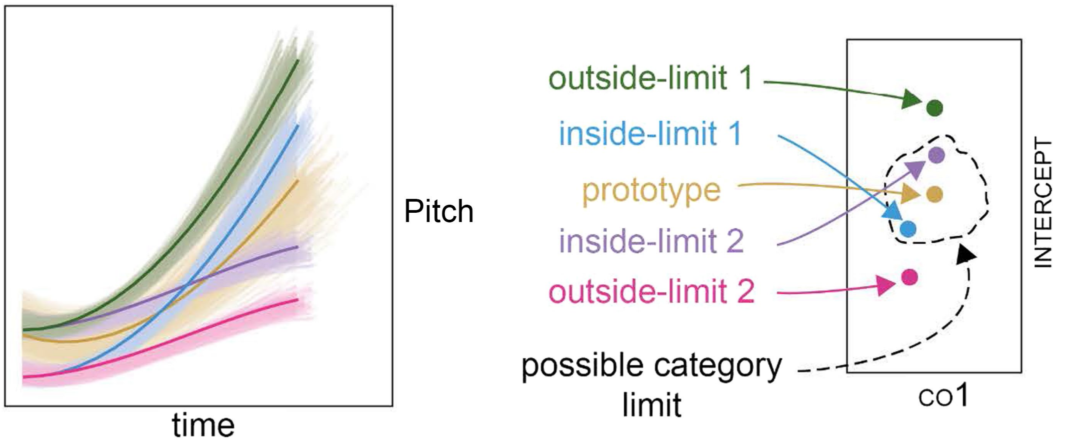


FIGURE 3
The contour shapes for the referents and neighboring exemplars; gold-prototype referent and the corresponding prototype referent set, blue and purple-two inside-limit non-prototype referents and the corresponding non-prototype referent sets; green and pink-two outside-limit non-prototype referents and the corresponding non-prototype referent sets.

intercept) to the polynomial function. Then, gating criteria were applied to ensure that all the synthesized pitch contours would be interpreted as L*H accents. The criteria were that there must be a low plateau of at least 40 ms at the beginning of the contour, where the maximum rise during the plateau was 8 Hz. These criteria were arrived at through informal investigation of the relevant just noticeable differences in conducting ToDI annotation (Gussenhoven, 2005).

After the entire process, the prototype neighborhood created using the bigger star-burst (hereafter the prototype referent set) contained 967 items (see the gold-colored blurred contours in the left panel of Figure 3). Each of the non-prototype neighborhoods (hereafter the inside-limit or outside-limit non-prototype referent sets; see the blue, purple, green and pink-colored blurred contours in the left panel Figure 3) and the prototype neighborhood created using the smaller star-burst (hereafter the near prototype referent set) contained approximately 250 items. A random sample of 150 items was made from each set. Each individual stimulus was created by re-synthesizing the prototype *via* a scripted process, using PSOLA implementation in Praat (Boersma and Weenink, 2012). This gave a separate sound file for each stimulus that consisted of the prototype with the pitch replaced with a curve described by the SOCoPaSul parameters. The inputs to the script to create each individual stimulus were the pitch contours calculated in step 4, and the degree of duration difference between the prototype and the stimulus was calculated.

Experiment 1: Goodness rating of resynthesized stimuli

To test for the 'gradient goodness' symptom of the perceptual magnet effect, a goodness rating experiment was conducted. Participants listened to the stimuli over headphones, and gave ratings on a five point equal-appearing interval scale from 'bad example' to 'good example' of Mi spoken with a rising melody.

Participants and materials

Ten native speakers of Dutch (6 females, mean age: 22;2) took part in this experiment. They were students at Utrecht University at the time of testing. All participants rated the prototype referent set. They each additionally rated one of the non-prototype referent sets (two participants each for the inside-limit sets, three participants each for the outside-limit sets). This meant that each participant rated 1,117 items in total.

Procedure

The experiment was conducted in a web browser using the jsPsych library (de Leeuw, 2015) and the Django Python web application framework (Holovaty and Jacob Kaplan, 2009). It took place in a quiet classroom equipped as a language lab with computers and good quality headphones.

The participants were instructed by means of a slide presentation (also implemented in a web-browser) that they read under the supervision of the experimenter. The key instruction was "determine how typical the rising melody of each example sounds in Dutch." After the participants were instructed, they did six practice trials to familiarize themselves with the experimental task. The practice trials used the prototype, two items from near the prototype and two items from far from the prototype to familiarize the participants with the extent of variation in the dataset. The presentation order of the items assigned to each participant were randomized by the computer, meaning items from the prototype referent set and from the other non-prototype referent sets were mixed together. Each trial began with the presentation of one item over the headphones. The participants used the mouse to select a rating on a five-point equal-appearing-interval scale, with labels 'slecht voorbeeld' (bad example) and 'goed voorbeeld' (good example). They could click multiple times to adjust their evaluation if they wished, and listen up to twice additionally to the stimulus by clicking the 'luister' (listen) button. When they were happy with the evaluation they had assigned, they clicked, the volgende, (next) button to proceed to the next trial. The interface used is depicted in Figure 4.

The participants completed the rating task in one 1-h appointment and one 40-min appointment on sequential days. The task was broken into blocks of approximately 20 min, with three blocks on the first appointment, and two blocks on the second. There was a mandatory 5 minute break between blocks. After the final block, the participants performed an unrelated task for another study.

Statistical analysis and results

To prepare data for analysis, each participant's scores were *z*-normalized, removing variation caused by different participants using subtly different anchors in their scales. This resulted in ratings that vary around 0 (the mean of a participant's scores), with positive evaluations being rated above 0 and negative evaluations below. The ratings given by the participants in the prototype referent set were moderately consistent (standardized Cronbach's $\alpha = 0.59$).

To test for gradient goodness, a mixed-effects linear regression model was first fitted using R Statistical Software (R Core Team, 2015) and the package lme4 (Bates et al., 2015) that predicted the mean normalized rating awarded to items in the prototype referent set by distance constructed as the Euclidean distance from the prototype to the item in (co1, co2, co3, intercept, and duration) space. Additional models were built to find out whether the fully specified model could be improved upon by removing some parameters from the calculation of the Euclidean distance. This was done *via* an "all-subsets" approach, by which all plausible models were constructed and then evaluated. Besides, we constructed and tested models using the "naturalistic" metrics typically used in the literature to characterize phonetic realization of pitch accents, i.e., the pitch register, the excursion size, the valley alignment and the duration.

We found that none of the "naturalistic" models using the conventional metrics of pitch accent variation account for the variation in rating as successfully as the best of the models incorporating the SOCoPaSul parameters (Supplementary Table 1), confirming our choice of using parametric modeling to quantify variations in the realization of a pitch accent. Notably, the model that excluded co3 outperformed the fully-specified model. As is depicted in Figure 1, the parameter co3 is the degree of influence that the cubic function contributes to the overall shape. Since the cubic function is sinusoidal, this parameter can be considered to control the degree of deviation from the overall curve at the extremities of the contour, adjusting the flatness of the plateaus at each end of the pitch accent.

Both the fully specified model ($r = -0.585$, $p < 0.01$) and the best-fitting model ($r = -0.609$, $p < 0.01$) clearly indicated a negative relationship between the distance of a token to the prototype and the goodness rating. The best-fitting model is depicted in Figure 5 and further reported in Table 1. As can be seen in Figure 5, the items closer to the prototype received significantly higher ratings than those further from it and the items from the prototype referent set received by and large the highest ratings, providing evidence for the gradient goodness symptom of the perceptual magnet effect.

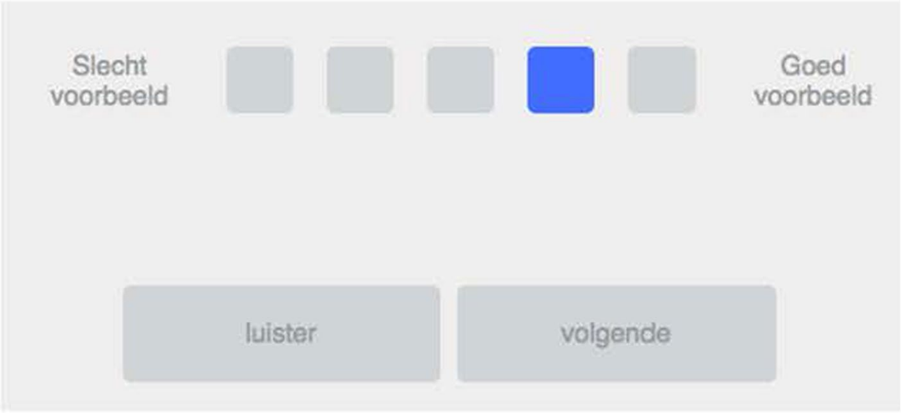


FIGURE 4
The interface used by the participants to input their ratings, with a rating of four selected, but not yet submitted.

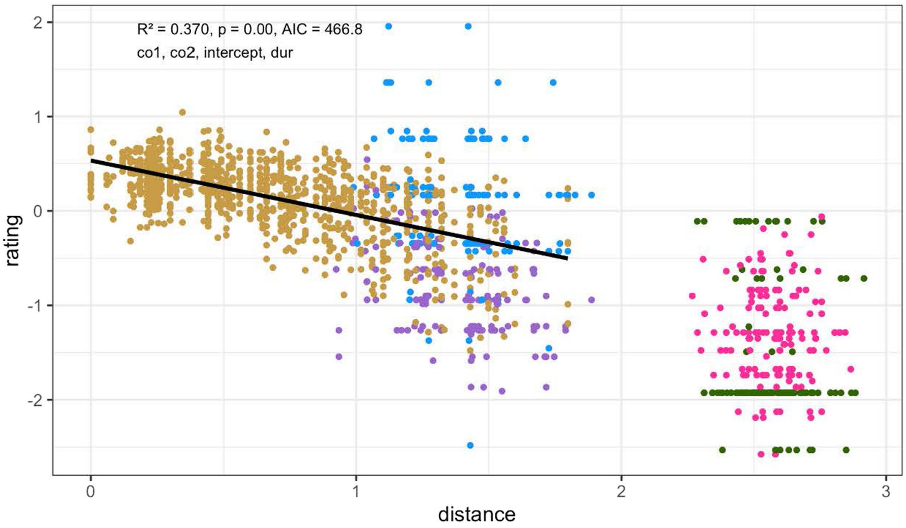


FIGURE 5
The model that explains the most variation characterizes the distance from the prototype in (co1, co2, intercept, and duration) space. The model is fitted on the 'goodness' dataset only, which is colored gold. The other colors represent the ratings on the non-prototype referents and neighboring exemplars. The *x*-axis depicts the distance between a rendition of L*H and the prototype of L*H (the '0' point). The *y*-axis shows the *z*-normalized scores of the goodness ratings, with 0 being the mean of a rater's scores, positive evaluations being rated above 0 and negative evaluations below 0.

TABLE 1 Summary of the best-fitting model for the goodness ratings of the prototype and its referent set.

	Estimate	Std. Error	*t* Value	Pr(>|t|)
(Intercept)	0.533	0.019	28.121	< 0.01
Euclidean distance (co1, co2, intercept, duration)	−0.576	0.024	−23.844	< 0.01

Experiment 2: Discrimination

An AB discrimination paradigm was employed to test for differential discrimination, presenting the tokens in pairs and asking participants to assess whether or not they heard a difference, similar to Fivela (2012) and Schneider and Möbius (2005). The aim of the task was to establish whether participants were able to detect the difference between a reference sound and a comparison sound.

Participants and materials

Fifteen native speakers of Dutch (12 females, mean: 21;11) took part in this experiment. They were students at Utrecht University at the time of testing and did not take part in Experiment 1.

All participants were tested on the near prototype referent set and on one of the four non-prototype referent sets in a single session.[2] As described in "The stimuli," each set contained 150 items, which were a random sample (the same for all participants assigned to that set) from the 250 items in each referent set. Thus, for each participant, there were 300 test trials (where the comparison sound differed from the reference sound). In the experimental trials, the reference sound in each stimulus pair was either the prototype referent or the non-prototype referent, and the comparison sound was a neighboring exemplar of these referents. In addition, for each type of reference sound, 30 control trials, where the reference sound was played twice, were included to assess the probability of false positives. Each participant was therefore tested on 360 trials. In each set, 75 of the 150 test trials used items taken from the first orbit of the star-burst pattern, 75 used items taken from the second orbit of the star-burst pattern. This means that, besides the control items where there was no difference between the reference sound and comparison sound, there were two levels of difference: small difference (the first orbit) and moderate difference (the second orbit).

Procedure

The experiment took place in the same quiet classroom setting as Experiment 1, equipped as a language lab with computers and good quality headphones.

[2] The prototype-referent set was not used in Experiment 2 because it contained far more items than the non-prototype-referent sets.

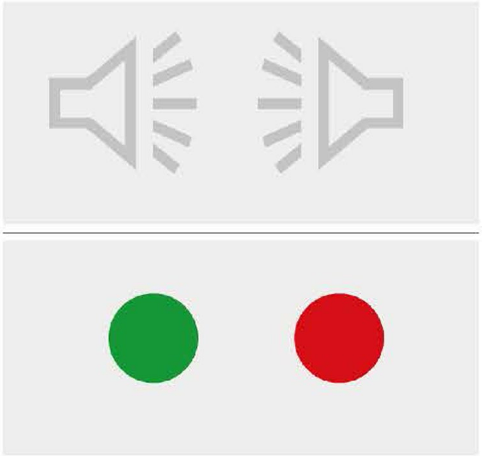


FIGURE 6
The graphics presented during the discrimination experiment. Upper panel: during playback. Lower panel: signal to respond.

The participants were divided into four groups, each of which were tested using a different non-prototype referent set. There were four participants in all groups except that tested with the green-colored outside-limit non-prototype referent set in Figure 3, which had three participants. Each participant was tested on pairs of items consisting of: (1) the reference sound and a comparison sound taken from the neighbors closest to the referent (small difference test trials, 41%), or (2) the reference sound and a comparison sound taken from the neighbors slightly further from the referent (moderate difference test trials, 41%) or (3) the reference sound repeated (control trials, 18%). Two blocks were conducted, one where the reference sound was the prototype (180 trials) and one where the reference sound was one of the four non-prototypes (180 trials). Block order and presentation order within each block were counterbalanced. Whether the target item appeared before the reference sound (AB order) or after the reference sound (BA order) was counterbalanced across participants within each group, so that the items that appeared in AB order for one participant appeared in BA order for the next (and *vice-versa*). The same software packages (jsPsych and Django) were used to implement this experiment as were used in Experiment 1.

The participants were instructed by means of a slide presentation that they read under the supervision of the experimenter. The key instruction was "determine whether you hear a difference between the two examples." A keyboard was labelled with a red sticker reading nee "no" on the M key, and a green sticker reading ja "yes" on the Z key. These keys were selected to force the participant to use both hands. The participants were instructed to press the "yes" key if they heard a difference, and to press the "no" key if they did not. On screen, whilst the two sounds were presented, a graphic of two speakers was presented (upper panel in Figure 6). After the offset of the second sound, this was replaced with a depiction of the green and red buttons, as signal to respond (lower panel in Figure 6).

Six practice trials were conducted under the supervision of the experimenter, using items from the prototype referent set that were not selected in the sample of experimental and control trials. These trials represented two control trials, two test trials from the "small difference" condition, and two test trials from the "moderate difference" condition.

The participants completed the task in one block of approximately 30 min. After the experiment, the participants performed an unrelated task for another study.

Statistical analysis and results

Generalization

The participants' discrimination responses on the test trials were coded as 'generalized' if they failed to detect a difference, or 'not generalized' if they succeeded in detecting a difference, following Kuhl (1991). As shown in Figure 7, there were more generalized trials in the near prototype condition than in the non-prototype condition in the two groups of participants tested on one of the within-limit non-prototype referent set. But the opposite pattern occurred in the two groups of participants tested on the outside-limit non-prototype referent set: greater generalisation in the non-prototype referent condition than the near prototype referent condition. Furthermore, increasing the difference (from + to ++ in Figure 7) between the comparison sound and the referent sound reduced generalisation for three of the four groups of participants. The participants thus appeared to perform in line with the predictions deriving from the differential discrimination symptom when tested on the within-limit non-prototype referent stimuli in addition to the near prototype referent stimuli. However, the participants differed substantially in their rate of generalisation in the near prototype referent stimuli, which we would expect to be consistent across groups. This implies that there were notable individual differences in the participants' performance, relative to which non-prototype stimuli were presented to them. This was subsequently confirmed when we plotted the differences between generalisation rates in the near prototype and non-prototype referent conditions, at the participant level (Supplementary Figure 1).

To take individual differences into account, we subsequently conducted mixed-effects binary logistic regression on the whole dataset using R Statistical Software (R Core Team, 2015) and the package lme4 (Bates et al., 2015). The outcome variable was binary (generalized, coded 1, or not generalized, coded 0). The fixed factors were the prototypical status of the reference sound (prototype and non-prototype) and distance of the comparison sound from the reference sound (small and moderate), the random factor was participant nested within group. We began with the random effects model, where only random factors were included. The models with each of the fixed effects on their own, both fixed effects combined, and both fixed effects and their

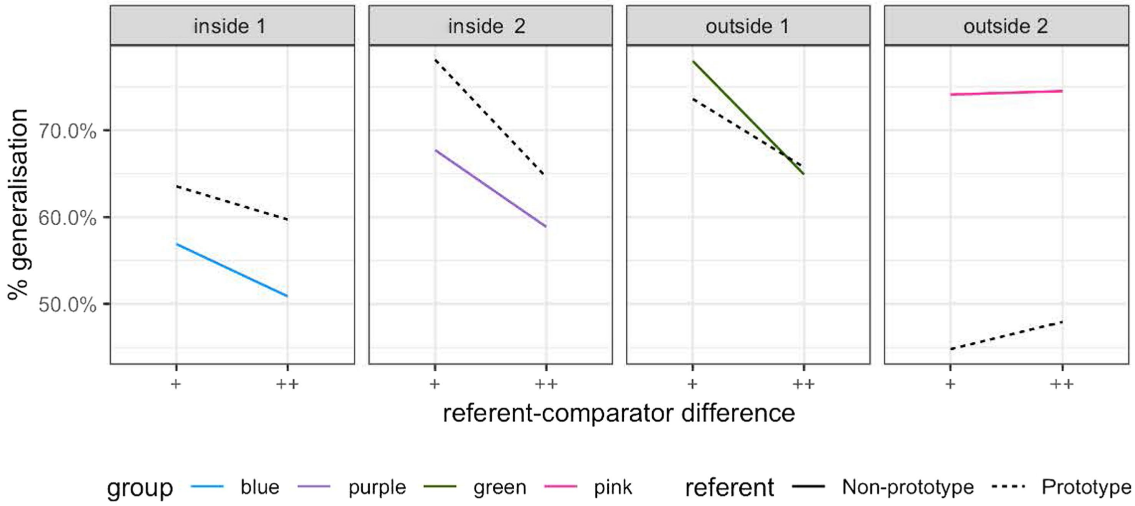


FIGURE 7
Generalization (misses) in the prototype (dashed) and non-prototype (solid) conditions in Experiment 1.

interaction, were tested in a stepwise fashion. After each iteration, fixed effects that did not represent a significant improvement over the previous model (as assessed by comparing the Bayesian information criterion) were excluded. The model with the best fit contained both fixed effects, but excluded their interaction. The criteria used to compare the candidate models are presented in Table 2, and the model with the best fit is shown in Table 3. The model demonstrates that when the trial was in the non-prototype referent condition, there was a significantly smaller chance of generalization, after controlling for variation between participants, than in the near prototype referent condition ($p < 0.01$). For a trial in the near prototype referent condition, moving from the baseline (small difference) in the difficulty dimension to moderate difference (that is, making the trial "easier") actually increased generalization. This finding was rather unexpected and contra what emerged at the group level (Figure 7) and will be revisited in the "General discussion" section.

Response accuracy

Because the result on generalization was coupled with notable differences in performance between the participants, we decided to conduct an explorative analysis on response accuracy in detail. Kuhl (1991) observed in her data that the response accuracy was substantially larger in the non-prototype referent condition than in the near prototype referent condition, in line with the pattern in generalization.

We used the mixed-effects logistic regression modeling technique applied to the generalization data to test for patterns in response accuracy, using R Statistical Software (R Core Team, 2015) and the package lme4 (Bates et al., 2015). In contrast to the models testing generalization, the control trials were included in these models. Therefore, the difficulty factor gained a third level, "no difference," which became the baseline. The same all-subsets procedure was used to generate and compare models.

The criteria used to compare the candidate models are presented in Table 4. As can be seen, the best fitting model included main effects of the factors prototypicality and difficulty, and the interaction of these two factors, in contrast to the generalization models (Table 5).

The main effect for prototypicality was such that the accuracy of trials in the non-prototype referent condition was significantly better than in the near prototype referent condition, in line with the effect of prototypicality on generalization. Difficulty had a surprising main effect: performance was significantly worse in the small difference and moderate difference conditions than in the no-difference condition. Intuitively, an increase in the distance between reference and comparison sound in each stimulus pair should result in improved performance, as the task becomes easier. That this is not the case suggests that the rejection of false positives may be inherently easier than detection of differences. The interactions were also significant; indicating that performance in trials that combined the non-prototype referent condition and the difference-detection task was significantly worse than would be predicted by the main effects alone.

Interim summary

The results of mixed effects logistic regression for generalization supports the presence of the differential discriminability symptom. In combination with the finding of gradient goodness in Experiment 1, these results indicate a perceptual magnet effect, and therefore evidence that L*H has internal structure. But we also observed notable individual variation in the discrimination of different groups of participants. Because these groups were

TABLE 2 The criteria used to compare the candidate models for the discrimination dataset in Experiment 2.

Fixed factors	Df	AIC	BIC	logLik	Deviance	Chisq	Chi Df	Pr(>Chisq)
(Only random factors)	4	5351.463	5377.111	−2671.732	5343.463	NA	NA	NA
Prototypicality	5	5333.660	5365.720	−2661.830	5323.660	19.803	1	< 0.01
Difficulty	5	5339.117	5371.176	−2664.558	5329.117	0.000	0	
Prototypicality + difficulty	6	5319.654	5358.125	−2653.827	5307.654	21.463	1	< 0.01
Prototypicality * difficulty	7	5321.502	5366.385	−2653.751	5307.502	0.152	1	0.696

TABLE 3 Overview of the best-fitting model for the discrimination dataset in Experiment 2.

	Estimate	Std. Error	*z* Value	Pr(>|z|)
(Intercept)	−0.570	0.243	−2.346	≤ 0.05
Prototypicalitynon-prototype	−0.316	0.068	−4.674	< 0.01
Difficultymoderate difference	0.271	0.068	4.010	< 0.01

presented with different sets of non-prototype stimuli, we conducted Experiment 3 using a within-subject design to find out whether such a design could mitigate individual variation in discrimination.

Experiment 3: Discrimination in a within-subject design

Using stimuli from Experiment 2, we tested 16 native speakers of Dutch (9 female, estimated mean age: 21 ~ 22 years)[3] for discrimination performance in all four non-prototype conditions. They did not participate in Experiment 1 and Experiment 2 but were otherwise comparable to the participants in the first two experiments.

The stimuli were presented in a blocked fashion. Four blocks consisted of trials where the sounds were taken from the four non-prototype referent sets. Two blocks consisted of trials where the sounds were taken from the near prototype referent set. Presentation order within each block was randomized, and the order of the blocks was pseudo-randomized such that the two identical near prototype referent blocks were separated by at least one other block. Each block took around 13 min, and the participants were obliged to take a three-minute pause and did a small paper-pencil based questionnaire on lexical semantics between each block to minimize participant fatigue and boredom.

In Experiment 2, a computer-equipped classroom was used, with multiple participants being tested simultaneously, using low-specification headphones. Experiment 3 was conducted in sound-isolated booths with high-specification headphones, providing the participants a much less distracting environment. Each testing session lasted about 120 min, including a practice session (see "Procedure" under the section Experiment 2: Discrimination).

Response accuracy

Figure 8 shows the percentage of trials with a correct response for each participant separately. As can be seen, only participant 18 clearly displayed the differential discrimination symptom of the perceptual magnet effect when we compared the average performance across the three different sorts of trials, i.e., those with no difference, those with a small difference, and those with a moderate difference between the reference sound and the comparison sound.

The absence of the differential discrimination symptom across the dataset is surprising given the result of Experiment 2. We thus checked for patterns introduced by the methodological changes between this experiment and Experiment 2. Specifically, in this experiment, all but three participants (due to technical problems) rated two blocks of sounds sampled from the near prototype neighborhoods. To rule out a possible training effect, we plotted the participants' performance for the two blocks of near prototype trials separately. As can be seen in Figure 9, there was a small performance difference between the first near prototype block and the second, but there was no clear pattern of better performance in the second session which would imply a training effect.

To make a more direct comparison with the data from Experiment 2, we plotted the participants' performance in the near prototype referent condition and inside-limit non-prototype referent conditions, excluding results from the second block of the near prototype referent condition (Supplementary Figure 2). This therefore simulated more closely the task of the participants in Experiment 2. But the results gathered were broadly similar to those obtained with two blocks of stimuli from the near prototype referent set (Figure 9).

λ-Center as a measure of discrimination performance

Schneider and Möbius (2005) and Schneider et al. (2009) used the λ-center metric to quantify participant success in discrimination tasks, instead of direct accuracy proportions

[3] Due to loss of information on the participants' age, we estimated their mean age based on the information on the participants in Experiment 1 and Experiment 2, who were recruited from the same student population and were comparable in academic background and age.

TABLE 4 The criteria used to compare the candidate models for the response accuracy dataset in Experiment 2.

Fixed factors	Df	AIC	BIC	logLik	Deviance	Chisq	Chi Df	Pr(>Chisq)
	4	6876.495	6902.872	−3434.248	6868.495	NA	NA	NA
\textsc{prototypicality}	5	6868.264	6901.235	−3429.132	6858.264	10.231	1	< 0.01
\textsc{difficulty}	6	6795.255	6834.820	−3391.627	6783.255	75.010	1	< 0.01
\textsc{prototypicality} + \textsc{difficulty}	7	6782.216	6828.375	−3384.108	6768.216	15.039	1	< 0.01
\textsc{prototypicality} + \textsc{difficulty} + \textsc{prototypicality:difficulty}	9	6778.082	6837.429	−3380.041	6760.082	8.134	2	≤ 0.05

TABLE 5 Summary of the best fitting model for the response accuracy dataset in Experiment 2.

	Estimate	Std. Error	*z* Value	Pr(>|z|)
(Intercept)	0.687	0.172	3.984	< 0.01
Prototypicalitynon-prototype	0.307	0.148	2.069	≤ 0.05
Difficultysmall difference	−1.192	0.119	−10.009	< 0.01
Difficultymoderate difference	−0.969	0.121	−7.973	< 0.01
Prototypicalitynon-prototype:difficultysmall difference	−0.617	0.173	−3.563	< 0.01
Prototypicalitynon-prototype:difficultymoderate difference	−0.572	0.175	−3.278	< 0.01

(generalization). The λ-center metric is a concept taken from Signal Detection Theory (Wickens, 2002) and seeks to equalize the performance of different listeners by quantifying the individual response criterion of each listener (i.e., the amount of difference between the two stimuli that the listener requires for them to report a difference). The λ-center is a quantification of that response criterion, using a Gaussian transformation of the proportions of correct hits vs. signal trains and false alarms vs. noise trials. Lower λ-center values indicate better discrimination performance.

Given that the generalization metric yielded no evidence for the differential discrimination symptom of the perceptual magnet effect in Experiment 3, we decided to analyze the data using the λ-center metric. The λ-center analysis detected better performance near the prototype referent than away from it when averaging across all other factors, against predictions of the perceptual magnet effect, as shown in Figure 10. However, these differences were not statistically significant (inside-limit non-prototype referent set vs. near prototype referent set: $F(1,29) = 2.57$ $p = 0.1197$, outside-limit non-prototype referent set vs. near prototype referent set: $F(1,29) = 1.62$ $p = 0.2135$). When examining the participants' responses during the first near prototype referent block, the difference between the near prototype referent condition and the non-prototype referent conditions appeared to be slightly more pronounced, with more successful discrimination in the neighborhood of the prototype. Nevertheless, the differences did not reach statistical significance (inside-limit non-prototype referent set vs. prototype: $F(1,29) = 3.21$ $p = 0.0834$, outside-limit non-prototype referent set vs. near prototype referent set: $F(1,29) = 2.2$ $p = 0.149$).

Interim summary

Unexpectedly, Experiment 3 failed to replicate the results of Experiment 2, in spite of the within-subject design and more favorable acoustical conditions, and arguably more sensitive analysis metric. The approximately equal performance on all conditions suggests that extensive exposure to stimuli with high context variance may influence listeners' discrimination within an intonational category, different from findings on the within-category discrimination of vowels (Iverson and Kuhl, 2000). We will revisit this finding in "General discussion".

General discussion

This study is concerned with the question whether intonational events are speech categories like phonemes and lexical tones. Categoricality of phoneme categories has been experimentally studied by testing discrimination sensitivity peaks at phonemic boundaries (the phoneme boundary effect) and poor discrimination sensitivity within phonemic boundaries, reaching minima near best exemplars of a category (the perceptual magnet effect). In past work, researchers have studied categoricality of pitch accents and boundary tones by examining an intonation boundary effect, the equivalent of the phoneme boundary effect, and to a lesser extent the perceptual magnet effect in the perception of intonational categories. Both lines of research have yielded mixed results. However, animal studies and research on humans using non-speech stimuli have shown that a boundary effect or categorical perception is not necessarily related to categoricality of speech categories. We have thus used improved methodology to examine whether pitch accents have domain-general internal structure of categories by testing the

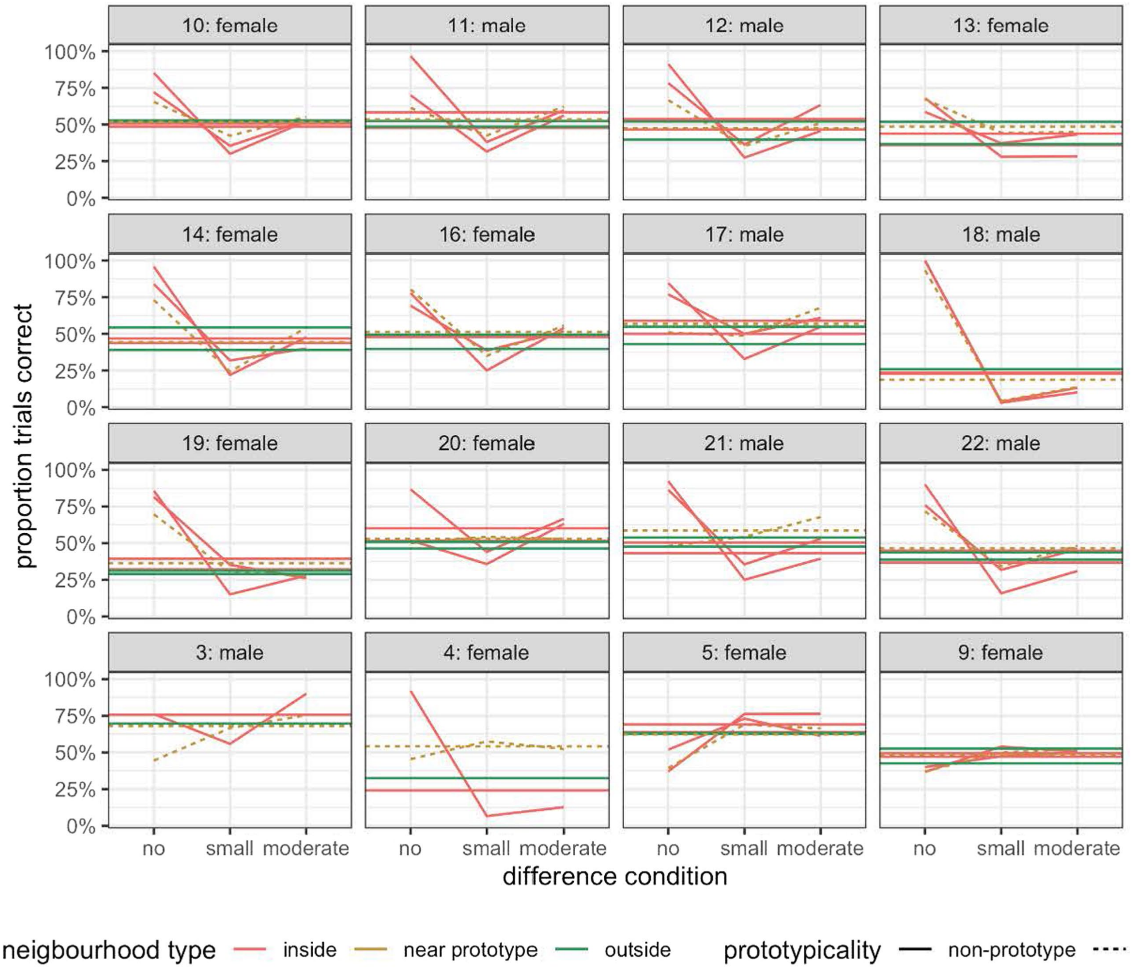


FIGURE 8
The percentage of trials with a correct response for each participant in Experiment 3.

two symptoms of the perceptual magnet effect in the perception of the Dutch L*H pitch accent: gradient goodness and differential discriminability.

Our results of the goodness rating (Experiment 1) demonstrate clearly that the gradient-goodness symptom of the perceptual magnet effect is present in the Dutch L*H pitch accent. The discrimination results of participants tested on the within-limit non-prototypes (Experiment 2) demonstrate that the differential discriminability symptom of the perceptual magnet effect is also present. Thus, the perceptual magnet effect is a feature of the Dutch L*H pitch accent, supporting its postulated categoricality in the phonology of Dutch intonation (Gussenhoven, 2005). This result has both theoretical and methodological implications for the debate on the phonological status of intonation. Theoretically, it suggests that the categoricality of intonational events may not be as controversial an issue as has been perceived on the basis of research examining the intonation boundary effect. Methodologically, it shows the potential of studying the categoricality of other pitch accents and boundary tones by examining the internal structure of the postulated category. Furthermore, that the model using the SOCoPaSul parameters to characterize perceptual distance was more successful than the model using the 'classic' quantifications of contour shape variation supports the view that there is merit in such a parametric approach to model intonation contours, and for interpreting the parameters as the dimensions of perceptual space.

However, the evidence gathered appears not to be as strong as that reported for phonemes. For example, the generalization rate was much higher in our study that for the discrimination of vowels (Kuhl, 1991), meaning that it might be considerably more difficult to detect differences between exemplars of pitch accents than exemplars of vowels. Another possibility is that the tone-shift technique used in Kuhl (1991) might be less demanding because it does not require participants to retain the first stimulus in memory for comparison with the second. Furthermore, the general discrimination accuracy was also much lower in this investigation than in Kuhl (1991) (here, in the order of 30–60% correct, rather than the accuracy rates of more than 75% in Kuhl's study). These differences in the degree of the perceptual magnet effect between intonational events and phonemes suggest that different types of speech categories may differ in the degree of categoricality.

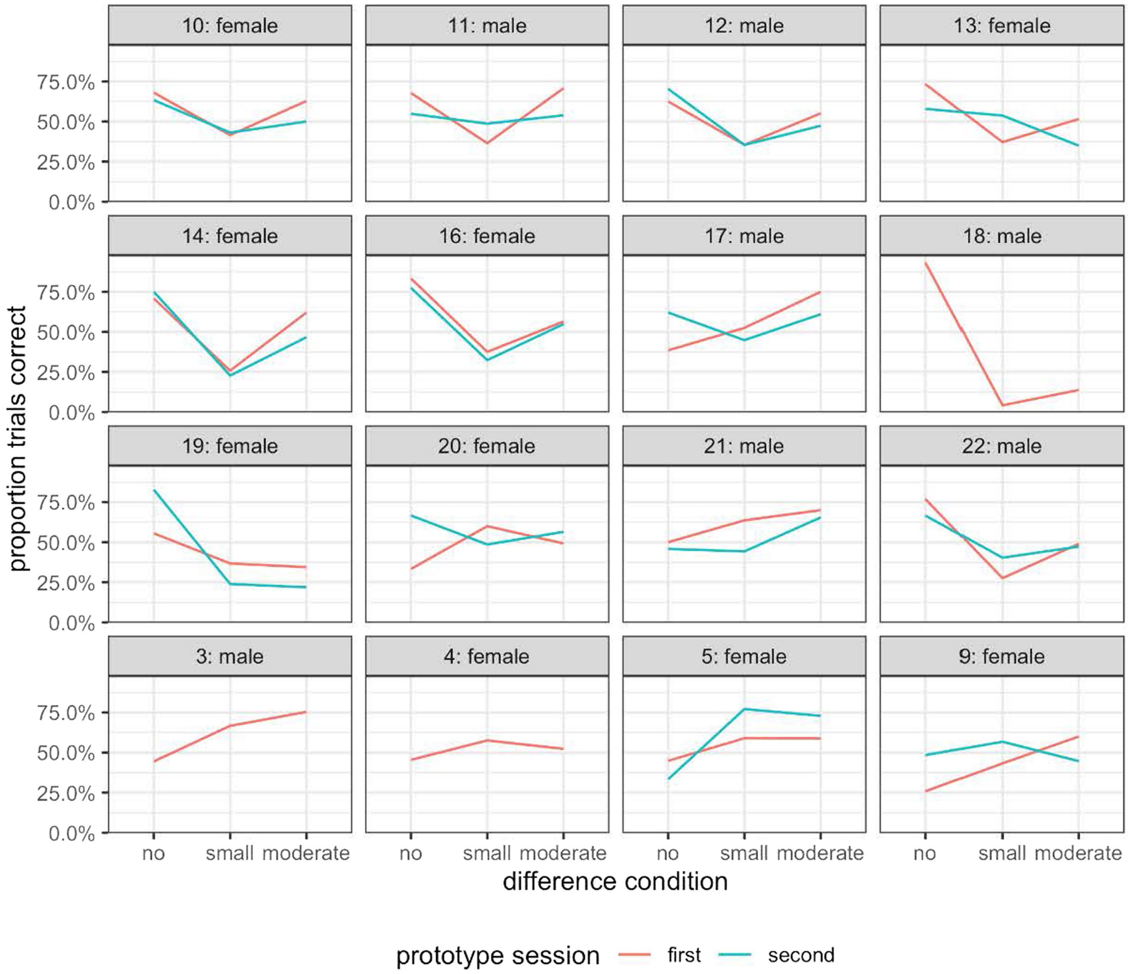


FIGURE 9
The percentage of trials with a correct response for each participant in the first and second half of the prototype-referent condition in Experiment 3.

The presence of a main effect for the factor difficulty (or acoustical distance between two stimuli in a pair) in the analysis on generalization in Experiment 2 calls for attention because it is in the opposite direction to that logically expected, i.e., more generalization in the presence of a larger acoustic distance. This finding is difficult to explain. It is perhaps the case that this pattern emerges because of the outside-limit non-prototypes. The participants in one of the outside-limit non-prototypes conditions (the 'green' condition) exhibited surprisingly low generalization in the small difference condition and greater generalization in the moderate difference condition. The small sample size and the between-subject design make it impossible to identify reasons why the participants in that group performed differently from the other groups, including the group in the other outside-limit non-prototypes conditions (the 'pink' condition). In future research, increasing the number of participants, including checks on factors that can potentially influence pitch perception such as musicality (Schön et al., 2004; Ong et al., 2020) may contribute to a clearer understanding of individual differences in task performance and possibly also mental representation of intonational events.

Experiment 3 was conducted to address the above-mentioned issues using a with-subject design. The result was rather unexpected. Instead of showing stronger evidence for differential discriminability, the participants showed no statistically significant differences in discrimination between the near prototype condition and the non-prototype condition. There are, however, some crucial procedural differences between Experiment 2 and Experiment 3. Namely, in Experiment 3, the participants were presented with many more stimuli (1,080 pairs of stimuli in Experiment 3 vs. 360 pairs of stimuli in Experiment 2) and tested in a much longer session (120 min in Experiment 3 vs. 30 min in Experiment 2) in sound-isolated booths with high-specification headphones. These differences raise the question whether the results were caused by a lack of engagement with the task in the participants. However, in a discrimination task, engagement can also mean that participants attend to subtle differences in a pair of stimuli and manage to discriminate to the same degree across pairs of stimuli, regardless of the acoustic distance between the two stimuli in each pair. This interpretation of participant

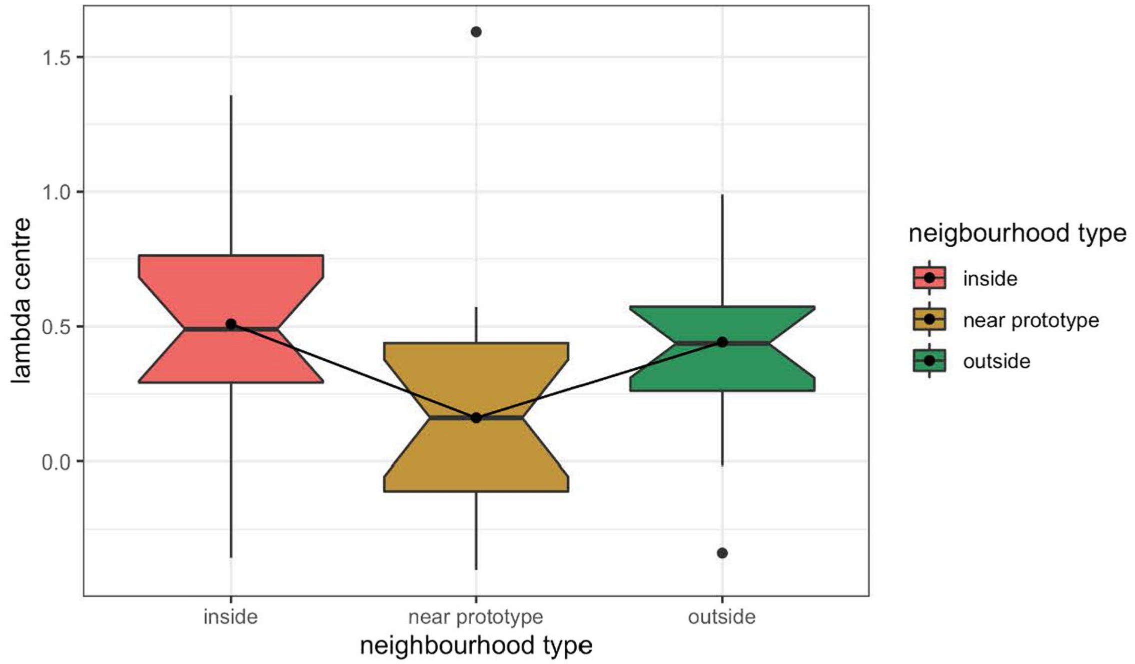


FIGURE 10
λ-Center for discrimination performance in the prototype referent condition (middle) and in the inside-limit non-prototype referent condition (left) and outside-limit non-prototype referent condition (right) in Experiment 3.

engagement appears to be in line with the data of Experiment 3. When we only plotted the participants' performance in the near prototype referent and inside-limit non-prototype referent conditions, simulating the task of the participants in Experiment 2, the results were broadly similar to the results based on the entire stimuli. This suggests that possible participant fatigue-triggered disengagement cannot explain the results of Experiment 3.

Our result may thus imply that very extensive exposure to a large number of exemplars of a hypothetical intonational category coupled with high context variance may influence listeners' discrimination within an intonational category. This in turn suggests that the state of perceptual magnet effect can be transient as a result of intensive exposure to high context variance. The question arising is whether it is specific to intonational events like pitch accents.

Neural network simulations of distributional learning shows that the transience of the perceptual magnet effect can also occur during the learning of phoneme categories. Using deep Boltzmann machines, Boersma (2019) studied the emergence of phoneme categories as a result of auditory-driven distributional learning of spectral content alone in a simulated first-language learner. He found that the stimulated learner showed a perceptual magnet behavior along a two-dimensional continuum (i.e., F1 and F2 of five vowels) after having listened to 1,000 pieces of data but this behavior faded away as more pieces of data were heard. However, the perceptual magnet behavior seems to be stable and insensitive to the amount of auditory exposure in simulated distributional learning of Mandarin lexical tones. Using the same neutral network simulation,[4] modeled the distributional learning of four Mandarin Chinese lexical tones on a three-dimensional continuum (onset pitch, medial pitch, offset pitch, pitch contour, sound-meaning mapping). They found that the simulated learner's perceptual magnet behavior was at its peak after having heard 1,000–1,500 pieces of data, started to decrease afterwards but stabilized after the presentation of 200,000 pieces of data. Together with these findings, findings from Experiment 3 posit a striking difference between pitch accents and phonemes as speech categories on the one hand and lexical tones on the other hand. Future experimental research on the perceptual magnet effects in the perception of lexical tones by native speakers in languages like Mandarin Chinese will be both valuable and necessary in order to attain a clearer understanding of perceptual magnet effects as a feature of tonal categories and the differences between pitch accents and lexical tones. Further research is also needed to tease apart the influence of extensive auditory exposure and high context variance on listeners' within-category discrimination.

Limitations

The current study is the first of its kind and inevitably has methodological limitations, which should be taken into account

[4] Yang, J. (2020). *Distributional Learning of Mandarin Lexical Tones in Bidirectional Deep Neural Network*. Unpublished report. Amsterdam: University of Amsterdam.

when generalizing its results and in follow-up research. First, the sample size was small, in particular in Experiments 1 and 2. This did not allow a more balanced distribution of male and female participants. Second, the participants listened to a large number of stimuli that were rather similar to each other in Experiments 1 and 3. Although we took measures to mitigate participant fatigue and boredom by inserting short obligatory pauses between blocks of stimuli, it would have been better to have longer breaks and not to conduct unrelated tests during breaks (Experiment 3). Third, the participants were not given a definition of the Dutch rising pattern. Neither were they told that the rise could have either a low or a mid-high start. The participants thus worked with their own notion of rising patterns. In spite of not being given a definition of the low rise under investigation, the participants in Experiments 1 and 2 showed clear evidence that they rated the postulated prototypical tokens of the low rise in the way that we expected based on the perceptual magnet effect. This may in turn suggests that native speakers of Dutch interpret a typical Dutch rising pattern as a low rise. Nevertheless, it would have been recommendable to make clear to the participants what kind of rise they were supposed to listen for by giving examples of Mi in the 'default testing' situational contexts. Finally, due to the monosyllabic nature of the stimuli and their being spoken as isolated utterances, the question arises as to whether we have tested the perceptual rating and discrimination of instances of the nuclear contour L*H H%. In line with Gussenhoven's (2005) description of rises and the boundary tones and Gussenhoven and Rietveld (2000), who used monosyllabic target words in sentence-final position to study the behavior of H* (as in H*L L%) and L* (as in L*H H%), we believe that the variation created in our stimuli does not change the identity of the boundary tone, which is always H% (see Figure 3), but it does change the valley alignment and shape of the rise before it reaches the target of H% and can hence influence the perceived pitch accent category. We thus argue that our results are pertinent to the categoricality of L*H, not that of the entire contour. Nevertheless, future research using multisyllabic stimuli is needed to validate the results of the current study.

Conclusion

To conclude, our study has put forward the first evidence for the categoricality of the Dutch L*H pitch accent by examining the perceptual magnet effect. This approach shows promise in future research as a means to investigate the categoricality of other pitch accents in Dutch and intonation events in other languages. It is, however, important to take into account that the perceptual magnet effect in perception of intonation may be sensitive to extensive auditory exposure and high context variance.

Data availability statement

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

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements at the time of testing. The patients/participants provided their written informed consent to participate in this study.

Author contributions

JR and AC designed the study and contributed to the interpretation of the results and the writing of this manuscript. JR conducted the study and analyzed the data. All authors contributed to the article and approved the submitted version.

Funding

The research was partially funded by a Talent grant awarded to AC by Utrecht University's strategic theme 'Dynamics of Youth'.

Acknowledgments

We thank Louis ten Bosch for the useful comments on experimental methodology at an early phase of the preparations, Uwe Reichel for comments on the orthogonal polynomial stylization approach adopted, and Huub van den Bergh and Tom Lentz for advice on statistical matters and on the interpretation and presentation of the results.

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.911349/full#supplementary-material

References

Aaltonen, O., Eerola, O., Hellström, Å., Uusipaikka, E., and Lang, A. H. (1997). Perceptual magnet effect in the light of behavioral and psychophysiological data. *J. Acoust. Soc. Am.* 101, 1090–1105. doi: 10.1121/1.418031

Arvaniti, A. (2011). "The representation of intonation," in *Companion to Phonology*. eds. M. van Oostendorp, C. Ewen, B. Hume and K. Rice (Hoboken, NJ: Wiley-Blackwell).

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).

Arvaniti, A., and Ladd, D. R. (2009). Greek Wh-Questions and the Phonology of Intonation. *Phonology* 26, 43–74. doi: 10.1017/S0952675709001717

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme 4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Beckman, M., and Pierrehumbert, J. B. (1986). Intonation structure in Japanese and English. *Phonol. Yearb.* 3, 255–309.

Boersma, P. (2019). "Simulated distributional learning in deep Boltzmann machines leads to the emergence of discrete categories." in *Proceedings of the 19th International Congress of Phonetic Sciences*, pp. 1520–1524.

Boersma, P., and Weenink, D. (2012). *Praat*. Amsterdam: University of Amsterdam.

Caspers, J. (2000). Experiments on the meaning of four types of single-accent intonation patterns in Dutch. *Lang. Speech* 43, 127–161. doi: 10.1177/00238309000430020101

Caspers, J., and van Heuven, V. J. (1993). Effects of time pressure on the phonetic realization of the Dutch accent-lending pitch rise and fall. *Phonetica* 50, 161–171.

Chen, A. (2003). Reaction time as an indicator of discrete Intonational contrasts in English. *EUROSPEECH*, 97–100. doi: 10.21437/Eurospeech.2003-60

Chen, A., Chen, A., Kager, R., and Wong, P. C. M. (2014). "Rises and falls in Dutch and mandarin Chinese." in *Proceedings Fourth International Symposium on Tonal Aspects of Languages*, pp. 83–86

Davis, K. D., and Kuhl, P. K. (1994). Tests of the perceptual magnet effect for American English/k/and/g/. *J. Acoust. Soc. Am.* 95:2976.

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

Fivela, B. G. (2012). "Meanings, shades of meanings and prototypes of Intonational categories," in *Prosody and Meaning*. eds. P. Prieto and G. Elordieta (Berlin: De Gruyter Mouton), 197–237.

Francis, A. L., Ciocca, V., Ma, L., and Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *J. Phon.* 36, 268–294. doi: 10.1016/j.wocn.2007.06.005

Grabe, E., Kochanski, G., and Coleman, J. (2007). "Connecting intonation labels to mathematical descriptions of fundamental frequency," in *Language and Speech* (Thousand Oaks, Los Angeles, CA: SAGE Publications), 50, 281–310.

Gussenhoven, C. (1984). *On the grammar and semantics of sentence accents*. Dordrecht: Foris.

Gussenhoven, C. (1999). Discreteness and gradience in intonational contrasts*. *Lang. Speech* 42, 283–305.

Gussenhoven, C. (2002). Intonation and interpretation: Phonetics and phonology. *Speech Prosody* 2002, 47–57.

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

Gussenhoven, C. (2005). "Transcription of Dutch intonation," in *Prosodic Typology: The Phonology of Intonation and Phrasing*. ed. S.-A. Jun (Oxford: Oxford University Press), 118–145.

Gussenhoven, C., and Rietveld, T. (2000). The behavior of H* and L* under variations in pitch range in Dutch rising contours. *Lang Speech* 43, 183–203. doi: 10.1177/00238309000430020301

Gussenhoven, C., and van de Ven, M. (2020). Categorical perception of lexical tone contrasts and gradient perception of the statement-question intonation contrast in Zhumadian mandarin*. *Lang. Cogn.* 12, 614–648. doi: 10.1017/langcog.2020.14

Hallé, P. A., Chang, Y., and Best, C. T. (2004). Identification and discrimination of Mandarin Chinese Tones by Mandarin Chinese vs. French Listeners. *J. Phon.* 32, 395–421. doi: 10.1016/S0095-4470(03)00016-0

Holovaty, A., and Jacob Kaplan, M. (2009). *The Defini-tive Guide to Django: Web Development Done Right*. New York City: Apress.

Iverson, P., and Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: do they arise from a common mechanism? *Percept. Psychophys.* 62, 874–886. doi: 10.3758/BF03206929

Kluender, K. R., Diehl, R. L., and Wright, B. A. (1988). Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *J. Phon.* 16, 153–169. doi: 10.1016/S0095-4470(19)30480-2

Kuhl, P. K. (1991). Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. *Percept. Psychophys.* 50, 93–107.

Kuhl, P. K., and Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science* 190, 69–72. doi: 10.1126/science.1166301

Kuhl, P. K., and Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VaT stimuli. *J. Acoust. Soc. Am.* 63, 905–917. doi: 10.1121/1.381770

Lacerda, F. (1995). "The perceptual-magnet effect: an emergent consequence of exemplar-based phonetic memory." In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, pp. 140–147.

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

Ladd, D. R., and Morton, R. (1997). The perception of Intonational emphasis: continuous or categorical? *J. Phon.* 25, 313–342.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol.* 54, 358–368.

Lively, S. E., and Pisoni, D. B. (1997). On prototypes and phonetic categories: a critical assessment of the perceptual magnet effect in speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 1665–1679. doi: 10.1037/0096-1523.23.6.1665

Macmillan, N. A., Goldberg, R. E., and Braida, L. D. (1988). Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua. *J. Acoust. Soc. Am.* 84, 1262–1280. doi: 10.1121/1.396626

Ong, J. H., Wong, P. C., and Liu, F. (2020). Musicians show enhanced perception, but not production, of native lexical tones. *J. Acoust. Soc. Am.* 148, 3443–3454. doi: 10.1121/10.0002776

Pierrehumbert, J. B. (1980). "The phonology and Phonet-ics of English intonation," in *PhD thesis, Massachusetts Institute of Technology*.

Pierrehumbert, J. B., and Steele, S. A. (1989). "Categories of tonal alignment in English," in *Phonetica* (Berlin: Mouton De Gruyter), vol. 46, 181–196.

Prieto, P. (2004). "The search for phonological targets in the tonal space: H1 scaling and alignment in five sentence-types in peninsular Spanish," in *Laboratory Approaches to Spanish Phonology*. ed. T. L. Face (Berlin: Mouton De Gruyter), 29–59.

Prieto, P. (2012). "Part I: experimental methods and paradigms for prosodic analysis," in *Handbook of Laboratory Phonology*. eds. A. C. Cohn, C. Fougeron, M. and K. Huffman (Oxford: Oxford University Press), 527–547.

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

Reichel, U. D. (2011). "The CoPaSul Intonation Model," in *Elektronische Sprachverarbeitung*. eds. B. J. Kröger and P. Birkholz (Dresden: TUDpress Verlag der Wissenschaften GmbH), 341–348.

Savino, M., and Grice, M. (2011). "The perception of negative bias in Bari Italian questions," in *Prosodic Categories: Production, Perception and Comprehension*. eds. S. Frota, E. Gorka and P. Prieto (Berlin: Springer), 187–206.

Schneider, K., Dogil, G., and Möbius, B. (2009). German boundary tones show categorical perception and a perceptual magnet effect when presented in different contexts. *INTERSPEECH*, 2519–2522. doi: 10.21437/Interspeech.2009-664

Schneider, K., and Möbius, B. (2005). Perceptual magnet effect in German boundary tones. *INTERSPEECH*, 41–44. doi: 10.21437/Interspeech.2005-34

Schön, D., Magne, C., and Besson, M. (2004). The music of speech: music training facilitates pitch processing in both music and language. *Psychophysiology* 41, 341–349. doi: 10.1111/1469-8986.00172.x

Walsh, M., Schweitzer, K., and Schauffler, N. (2013). Exemplar-based pitch accent categorisation using the generalized context model. *INTERSPEECH*, 258–262. doi: 10.21437/Interspeech.2013-79

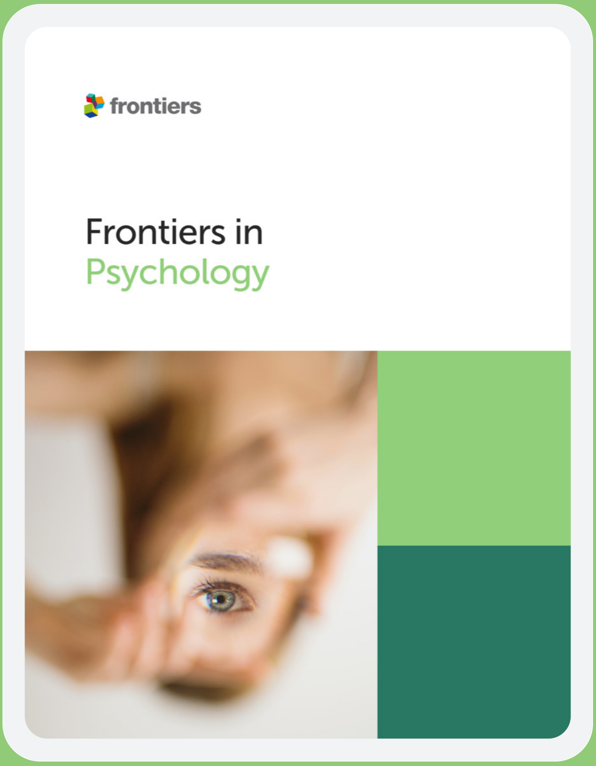
Wickens, T. D. (2002). *Elementary Signal Detection Theory*. Oxford: Oxford University Press.

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., and Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nat. Neurosci.* 10, 420–422. doi: 10.1038/nn1872

Wood, C. C. (1976). Discriminability, response bias, and phoneme categories in discrimination of voice onset time. *J. Acoust. Soc. Am.* 60, 1381–1389. doi: 10.1121/1.381231

Xu, Y. (2013). "ProsodyPro–a tool for large-scale systematic prosody analysis." in *Proceedings of TRASP*. eds. B. Bigi and D. Hirst (France: Aix en Provence).

Xu, Y., Lee, A., Prom-On, S., and Lui, F. (2015). Explaining the PENTA model: a reply to Arvaniti and Ladd. *Phonology* 32, 505–535. doi: 10.1017/S0952675715000299



frontiers

Frontiers in
Psychology

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact