frontiers | Research Topics

Science, technology and art in the spoken expression of meaning

Edited by
Plinio Almeida Barbosa, Sandra Madureira
and Åsa Abelin

Published in
Frontiers in Communication
Frontiers in Psychology

ISSN 1664-8714
ISBN 978-2-8325-3271-3
DOI 10.3389/978-2-8325-3271-3

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

Science, technology and art in the spoken expression of meaning

Topic editors
Plinio Almeida Barbosa — State University of Campinas, Brazil
Sandra Madureira — PUCSP, Brazil
Åsa Abelin — University of Gothenburg, Sweden

Citation
Barbosa, P. A., Madureira, S., Abelin, Å., eds. (2023). *Science, technology and art in the spoken expression of meaning*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-3271-3

Table of contents

TYPE Editorial
PUBLISHED 04 August 2023
DOI 10.3389/fcomm.2023.1257945

OPEN ACCESS

EDITED AND REVIEWED BY
Xiaolin Zhou,
Peking University, China

*CORRESPONDENCE
Plinio Almeida Barbosa
✉ pbarbosa@unicamp.br

RECEIVED 13 July 2023
ACCEPTED 21 July 2023
PUBLISHED 04 August 2023

CITATION
Barbosa PA (2023) Editorial: Science, technology and art in the spoken expression of meaning. *Front. Commun.* 8:1257945. doi: 10.3389/fcomm.2023.1257945

Editorial: Science, technology and art in the spoken expression of meaning

Plinio Almeida Barbosa*

Department of Linguistics, State University of Campinas, Campinas, Brazil

KEYWORDS

speech research, vocal aesthetics, prosodic, prosodic meanings, speech technology

Editorial on the Research Topic
Science, technology and art in the spoken expression of meaning

The set of papers in this Research Topic on Science, Technology and Art in the Spoken Expression of Meaning covers a broad spectrum of pragmatic meanings conveyed by voice quality, speech prosody and body gestures in the contexts of human-human and human-machine interactions.

In human–human interaction, the acoustic characteristics of attitudes are explored together with the prosodic attributes of oral poetry and storytelling in different African languages, and the shapes of pragmatically distinct meanings of questions in French and Italian. Multimodality is investigated for the role of facial features in clear speech and emotions.

Human-machine interaction is explored in such manner as to pave the way for a more anthropomorphic vocal expression on the part of machines including the complex perception of racial components, charisma and degrees of arousal in voice.

These pieces of work present research in languages as diverse as Mandarin, Wuxi, French, Italian, English, Yorùbá, Anyi, Ega, Estonian, and Brazilian Portuguese.

The paper by Ji et al. investigates the encoding of speaker (un)confidence in Wuxi dialect vowels. They show that this propositional attitude is revealed by segmental parameters such as F1 and F2, and by statistical descriptors of prosodic parameters such as F0, intensity and duration. The expression of these parameters interacts with specific tones, namely flat vs. counter tones. Their findings shed new light on the mechanisms of segmental and prosodic encoding of speaker confidence at the vowel level.

The paper by Liu et al. investigates the character voices of leading male characters in the TV series Empresses in the Palace. The authors found that the subordinated characters usually adopt a higher pitch or breathy voice whereas the dominant characters use a lower pitch or modal/creaky voice. In addition, cepstral peak prominence (CPP), F0, and H1-A3 are the key acoustic indicators to distinguish character voices. These results have impact for the entertainment industry, such as the choice of voices for characters in animated films.

Two papers explore the use of the voice for oral poetry and storytelling in African languages. The one by Akinbo et al. is an acoustic study of the vocal expressions of two genres of Yorùbá oral poetry. An original poem in speech mode was acoustically analyzed and the results showed that cepstral peak prominence (CPP), the Hammarberg index, and the energy below 500 Hz in voiced sounds distinguish the two genres of oral poetry and speech but are not as reliable as F0 height and vibrato. The paper by Gibbon studied the rhythm of storytelling in two Niger-Congo tone languages, Anyi and Ega. He showed that the interpretations of the rhythm patterns he found were related to turn interaction types, speech registers and social roles, distinct narration styles, and the genre difference between interactive narration and more formal narratives.

The work by Cresti and Moneglia reveals the correlations between melodic contours and question speech acts in Italian and French. They explore a classification of question illocutionary types present in two parallel corpora of informal speech, one in Italian and the other in French. Yes/no questions were found not only to end in canonical rising contours, but also decreasing ones (26% in French and 36% in Italian), representing 37% of all questions in French and 39% in Italian. A bit <10% of utterances are questions compared to declarative utterances, which are >50% in both languages. Partial questions are as frequent among questions as 26% in French and 38% in Italian. The authors also highlight the importance of questions performed through certain illocutionary types (tag-questions, alternative questions, and double questions) representing 5% in French and 9% in Italian.

The paper by Garg et al. investigates the facial cues associated with clear vs. conversational speech in Mandarin. By comparing movements of the head, eyebrows and lips associated with these two speech styles in Mandarin tone articulation, they examined the extent to which clear-speech modifications involve signal-based exaggerated facial movements or code-based enhancement of linguistically relevant articulatory movements. They found that head, eyebrow, and lip movements correlate with pitch-related variability. They also found, for all the four Mandarin tones, longer duration and greater movements of the head, eyebrows, and lips in clear speech in contrast with conversational speech.

The paper by Madureira and Fontes considers Laver's VPA settings under a sound-symbolic and synesthetic perspective by focusing on the auditory impressions these settings have on listeners' attributions of meaning and associations between vocal and visual features related to the expression of six basic emotions. Their results provide evidence of existing links between sound and meaning which has a relation to the biological codes proposed in the literature, phonetic metaphors, and the vocal and facial gestures involved in emotion expression.

The work by Holliday analyzed the acoustic properties and racialized judgments of four voices of the Siri assistant. A large set of American English listeners responded to questions about the synthetic speaker's sociolinguistic characteristics and personal features. Her evaluation showed that two of the voices were significantly more likely to be rated as belonging to a black speaker than the other two. Additionally, one of the voices judged as belonging to a black speaker was judged less competent, less professional, and funniest. VQ measures such as mean F0 and H1–A3c significantly affect the listeners' ratings of the voices and are correlated with perceptions of pitch and breathiness, respectively.

Another work on synthetic voices is the one by Pajupuu et al. which evaluated the likability of calm and energetic audio advertising styles transferred to Estonian synthesized voices. They used a corpus of advertisements created out of the reading of one text in the two advertisement style to show that not only the calm style was preferred, but also that it differed from the energetic one in acoustic features related to a lower, quieter, and more sonorous voice and a more neutral speaking style.

Finally, the paper by Fucinato et al. showed that charismatic speech features in robot instructions impacts in both team creativity and performance. For doing so, they compared the performance of the teams' activities upon reception of instructions from the robot in a "charismatic" speaking style vs. a neutral way of speaking. The results show that when the robot's speech is based on charismatic characteristics, it is significantly better at enhancing team creativity and performance.

The papers introduced here explore expressive uses of non-verbal language features. These features are multimodal in nature and play a very important communicative role. The main assumption underlying these studies is symbolism, which is manifested in vocal and body language gesture.

Author contributions

PB: Writing—original draft, Writing—review and editing.

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.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 12 December 2022
DOI 10.3389/fpsyg.2022.1028106

OPEN ACCESS

EDITED BY
Plinio Almeida Barbosa,
State University of Campinas, Brazil

REVIEWED BY
Ping Tang,
Nanjing University of Science and Technology, China
Åsa Abelin,
University of Gothenburg,
Sweden

*CORRESPONDENCE
Xiaoming Jiang
xiaoming.jiang@shisu.edu.cn

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

RECEIVED 25 August 2022
ACCEPTED 15 November 2022
PUBLISHED 12 December 2022

CITATION
Ji Y, Hu Y and Jiang X (2022) Segmental and suprasegmental encoding of speaker confidence in Wuxi dialect vowels.
Front. Psychol. 13:1028106.
doi: 10.3389/fpsyg.2022.1028106

Segmental and suprasegmental encoding of speaker confidence in Wuxi dialect vowels

Yujie Ji, Yanbing Hu and Xiaoming Jiang*

Institute of Linguistics, Shanghai International Studies University, Shanghai, China

Introduction: Wuxi dialect is a variation of Wu dialect spoken in eastern China and is characterized by a rich tonal system. Compared with standard Mandarin speakers, those of Wuxi dialect as their mother tongue can be more efficient in varying vocal cues to encode communicative meanings in speech communication. While literature has demonstrated that speakers encode high vs. low confidence in global prosodic cues at the sentence level, it is unknown how speakers' intended confidence is encoded at a more local, phonetic level. This study aimed to explore the effects of speakers' intended confidence on both prosodic and formant features of vowels in two lexical tones (the flat tone and the contour tone) of Wuxi dialect.

Methods: Words of a single vowel were spoken in confident, unconfident, or neutral tone of voice by native Wuxi dialect speakers using a standard elicitation procedure. Linear-mixed effects modeling and parametric bootstrapping testing were performed.

Results: The results showed that (1) the speakers raised both F1 and F2 in the confident level (compared with the neutral-intending expression). Additionally, F1 can distinguish between the confident and unconfident expressions; (2) Compared with the neutral-intending expression, the speakers raised mean f0, had a greater variation of f0 and prolonged pronunciation time in the unconfident level while they raised mean intensity, had a greater variation of intensity and prolonged pronunciation time in the confident level. (3) The speakers modulated mean f0 and mean intensity to a larger extent on the flat tone than the contour tone to differentiate between levels of confidence in the voice, while they modulated f0 and intensity range more only on the contour tone.

Discussion: These findings shed new light on the mechanisms of segmental and suprasegmental encoding of speaker confidence and lack of confidence at the vowel level, highlighting the interplay of lexical tone and vocal expression in speech communication.

KEYWORDS

Wuxi dialect, confidence, lexical tone, vocal production, formant frequency

Introduction

Imagine a situation where a student on a language-learning class asks the lecturer what a specific written word is pronounced because that word is printed in a visually-unrecognized manner. When responding to students, the lecturer may find themselves not sure what the word is. This is when the lecturer replies with his or her own pronunciation of the word to convey their knowledge toward how they evaluate that specific situation.

In daily interactions, speakers often assess whether the event they perceive is true and whether what they say is correct, and they show evidence on their evaluation of things in their statements. Speakers may use the epistemic modality to convey their feeling of (un)knowing about what is proposed (Swerts and Krahmer, 2005). Except for the modal auxiliaries and modal adverbs (Coates, 2012), epistemic modality encompasses a wide range of linguistic forms that feature a specific pattern of prosodic and paralinguistic cues, which are valuable resources for speakers to use to indicate the speaker's confidence or lack of confidence in the truth of the proposition expressed in the discourse. In face-to-face communication, human have an intuition about how confident our conversational partner is about what they are saying.

Vocal confidence expressions serve as "evidentiality" devices for inferring the reliability, correctness, or truth value of what is expressed from a speaker's tone of voice (Caffi and Janney, 1994; Jiang and Pell, 2015). In particular, a speaker's possession of confidence is typically encoded by external cues that provide evidence for the speaker's knowledge about the self-evaluated correctness or truth value of his own statements (London et al., 1970a,b, 1971; Scherer et al., 1973). In contrast, the speaker's lack of confidence or doubt (with only 50% certainty about whether what is said is true) indicates a person's negative attitude or hesitation about a fact or opinion, which is marked by cues that supply signs of untrustworthiness (the lack of moral value of showing remorse or taking responsibility for having done something wrong) or lack of credibility (the perceived believability of information that leads to the listener's feeling of trust; Kuhlen et al., 2015; Belin et al., 2017; Jiang and Pell, 2017).

Previous acoustic-phonetic studies have been conducted from different perspectives regarding whether confidence is defined according to the speaker intention or the listener perception. In the first group of study, speakers were instructed experienced to utter sentences in a confident vs. unconfident way, after which acoustic analysis was performed by measuring different prosodic characteristics of the speaker's voice based on which level of confidence the speakers' intended. The results showed that speakers often spoke with a higher pitch and at a greater intensity when they intended to be confident (Scherer et al., 1973; Van Zant and Berger, 2020). In a second set of works, the same group of vocal stimuli was judged on speaker confidence by an independent group of listeners, and the acoustic analysis was performed based on the regrouping of the stimuli according to the listener's perception. Results showed a distinct pattern of pitch, intensity, and temporal features according to the perceived levels of confidence: the confident expressions were highest in the variation of fundamental frequency (f0), mean amplitude, and amplitude range, but were lower than the unconfident ones in the mean f0, emphasizing the set of acoustic features that listener showed the sensitivities to Jiang and Pell (2014, 2017). In addition, a smaller set of studies directly manipulated the acoustic parameters of the speech and assessed the listener's perceived confidence. These studies showed that the lower pitch can elicit perceptions of higher confidence (Guyer J. 2016).

Differential approaches to determining acoustic-phonetic features based on speaker intended expression or listener perception is how speech materials are selected for acoustic analysis. In the former approach of analysis, the study utilized listeners' perception to validate that the differences in the acoustic features are indeed attributed to the speaker intention. According to the latter approach, the material was regrouped based on perception results, and the regrouped stimuli could only reflect what listeners' commitment but not speakers' own intention. Additionally, while in ideal cases, the speaker and the listener are convergent in the use of communicative cues, in many cases, such convergence is not reached and the encoding and the decoding processes seem to rely on a partially-independent set of cues (Jiang and Pell, 2016, 2017, 2018). In Brunswik's lens model (Brunswik, 1956), acoustic cues in the voice are understood by listeners as probabilistic and partly redundant. The accurate perception of speaker confidence usually depends on both verbal and vocal cues, which can be weighed differently by listeners (Jiang and Pell, 2016). Crucially, listeners are thought to rely on these cues in a partly interchangeable manner (Juslin and Laukka, 2003).

While epistemic and social meanings have been demonstrated to be encoded at the suprasegmental level of speech, they are often found to occur at the segmental level in a much smaller spoken unit. In a study by Laukka et al. (2005) on vocal emotion, it was noted that the first formant (F1) of the stable portion of vowels can predict the level of affective activation. Another study revealed that the first and second formant (F2) of the vowels was influenced by different affect dimensions (Goudbeek et al., 2009). For example, monophthongs of higher-level of arousal resulted in a higher mean F1 than those of lower-level of arousal, whereas monophthongs of positive valence resulted in higher mean values of F2 than those of negative valence. It was also found that adults who stuttered had significantly greater F2 frequency fluctuations when speaking in situations that elicited increases in arousal and unpleasantness. They also showed that those who did not stutter showed little change in F2 fluctuations across varied emotion categories (Bauerly, 2018). Despite that the emotional and epistemic meaning of speech could differ in many aspects, someone argues that the expression of emotions enables speakers to communicate powerful messages to others, which in turn may have a consequence on their attitudes and perceived stances (Guyer J. J. 2016). Delivering an emotional message in persuasive vs. neutral manner altered the voice onset time of consonants (Banzina, 2021; Jiang and Lu, 2021). The complex interplay between the emotions expressed in the voice and the speaker confidence toward the emotional messages suggests that the alterations in the formant frequencies may also be shown in the confidence-related speech. Thus, the present study aimed to examine whether the levels of speaker confidence can be encoded at the segmental level.

More interestingly, the acoustic realization of lexical tones and vocal expressions of social information could involve similar mechanisms. Not only the intonation that conveys social information is realized by acoustic parameters such as the level and variation of f0, but also is the lexical tone in tonal language reflected in the nature of f0 (Eady, 1982; Cutler and Chen, 1997). In tonal languages, the lexical tone is treated as pitch patterns as a contrastive feature. One intriguing aspect of tonal language/dialect is that the intonation system is independent from the lexical tone, although both elements can be expressed by the f0 contour to symbolize the change. This means that some acoustic features such as the f0 contour carries the identifying functions of both linguistic and paralinguistic information. To understand the encoding mechanism of vocal expressions in speech communication, it is essential to investigate the interplay between the lexical tone and intonation of certain emotional or pragmatic function in the context of tonal language/dialect.

Despite that many studies examined the acoustic realizations of different lexical tones (e.g., pitch contour, and duration), there were rare investigations on how the lexical tone could modulate the way social information is encoded in the expressive tone at the segmental level. Chao (1933) proposed two items to distinguish two interplay types of tone and intonation addition patterns: simultaneous addition and successive addition. The simultaneous addition refers to the tones that are the algebraic sums of two factors: the original lexical tone and the sentence intonation proper. The successive addition refers that a rising or falling intonation of a clause is not added simultaneously to certain syllables but added on successively after the lexical tones are completed. The function of the successive addition boundary tone is to express the speaker's emotion rather than to convey linguistic contents. An empirical study investigated how the lexical tone and affective tone interacted in Mandarin Chinese, using monosyllabic emotional utterances as materials (Li et al., 2011). It was found that the tonal space (with all f0 values mapped into a five-point scale), the edge tone (the pattern among the tone and the intonation being added up in emotional speech), and the length of monosyllabic materials differed greatly between seven emotions. In other words, the f0 pattern of lexical tones was affected by the emotional intonation. Furthermore, researchers pointed out that boundary tones of emotional intonation are more appropriately characterized by both traditional boundary tone features and successive addition tone features (Li et al., 2012). For instance, the "disgusting" sound had a "falling" addition tone following the lexical tone of the last syllable, assembled as successive addition tones. These analyses or findings have strongly suggested that the lexical tone and the expressive tone co-constrain the acoustic encoding of social information in speech at the suprasegmental level (e.g., f0 features).

Similar to Mandarin Chinese, Wuxi Dialect, as a member of Wu dialects, has a rich segmental system that consists of 27 consonants, 44 vowels, and eight tones. For instance, Wuxi vowels contain 19 monophthongs, 21 diphthongs, and four triphthongs (Wen, 1996; Cao, 2003). Considering that the Wuxi dialect also has a rich system of tones, it is likely that the vocal expression of confidence in this dialect could also show a pattern of successive or simultaneous addition to the lexical tones. In particular, the tonal context (a flat tone or a contour tone) could modulate the acoustic encoding of vocally-expressed confidence in the Wuxi dialect.

Some studies reported acoustic encoding of vocal expressions from a limited number of speech materials spoken by a larger number of speakers (Pell et al., 2009; McAleer et al., 2014; Ponsot et al., 2018), which had the advantage of considering inter-speaker variability to reveal a generalizable pattern across speakers. However, it could also suffer from poor generalizability across items concerning limited materials. Other reports focused on a larger number of materials spoken by a smaller number of speakers (typically 4–8 speakers, e.g., Pell et al., 2009; Liu and Pell, 2012; Hellbernd and Sammler, 2016; Jiang et al., 2017, 2020; Caballero et al., 2018). The method using numerous materials from limited speakers has the advantage of better generalizability across spoken materials and the disadvantage of lack of inter-speaker variability.

In the present study, 20 different vowels were included for the analysis, which aimed to increase the generalizability across the vowel acoustic space. Additionally, four speakers (two males and two females, from the middle-aged to the elderly) were chosen for this study. To compensate for the relatively lower generalizability across speakers, the speakers were selected to increase the speaker variations in social identities (such as biological sexes and ages) as much as possible.

Considering that related previous studies were mainly focused on the suprasegmental level of vocal expression in sentences in non-tonal languages such as English, the present study aimed to investigate how tonal-language (i.e., Wuxi dialect) speakers encode social intentions in their voices at both segmental and suprasegmental levels. To achieve our purpose, therefore, we generated a corpus with four native Wuxi dialect speakers (i.e., two females and two males) expressing different levels of confidence (three levels: confident, unconfident, and neutral-intending) in different lexical tones (two levels: flat vs. contour tone) through vowels. We measured which acoustic cues individuals use at the segmental level (and also at the suprasegmental level) to encode confidence levels in their voices and tested the way in which these acoustic-phonetic features of confidence were influenced by lexical tones.

The acoustic features the present study focused are as followed. The segmental features included the first two formants (F1 and F2; Laukka et al., 2005; Goudbeek et al., 2009; Ji and Jiang, 2021; Salais et al., 2022); the suprasegmental features included (1) The fundamental frequency (f0); (2) The sound intensity (dB); and (3) Duration (Scherer et al., 1973; Jiang and Pell, 2014, 2017; Guyer

J. 2016; Van Zant and Berger, 2020). The present study focused on these acoustic parameters because the social information intended in the speaker voice has been associated with these segmental and suprasegmental features in related studies.

Previous studies showed the association between the increased arousal and higher mean F1 and the association between the positive potency and higher mean F2 in the speech. It is expected that confident and unconfident voices could lead to higher F1 compared to neutral voices, and confident voices could be associated with higher F2 than neutral and unconfident voices. Given that prosodic features such as pitch and intensity have been shown to reliably differentiate different speaker confidence, it is expected that the confident voice would show lower fundamental frequency and greater intensity as compared with the unconfident voice, as such finding can be extended from English at the sentence level to a type of eastern Chinese-dialect at the level of a smaller segmental unit. Considering that lexical tone plays a role in the expression of emotions in previous studies, an interaction between speakers' intended confidence and lexical tone is expected to occur in the size of f0, which means that the lexical tone affects the acoustic representation of vocal expression of speaker confidence.

Materials and methods

Participants

Four native Wuxi-dialect speakers were invited to produce sentences in different levels of confidence in their native dialect. Only middle-aged and elderly adults but not young adults were selected as speakers because studies have shown that the Chinese dialect pronunciation remained more stable in the middle-aged and elderly populations (Liu and Chen, 2018; Zhang, 2020). Moreover, speakers of certain age ranges were selected to increase the generalizability of the findings by increasing the speaker variations in social identities (such as biological sexes and ages). All speakers (Mean Age = 64.25 years, SD = 17.10 years, and two females) were all born and raised up in Wuxi, a city located in the Jiangsu Province in China, where local residents speak Wuxi Dialect as their native tongue (Cao et al., 2003). All speak Mandarin but did not pick it up until 5 years old. All reported to speak only Wuxi dialect at home and use dialect to communicate more often than Mandarin during work. None of the speakers had lived outside of Wuxi consecutively for over 2 years. The mean self-reported proficiency of the four speakers was 6.25 (SD = 0.5) for speaking and 6.75 (SD = 0.5) for listening Wuxi Dialect, and was 5.25 (SD = 1.5) for speaking and 7 (SD = 0) for listening Mandarin Chinese (out of seven-point scale, with 1 the least proficient and 7 the most proficient). All reported to receive formal education for 12 years. All speakers reported to have normal hearing and none had suffered any previous neurological or speech disorders. Speakers were not selected for having previous training or experiences in professional acting or public speaking. This study was approved by the Ethics Committee of the Institute of Linguistics from the Shanghai International Studies University.

Materials

To eliminate the potential effect of local consonants on the subsequent vowels, word materials for production were selected with zero-consonant. Vowels were selected exhaustively based on the phonological system of the Wuxi dialect (Cao, 2003) to enrich the types of vowel materials and increase the degree of vowel variation. The selected totally 20 vowels consisted in 10 monophthongs (i, u, y, ɚ, a, ʌ, ʊ, ɛ, ã, ɒ̃) and 10 diphthongs (ia, ua, iʌ, yʊ, uɛ, ei, əɯ, iã, uã, uɒ̃). Despite covering such a variety of vowels, the present study was interested in the overall patterns of different vowels at different levels of confidence instead of the differences in the vowels themselves. Two lexical tones (i.e., the flat tone and the contour tone) were selected as target tones (see Figure 1). These two tones were chosen for two reasons: First, all vowels can be produced in both flat tone and contour tone contexts, to ensure vowels in target tones correspond to real words in Wuxi dialect to the maximal extent[1]; Second, these two tones are representative in terms of the fundamental frequency patterns, with the flat tone having a stable fundamental frequency throughout the vowel, and the contour tone having a constantly changing fundamental frequency throughout the vowel. A previous study demonstrated that formant peaks contributed to the high level tone and the third tone in Mandarin Chinese (Zhang et al., 2021). In Wuxi dialect, the flat tone is similar to the high-level tone in Mandarin while the contour tone is similar to the third tone in Mandarin which starts at the low tone with a slight fall and then rises to a high pitch.

Altogether, 40 different words were selected to form the production list for the elicitation (Supplementary Table S1, see Supplementary materials). Carrier sentences were created such that each word was embedded in a sentence "This word is 'X'," and to ensure the zero-consonant vowels were preceded by a local linguistic context which was semantically neutral. Therefore, in total there were 480 stimuli (4 speakers × 3 confidence levels × 20 vowels × 2 lexical tones).

Recording and elicitation procedure

Speakers were seated in a quiet room in front of the TroyStudio portable sound absorption equipment which aimed at reducing

[1] All but three vowels ([ɚ, uã, ia]) corresponds to a real word in flat tone (see Supplementary Table 1 in Supplementary material). During production, speakers were shown words for both flat and contour conditions. For the vowels with no word mapping in the flat tone, the speaker was shown words in the counterpart of the contour tone but was instructed to produce that in the flat tone of voice.

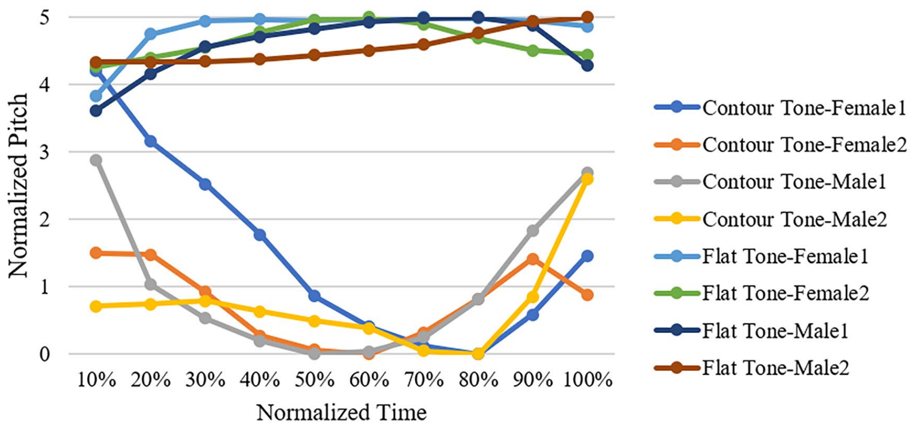

FIGURE 1
Four speakers neutral-intending expression of vowel /a/ with contour tone and flat tone normalized in a five-point scale.

sound reverberation and environmental noise. The vocal stimuli were recorded by a TASCAM-DR-07X recorder (with a sampling rate of 44.1 kHz, 16 bit, mono, input level of −9 dBV). The distance between mouth and the microphone was approximately 15 cm and was ensured for each speaker. To facilitate the production of the vocally-expressed confidence, speakers were instructed to produce each sentence twice with a certain level of confidence by responding to the same question from a native Wuxi dialect female confederator in a mini-dialog format (e.g., Question: What is the word? Answer: This word is "X"; Jiang and Pell, 2014, 2017, 2018). The target vowel was the new information in the answer which corresponded to the wh-constituent in the questions, which aimed at inducing natural vowels for subsequent acoustic analysis (Waters et al., 2021). The question was asked in a neutral tone of voice. The answer was produced in a certain level of confidence. The speakers were instructed to articulate the word clearly and to communicate the target level of confidence directly to the confederator and to avoid simply reading out the sentences.

The vocal stimuli were recorded in separate blocks, in each of which a certain intended level of confidence was elicited. Such procedure has proven successful to elicit a stable level of speaker expression across sentences. In the confident condition, the speakers were instructed to produce the sentence with 100% certainty that the word they said in the sentence was true. In the unconfident condition, the speakers produced the sentence with the knowledge that only in 50% cases the word they said was true. The unconfident expression was not elicited through questions given that the encoding of linguistic question was not the same as the vocal expression of lack of speaker confidence. For instance, the speaker could simply lengthen the production of certain constituents to mark their lack of confidence (Jiang and Pell, 2017). To elicit a condition which lacked in any level of explicitly-encoded speaker confidence, speakers were also instructed to produce a corresponding set of neutrally-intending sentences. In this condition, the speakers were encouraged to produce utterances "without feeling any particular emotion or attitude" toward the content of the sentence. At no time did the confederator provide an explicit model of how intended target meanings should be expressed. For confident and unconfident blocks, the speaker was additionally instructed to convey the intended level of confidence throughout the sentence. The order of the three recording blocks (confident, unconfident, and neutral) were randomized across speakers with the exception that the block for the neutrally-intending expressions always preceded the blocks of confident and unconfident expressions. Breaks were inserted between blocks to ensure a successful transition between modes of different levels of confidence. The repetition of each sentence was initially evaluated by a native Wuxi-dialect speaker to select the best exemplar per item/speaker, based on her intuition to decide which item better conveyed the intended target level of confidence, and to discard the items that sounded unnatural and/or had speech errors.

To ensure that the three levels of speaker's intended confidence were perceived as different, 16 participants who did not participate in the production task listened to each vowel and rated the speaker's level of confidence on a seven-point scale (1 = not at all confident; 7 = very much) for all stimuli. The mean rating was 3.93 (SD = 1.62) for the unconfident expression, 4.20 (SD = 1.46) for the neutral expression and 4.57 (SD = 1.51) for confident ones. One-way ANOVA showed that the three levels of speaker's intended confidence was perceptually different [F (2,7,526) = 118.42, $p < 0.001$; Bonferroni post-test, $ts > 6.43$, $ps < 0.001$].

Data analysis

Based on the preliminary screening, a total of 477 recordings including both monophthongs and diphthongs were subjected to further analysis, with one diphthong of the flat tone produced in the confident expression of one female speaker and one monophthong of two lexical tones produced in the unconfident expression of the other female speaker were discarded due to pronunciation errors.

Both the segmental features that distinguish vowel units and the suprasegmental features that are superimposed on these units were analyzed on target vowels in order to show different levels of acoustic features of vowels expressed in different intended levels of confidence.

Segmental features

To quantify F1 and F2, we labeled the stable articulation of the vowels based on the selected stimuli in TextGrid with Praat (Version 6.1.52) before extracting the mean values of F1 and F2. For the monophthongs, the stable articulations were labeled; whereas for the diphthongs, the stable articulations of the first and the second vowels were separately labeled. The Praat script[2] was adapted to extract mean formants (Hz) of the stable section of the particular vowels labeled in the Textgrid Tier for both monophthongs and diphthongs.

Prosodic features

The prosodic features included: the mean fundamental frequency (mean f0, in Hz), the range of fundamental frequency (f0 variance, in Hz), and the mean sound intensity (mean intensity, in dB), the range of sound intensity (intensity variance, in dB) for both monophthongs and diphthongs, duration (in ms) for monophthongs only[3]. The same stable parts for the vowels as in the analysis of segmental features were used to obtain prosodic features except duration. The entire vowel articulation was labeled to define the duration for monophthongs. Formant and prosodic values extracted from the first and the second vowels of the same diphthong were treated as two separate parts. The *ProsodyPro* tool (Xu and Prom-On, 2014) was used to extract duration, intensity (mean intensity, maximum intensity, and minimum intensity) and fundamental frequency (mean f0, maximum f0, and minimum f0) of the vowel stimuli. The intensity range and the f0 range were then calculated by subtracting the minimum value from the maximum value.

A normalization procedure was applied to all prosodic features of each stimuli before comparing between speakers (Pell et al., 2009; Liu and Pell, 2012; Jiang and Pell, 2017). The mean fundamental frequency of each speaker's articulation naturally differs, and the absolute differences in f0 range vary as an index of the speaker's meanf0. There is evidence that when speaking in a non-emotional manner, each speaker has to a highly stable "resting frequency" or end-point f0 at the end of their utterances which is characteristic for that individual (Menn and Boyce, 1982; Pell et al., 2009). In order to correct for the individual difference in a speaker's mean voice pitch, all f0 measures (mean, maximum, and minimum f0) were normalized in relation to the individual "resting frequency" of each speaker (i.e., the average minimum f0 value of all neutral stimuli produced by that speaker). Measures of normalized f0 range were then calculated by subtracting the normalized minimum f0 values from the normalized maximum f0 values. The same method was applied to the normalization of the intensity values of each speaker. The normalized duration for monophthong or diphthong was obtained in relation to the individual "resting production length" of each speaker (i.e., the average mean duration of all neutral stimuli for monophthongs or diphthongs produced by that speaker).

Statistical analysis

Statistical modelings were conducted for segmental and prosodic features separately. Considering the correlations among our dependent variables (Jiang and Pell, 2014, 2017), multiple ANOVA (MANOVA) were used to reduce the joint error rate and to achieve greater statistical power compared to a series of ANOVA tests (Matuschek et al., 2017; see also https://statisticsbyjim.com/anova/multivariate-anova-manova-benefits-use/). To ascertain whether speaker confidence differed in the linear composition of acoustic features, MANOVAs were conducted on the linear composition of formant features and of suprasegmental features (f0 and intensity values) separately.

To determine the effects of Lexical Tone, Speaker Confidence and their interaction(s) on each independent acoustic feature, linear mixed effects models (LMMs) were separately conducted on each segmental and suprasegmental feature. The model selection procedure started with a baseline model including only by-subject and by-vowel item random intercepts. Predictors were then added in a step-wise fashion to determine the model fit. Model comparisons were conducted using chi-squared tests of model log-likelihoods. The predictor was dropped from the model when it did not yield significant improvement in the model comparison (Ip and Cutler, 2020). The AICs (Akaike Information Criterion) of added models were compared. Compared with the baseline model, the best fitting model contained significant effect of Lexical Tone, Speakers Confidence and their interaction for model of F1 [$\chi^2(2) = 7.42$, $p = 0.025$], F2 [$\chi^2(2) = 11.11$, $p = 0.049$], mean f0 [$\chi^2(2) = 19.00$, $p < 0.001$], range of f0 [$\chi^2(2) = 17.04$, $p < 0.001$], mean intensity [$\chi^2(2) = 11.96$, $p < 0.001$], and range of intensity [$\chi^2(2) = 8.20$, $p = 0.012$]. The fixed factors were Lexical Tone and Speakers Confidence. The random factors were Subjects and Vowel Items.

y[4] ~ lexical tone*levels of confidence + (1|Subject) + (1|Item)

All data were analyzed using linear mixed effects models (LMMs) within the *lmerTest* packages of R (Version 3.1.3, https://github.com/runehaubo/lmerTestR). Considering the sample size

[2] https://github.com/feelins/Praat_Scripts/tree/master/10-get_duration_and_formant

[3] We did not include duration values of diphthongs in statistic models because labeling the transition boundary and the boundaries for the stable portion of the vowel articulation could be arbitrary.

[4] y refers to the dependent factor (the acoustic features, e.g., F1, F2, f0, intensity and duration) in each model.

per speaker confidence per lexical tone was 120[5] for all models except for the model of duration ($n = 80$[6]), the p-values for fixed effects were tested by parametric bootstrapping[7] using function *mixed()* from R package "afex" (nsim = 10,000; Singmann, 2019).

Considering the complexity of acoustic parameters in the LMMs, the current study put the results of statistics results into tables to ensure the conciseness and intuitiveness of the results.

Results

Segmental features

Table 1 demonstrated the mean F1 and F2 values computed for all vowels across lexical tones and levels of speaker confidence. The MANOVA on the linear combination of the two formant parameters showed a significant effect of Speaker Confidence [Pillai's Trace =0.03, F (2,702) =5.30, $p < 0.001$, $\eta^2_p = 0.01$]. The models for the effect of Lexical Tone did not reach significance [Pillai's Trace = 0.001, F (1,708) =0.31, $p = 0.735$, $\eta^2_p = 0.0008$].

To ascertain the potential effect of Speaker Confidence and its interaction with Lexical Tone, the LMMs were separately built on mean values of F1 and F2 (see Table 2). The F1 model revealed a significant main effect of Speaker Confidence, suggesting that the confident expression revealed a larger F1 than the unconfident and the neutral-intending expression, and the unconfident did not differ from neutral-intending expression (see Figure 2A).

The F2 model revealed a significant main effect of Speaker Confidence, suggesting that the confident expression revealed a larger F2 can only be seen between confident vs. neutral-intending expression (see Figure 2B).

In summary, the speakers raised both F1 and F2 in the confident level (compared with the neutral-intending expression). Additionally, F1 can distinguish between the confident and unconfident expressions.

TABLE 1 Mean and SD of mean F1 and F2 values (in Hz) in different lexical tones averaged between speakers.

	F1		F2	
	Flat tone	Contour tone	Flat tone	Contour tone
Confident[a]	687.22	629.78	1667.49	1720.40
	(291.85)	(305.07)	(562.71)	(589.64)
Unconfident	543.66	572.55	1613.19	1635.03
	(273.19)	(295.99)	(581.39)	(611.89)
Neutral	567.27	574.89	1565.96	1593.67
	(273.36)	(283.19)	(614.03)	(610.40)

Standard deviations were shown in brackets. [a]Sample size per cell was 120, except that for vowels of a flat tone, the sample size was 118 for the confident expression and was 119 for the unconfident expression; for those of a contour tone, the sample size was 119 for the unconfident expression, given that non-standard incorrect pronunciations were discarded.

Prosodic features

We examined whether speakers utilized prosodic cues to express levels of confidence under two different lexical tones in the same two steps: MANOVAs and LMERs. In Table 3, the means and SDs for the prosodic values of vowels by all factor levels (lexical tones and levels of speaker confidence) are presented.

The MANOVA was first built for the effect of Speaker Confidence on the linear combination of four prosodic parameters, including mean f0, f0 range, mean intensity, and intensity range. The model showed a significant effect of Speaker Confidence [Pillai's Trace =0.26, F (2,707) = 25.96, $p < 0.001$, $\eta^2_p = 0.12$]. The MANOVA also showed a significant effect of Lexical Tone [Pillai's Trace = 0.61, F (1,708) =277.13, $p < 0.001$, $\eta^2_p = 0.61$]. Both Speaker Confidence and Lexical Tone significantly modulated the linear combination of the prosodic parameters.

To show the potential effect of Speaker Confidence and its interaction with Lexical Tone, the LMMs were separately built on each prosodic factor (see Table 4). The mean f0 model revealed a significant main effect of Speaker Confidence (see Figure 3A), suggesting that the mean f0 was largest in the unconfident expression, seconded by the confident, and was smallest in the neutral-intending expression. The model revealed a significant main effect of Lexical Tone, suggesting that the mean f0 was significantly larger in vowels of a flat tone than those of a contour tone. The Speaker Confidence x Lexical Tone interaction was significant (see Figure 4A). For vowels of a flat tone, the mean f0 differed among three levels of confidence, with the mean f0 largest in the unconfident expression, followed by the confident, and smallest by the neutral-intending expression; for those of a contour tone, the mean f0 was larger in the unconfident than in both the confident and the neutral expression and the confident did not differ from neutral-intending expression.

The f0 range model revealed a significant effect of Speaker Confidence (see Figure 3B), suggesting that the f0 range was significantly smaller in the neutral-intending expression than the confident and the unconfident expression and the confident did

[5] The sample size per cell for all features except for duration was 120 (=4 speakers * (10 monophthongs +10 diphthongs *2 parts of the vowels)). We divided the diphthong into two portions of the vowel to calculate the acoustic features from each separate vowel.

[6] The sample size per cell for the duration was 80 (=4 speakers * (10 monophthongs +10 diphthongs)).

[7] The parametric bootstrapping approach showed an advantage in dealing with statistic issues with a small-sample design (Fisher and Hall, 1991); Also see the link: https://www.millerwjr.com/all-projects/2018/3/10/non-parametric-bootstrap-in-r-wiping%20maintains%20an%20advantage%20over%20non-parametric%20bootstrappinghe%20smoothing%20effects%20offered%20by%20estimating%20the%20distribution.

TABLE 2 LME model performances for formant features.

Formant features	Effect	Chisq	P-value	Contrast	Estimate	SE[b]	*t*	P-value[a]	95%CI
F1	Lexical Tone	1.57	0.207	Contour—Flat					
	Speaker Confidence	27.89	***	Conf—Neut	87.5	20.4	4.29	***	[38.6,137.0]
				Conf—Unconf	99.5	20.5	4.87	***	[50.4,149.0]
				Neut—Unconf	12.0	20.4	0.59	1.00	[−37.0,61.0]
	Lexical Tone × Speaker Confidence	4.91	0.092						
F2	Lexical Tone	0.02	0.890						
	Speaker Confidence	7.63	0.026**	Conf—Neut	116.7	42.5	2.75	0.019	[14.7,218.6]
				Conf—Unconf	68.5	42.6	1.61	0.324	[−33.7,170.7]
				Neut—Unconf	−48.1	42.5	−1.13	0.773	[−150.1,53.8]
	Lexical Tone × Speaker Confidence	0.10	0.955						

[a]Significance levels under Bonferroni-corrections: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. [b]SE: standard error.

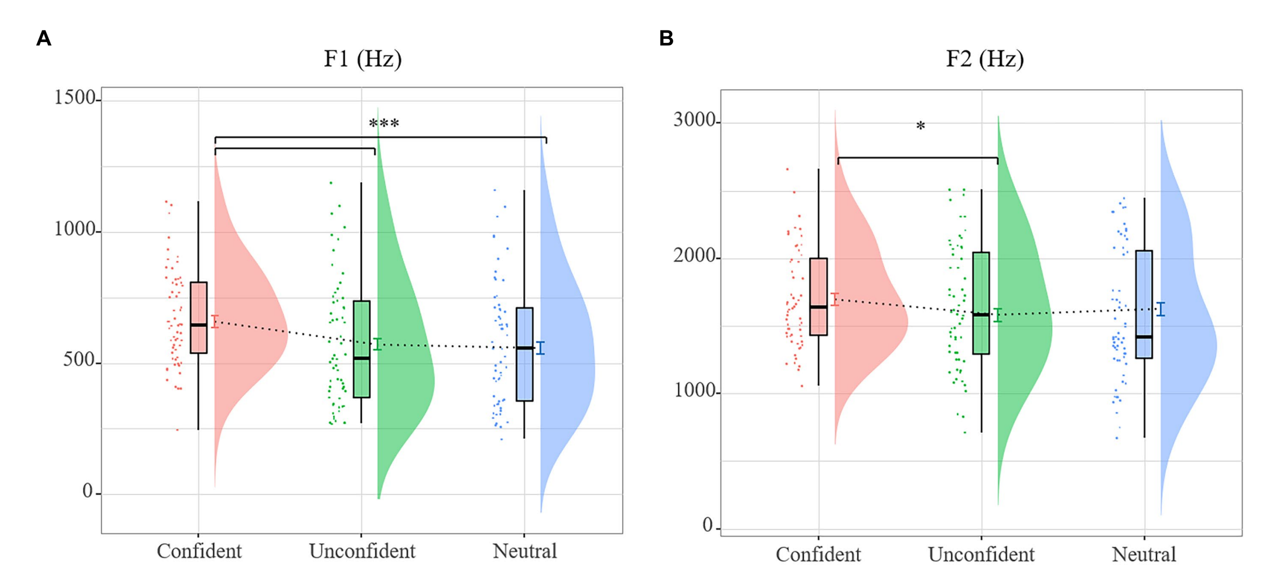

FIGURE 2
Raincloud plots for formant features showing the main effect of speaker confidence. **(A)** F1 and **(B)** F2 values per confidence level for all vowels.

not differ from unconfident expression. The model also revealed a significant effect of Lexical Tone, suggesting that the f0 range was larger in vowels of a contour tone than those of a flat tone. The f0 range model revealed a significant Speaker Confidence x Lexical Tone interaction (see Figure 4B). For vowels of a contour tone, the f0 range was smaller in the neutral-intending than in both the confident and the unconfident expression and the confident did not differ from unconfident expression; for those of a flat tone, the f0 range did not differ among three levels of speaker confidence.

The mean intensity model revealed a significant effect of Speaker Confidence (see Figure 3C), suggesting that the mean intensity was significantly larger in the confident than the unconfident and the neutral-intending expression. No significant difference was shown between the neutral-intending and the unconfident voice. The mean intensity model revealed a significant effect of Lexical Tone, with the mean intensity of vowels of a flat tone sounding more intense than those of a contour tone. Moreover, the mean intensity model showed a significant Speaker Confidence x Lexical Tone interaction (see Figure 4C). For vowels of a contour tone, the mean intensity was larger in the confident than the unconfident and neutral-intending expression. No significant difference was shown between the neutral-intending and the unconfident voice. But for those of a flat tone, the mean intensity differed among all three levels of speaker confidence, with the mean intensity largest in the confident expression, followed by the neutral-intending expression, and lowest by the unconfident.

The intensity range model revealed a significant effect of Speaker Confidence (see Figure 3D), suggesting that the intensity range was significantly larger in the confident than the neutral-intending expression. The main effect of Lexical Tone was not

TABLE 3 Means and standard deviations of the normalized pitch, intensity, and duration measures in different lexical tones averaged across speakers.

	Mean F0		F0 range		Mean intensity		Intensity range		Duration[b]	
	Flat tone	Contour tone	Flat tone	Contour tone	Flat tone	Contour tone	Flat tone	Contour tone	Flat tone	Contour tone
Confident[a]	0.46 (0.20)	0.003 (0.22)	0.13 (0.13)	0.29 (0.23)	0.10 (0.07)	0.05 (0.06)	0.08 (0.05)	0.09 (0.07)	0.88 (0.18)	1.17 (0.18)
Unconfident	0.59 (0.23)	0.02 (0.22)	0.13 (0.15)	0.30 (0.21)	0.05 (0.08)	0.02 (0.06)	0.08 (0.06)	0.07 (0.05)	0.94 (0.19)	0.94 (0.19)
Neutral	0.29 (0.16)	−014 (0.13)	0.12 (0.12)	0.18 (0.13)	0.08 (0.06)	0.02 (0.05)	0.08 (0.05)	0.07 (0.05)	0.84 (0.16)	0.84 (0.16)

Standard deviations were shown in brackets. [a]Sample size of f0 and intensity features was 120 (40 items for monophthongs and 80 items for diphthongs), except that: for vowels of a flat tone, the sample size was 118 for the confident expression, 119 for the unconfident and neutral-intending expression; for those of a contour tone, the sample size was 117 for the confident expression, 118 for the unconfident expression, and 119 for the neutral-intending expression given non-standard incorrect pronunciations. [b]Sample size of duration was 80, except that for monophthongs of a contour tone, the sample size was 78 for the unconfident expression; for those of a flat tone, the sample size was 79 for the unconfident expression.

significant. The mean intensity model also showed a significant Speaker Confidence × Lexical Tone interaction (see Figure 4D). For vowels of a contour tone, the intensity range was larger in the confident than the unconfident and the neutral-intending expression, with no difference between the latter two. For those of a lexical tone, the intensity range did not differ among all three levels of speaker confidence.

The duration model was performed on all vowels, with Speaker Confidence and Lexical Tone as two fixed factors, Vowel Item and Speaker as random intercepts. Vowel type (monophthong vs. diphthong) was included as the fixed covariate given that the durations of monophthongs and diphthongs were different. The model revealed a significant effect of Speaker Confidence (see Figure 3E), suggesting that the duration was significantly shorter in the neutrally-intending expression than the confident and the unconfident expression and no significant difference were shown between the latter two conditions. The model revealed a significant main effect of Lexical Tone, suggesting that the normalized duration was significantly larger in vowels of a contour tone than that of a flat tone. The interaction between Speaker Confident and Lexical Tone was not significant.

To conclude, compared with the neutral-intending expression, the speakers raised mean f0, had a greater variation of f0 and prolonged pronunciation time in the unconfident level, while they raised mean intensity, had a greater variation of intensity and prolonged pronunciation time in the confident level. Additionally, considering the interplay of lexical tone and intended confidence, the speaker modulated the mean f0 and mean intensity to a larger extent on the flat tone than the contour tone to differentiate between levels of confidence in the voice, while they modulated the range of f0 and intensity more on the contour tone than the flat tone.

Discussion

In this study, acoustic-phonetic features at both segmental and suprasegmental level were examined on vowels produced by native Wuxi dialect speakers in confident, unconfident and neutral tone of voice. We found that the intended speaker confidence can be encoded in the mean values of both the first and the second formant at the segmental level. In particular, the vowel spoken in a confident tone demonstrated a larger F1 than the one spoken in neutral and unconfident tones and a larger F2 than the one spoken in a neutral tone. For all vowels, both temporal and spectral prosodic features varied as a function of the intended speaker confidence. Both f0 and intensity measures were associated with the intended speaker confidence. In particular, the more confident the speakers' intended, the mean f0 was lower and the mean intensity was stronger. As long as the speaker encoded a certain level of confidence, whether confident or not, compared to a neutral tone, the f0 variation was larger and the intensity variation was lower. The speaker modulated the mean f0 and mean intensity to a larger extent on the flat tone than the contour tone to differentiate levels of confidence in voice but, while they modulated the range of f0 and intensity more on the contour tone than the flat tone.

This finding suggests that segmental and suprasegmental features in vowels can provide sufficient information to differentiate when the speakers' intended high vs. low confidence and when the speaker did or did not intend any emotion or confidence in the sound (Jiang and Pell, 2015). In addition, lexical tone modulated the acoustic encoding of speaker confidence levels in vowels. The speaker modulated mean f0 and mean intensity to a larger extent on the flat tone than the contour tone to differentiate between levels of confidence in the voice but modulated f0 range and intensity range more on the contour tone than the flat tone, suggesting a complex mechanism regarding how tone and vocal expression interplay with each other.

Encoding speaker confidence in formant features

While previous studies have mostly assigned critical roles of formant peaks in determining vowel identity (Barreda and Nearey, 2011), the current study extended this finding by demonstrating that the formant values can be associated with vocally-expressed confidence in speech production. In particular, speaking in a confident voice raised both F1 and F2.

Existing speech-articulatory models (Fant, 1960; Ladefoged et al., 1978) and empirical studies focusing on the relationship between formant frequencies and tongue positions (Lee et al.,

TABLE 4 LME model performances for normalized prosodic features.

Prosodic features	Effect	Chisq	*p* value[a]	Contrast		Estimate	SE[b]	*t*	*p* value	95%CI
Mean F0	Lexical Tone	700.06	***	Contour—Flat		−0.49	0.01	−34.91	***	[−0.51, −0.49]
	Speaker Confidence	166.69	***	Conf—Neut		0.16	0.02	9.28	***	[0.12,0.20]
				Conf—Unconf		−0.07	0.02	−4.04	***	[−0.11,-0.03]
				Neut—Unconf		−0.23	0.02	−13.36	***	[−0.27,-0.19]
	Lexical Tone × Speaker Confidence	19.02	***	Contour tone	Conf—Neut	0.14	0.02	5.89	***	[0.08,0.20]
					Conf—Unconf	−0.01	0.02	−0.58	1.00	[−0.07,0.04]
					Neut—Unconf	−0.16	0.02	−6.49	***	[−0.21,-0.10]
				Flat tone	Conf—Neut	0.17	0.02	7.24	***	[0.12,0.23]
					Conf—Unconf	−0.12	0.02	−5.14	***	[−0.18,-0.07]
					Neut—Unconf	−0.30	0.02	−12.41	***	[−0.36,-0.24]
F0 range	Lexical Tone	107.75	***	Contour—Flat		0.13	0.01	10.86	***	[0.11,0.15]
	Speaker Confidence	27.21	***	Conf—Neut		0.06	0.01	4.30	***	[0.03,0.10]
				Conf—Unconf		−0.01	0.01	−0.52	1.00	[−0.04,0.03]
				Neut—Unconf		−0.07	0.01	−4.82	***	[−0.10,-0.04]
	Lexical Tone × Speaker Confidence	16.61	***	Contour tone	Conf—Neut	1.10e-01	0.02	5.37	***	[0.06,0.16]
					Conf—Unconf	−1.50e-02	0.02	−0.73	1.00	[−0.06,0.03]
					Neut—Unconf	−1.25e-01	0.02	−6.09	***	[−0.17,-0.08]
				Flat tone	Conf—Neut	1.48e-02	0.02	0.72	1.00	[−0.03,0.06]
					Conf—Unconf	−5.38e-06	0.02	0.00	1.00	[−0.05,0.05]
					Neut—Unconf	−1.48e-02	0.02	−0.73	1.00	[−0.06,0.34]
Mean intensity	Lexical Tone	104.67	***	Contour – Flat		−0.05	0.00	−10.70	***	[−0.06, −0.04]
	Speaker Confidence	54.24	***	Conf—Neut		0.03	0.01	9.28	***	[0.15,0.04]
				Conf—Unconf		0.04	0.01	−4.04	***	[0.03,0.05]
				Neut—Unconf		0.01	0.01	−13.36	0.127	[−0.00,-0.02]
	Lexical Tone × Speaker Confidence	12.10	0.003	Contour tone	Conf—Neut	0.03	0.01	3.68	***	[0.01,0.05]
					Conf—Unconf	0.02	0.01	2.98	0.009	[0.00,0.04]
					Neut—Unconf	−0.01	0.01	−0.684	1.00	[−0.02,0.01]
				Flat tone	Conf—Neut	0.03	0.01	3.60	0.010	[0.01,0.05]
					Conf—Unconf	0.06	0.01	7.14	***	[0.04,0.07]
					Neut—Unconf	0.03	0.01	3.56	0.001	[0.01,0.05]
Intensity range	Lexical Tone	1.79	0.181	Contour—Flat						
	Speaker Confidence	10.79	0.005	Conf—Neut		0.01	0.00	3.25	0.004	[0.00,0.03]
				Conf—Unconf		0.01	0.00	2.16	0.094	[−0.00,0.02]
				Neut—Unconf		−0.00	0.00	−1.09	0.830	[−0.02,0.01]
	Lexical Tone × Speaker Confidence	7.95	0.020	Contour tone	Conf—Neut	0.03	0.01	3.95	***	[0.01,0.04]
					Conf—Unconf	0.02	0.01	3.38	0.002	[0.01,0.04]
					Neut—Unconf	−0.00	0.01	−0.56	1.00	[−0.02,0.01]
				Flat tone	Conf—Neut	0.00	0.01	0.66	1.00	[−0.01,0.02]
					Conf—Unconf	−0.00	0.01	−0.32	1.00	[−0.01,0.01]
					Neut—Unconf	−0.01	0.01	−0.98	0.988	[−0.02,0.01]
Duration	Lexical Tone	29.29	***	Contour—Flat		0.36	0.07	5.53	***	[0.23,0.49]
	Speaker Confidence	34.55	***	Conf—Neut		0.45	0.08	5.60	***	[0.26,0.64]
				Conf—Unconf		0.07	0.08	0.82	1.00	[−0.13,0.26]

(Continued)

TABLE 4 (Continued)

Prosodic features	Effect	Chisq	*p* value[a]	Contrast	Estimate	SE[b]	*t*	*p* value	95%CI
				Neut—Unconf	−0.38	0.08	−4.74	***	[−0.58,-0.19]
	Lexical Tone × Speaker Confidence	0.80	0.796						

[a]Significance levels under Bonferroni-corrections: *$p < 0.05$; **$p < 0.01$; and ***$p < 0.001$. [b]SE, standard error.

2016) have indicated that the first formant frequency (F1) was typically shown to reflect tongue height, and the F2 was related to the size of the frontal oral cavity or the degree of tongue advancement. The F1 was typically reduced when a high vowel such as /i/ or /u/ pulled the tongue out of pharynx, moved the tongue upward and subsequently increased the volume of the pharynx. The F2 frequency was reduced when the vowel like /a/ or /u/ was produced with the tongue moving far back in the oral cavity (Lieberman and Blumstein, 1988). However, such different articulatory mechanisms underlying F1 and F2 were blurred in a recent study comparing vowels under different consonant contexts (i.e., \h\ + Vowel+\d\ and \d\ + Vowel+\d\ in female speech), which did not demonstrate a universal correlation pattern between tongue positions and formant frequencies. It is shown that F2 is a much more complex reflection of tongue variation in both tongue height and tongue advancement while the F1 variation unambiguously reflects tongue height (Lee et al., 2016).

The relation between formant frequencies and speech articulatory mechanisms allows the possibility for the speaker to encode social-pragmatic meaning, in particular, different levels of confidence in the present study by modulating the articulatory structure and further by moving their tongue positions. Previous works has shown an association between formant placement and speaker emotion. The first and second formants in certain vowels /i/, /u/ and /a/ of 12 emotions varied as a function of the emotional dimension in the tone of voice. While the higher-arousal emotional states resulted in a higher mean values in F1 in all vowels, the positive valence resulted in higher mean values in F2 (Laukka et al., 2005; Goudbeek et al., 2009). The formant encoding of speaker emotion could reflect the articulatory to acoustic mapping. It is likely that the increased feeling of knowing in the confident voice (Guyer J. J. 2016; Jiang and Pell, 2017) could possibly elicit an increased arousal of the speaker, therefore modulating their efforts to articulate vowels by raising the F1 and F2. The formant-frequency values are effectively determined by vowel type (the inter-vowel variability) and vocal tract length (the intra-vowel variability; Turner et al., 2009). Human speakers lower formants by increasing apparent vocal tract length (VTL). They also use formant information to change their own perceived social attributes (e.g., body size, Pisanski et al., 2016) or to perceive the social attributes of others (e.g., speaker height, Barreda, 2016). Accordingly, the innovative finding of this study is that the speaker's level of confidence influences the change in formants, possibly due to their efforts to encode socio-pragmatic meanings. However, it has also been observed that changes in tongue/lip positions can affect vocal tract length changes. The position of three articulatory parameters appears to contribute significantly to the instantaneous length of the vocal tract: lip, tongue dorsum, and larynx height (Dusan, 2007). The question of whether the resonance peaks encoding the speaker confidence are modulated by the change in VTL or tongue/lip position awaits further explorations with physiological measurements (e.g., MRI). Therefore, although formant cues usually serve as a stable acoustic indicator for distinguishing vowel identity, speakers can encode vocal expression of confidence through these stable characteristics. It is noted that the effect size of the formant characteristics was smaller than that of the prosodic features in the present study, suggesting a relative contribution of segmental vs. suprasegmental features in encoding vocal dynamic cues of speaker confidence (Zhang et al., 2021).

Encoding speaker confidence in prosodic features

Previous studies have demonstrated the effects of confident voice expressions on suprasegmental features in English spoken sentences (Jiang and Pell, 2014, 2017). The neutrally-intending and confident-intending expression seemed to be differentiated in prosodic cues of vowels, however, the neutrally-intending expression was judged close to confident (Jiang and Pell, 2014) or comparable to confident expression in the believability judgment (Jiang and Pell, 2018). Even though, the perceptual consequences between confident and neutrally-intending voices can be perceptually more similar than between confident and unconfident ones, prosodic marking can be quite distinctive in confident and neutral-intending ones to achieve the speaker's high feeling of knowing (Jiang and Pell, 2017).

In a dialect with rich tonal possibilities, the suprasegmental pitch encoding of confidence in vowels showed similar mechanisms from that in the longer spoken units. The pattern of mean pitch in vowels of our current results as a function of the intended speaker's confidence resembled the same patterns in previous studies on sentences based on the perceived level of confidence, with both showing the highest normalized mean f0 in

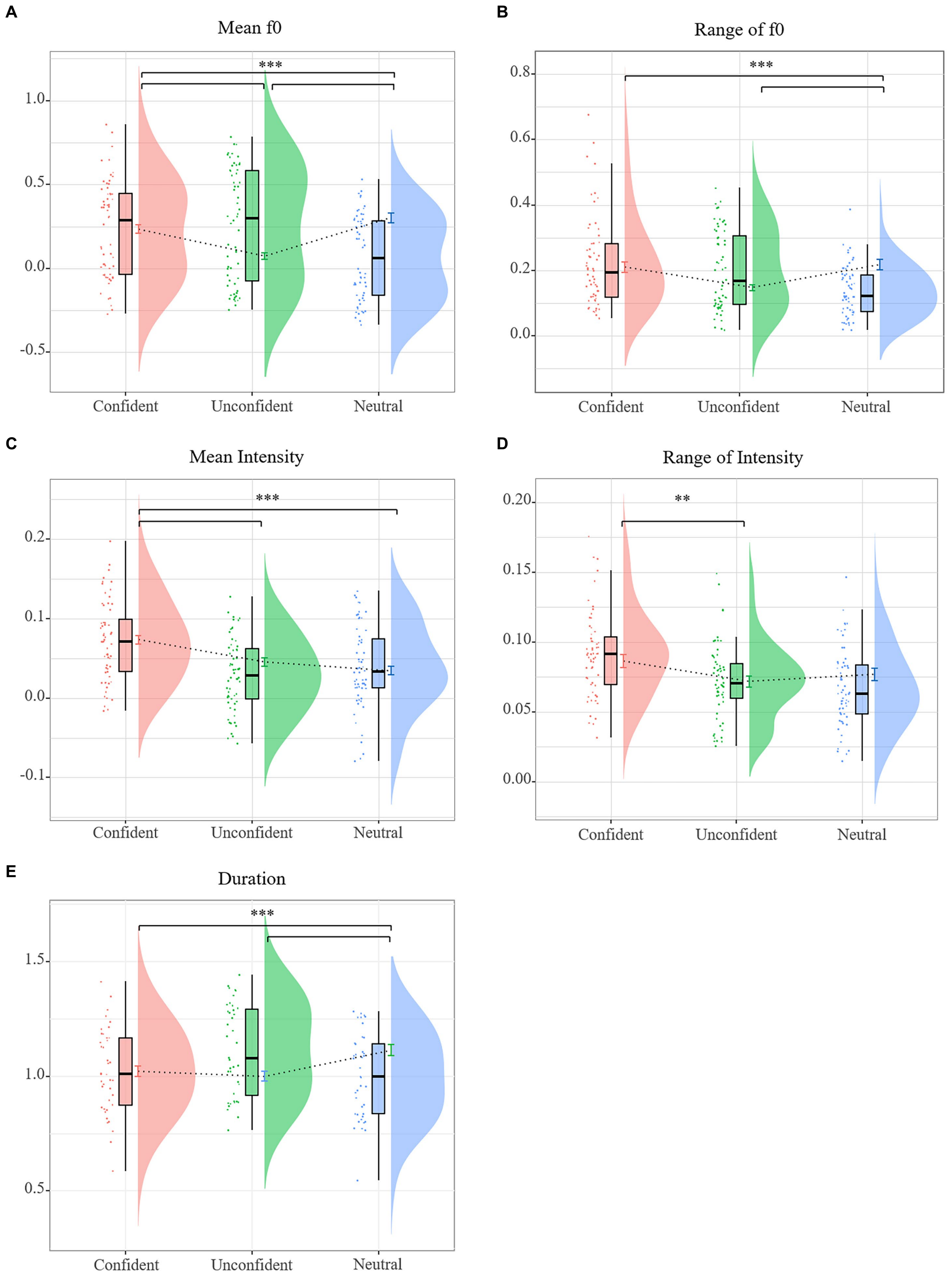

FIGURE 3
Raincloud plots for prosodic features showing the main effect of speaker confidence. **(A)** mean f0, **(B)** f0 range, **(C)** mean intensity, **(D)** intensity range per confidence level for all vowels, and **(E)** duration for monophthongs.

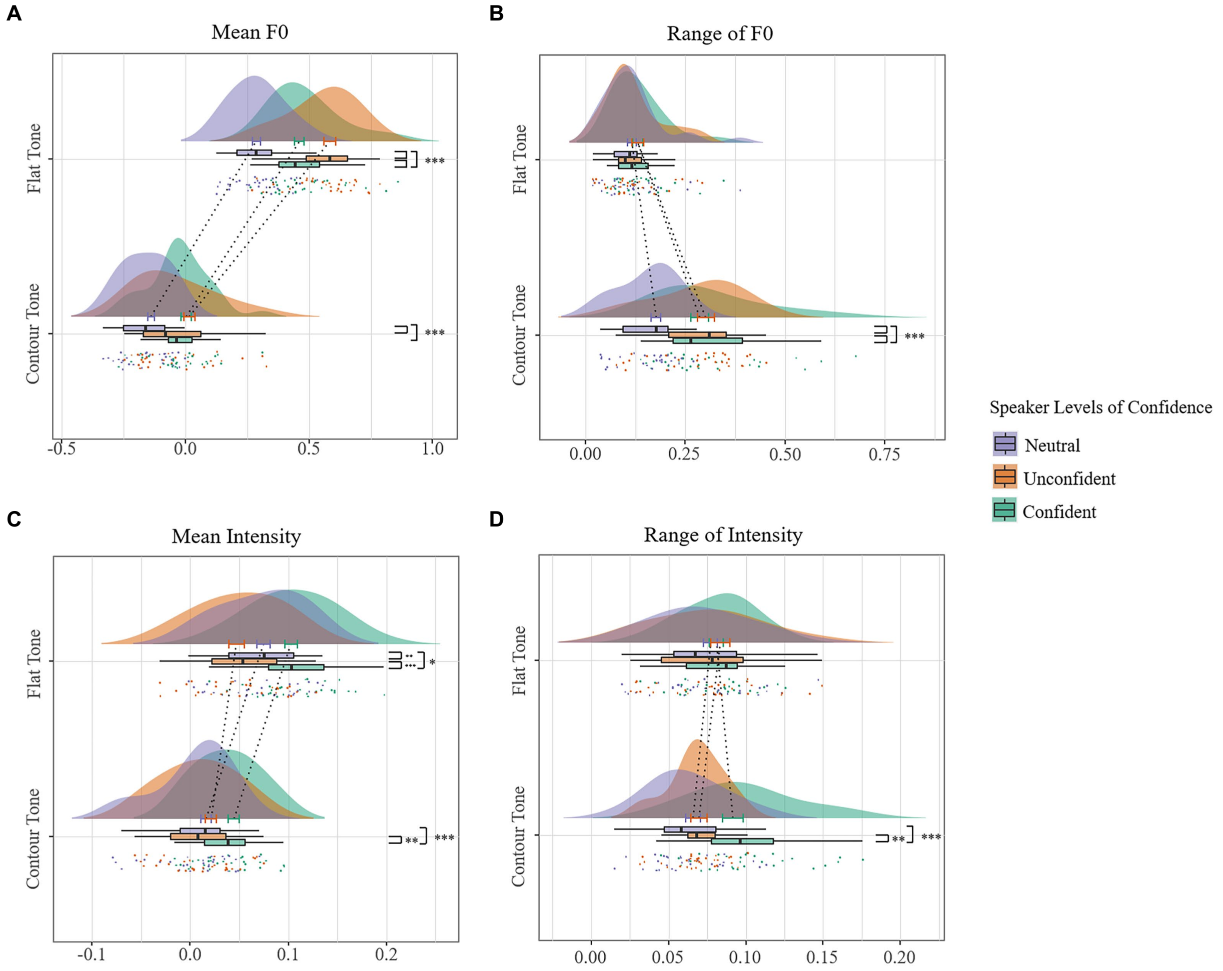

FIGURE 4
Raincloud plots for prosodic features showing the interaction of speaker confidence and lexical tone. **(A)** mean f0, **(B)** f0 range, **(C)** mean intensity, and **(D)** intensity range per lexical tone per confidence level for all vowels.

the unconfident level, followed by gradually decreased f0 over the confident and the neutral level. A similar pattern of f0 range also occurred in vowels. Speakers varied f0 to a larger extent when encoding confidence-related information in the voice. These findings suggested that speakers and listeners showed consistency regarding how fundamental frequency encodes speaker feeling of knowing no matter how long the stimuli are.

Past studies have revealed a strong relationship between a speaker's f0 variation and the perceived attractiveness (Xu et al., 2013), trustworthiness (McAleer et al., 2014), sarcasm (Jansen and Chen, 2020), and speakers' intended stress (Eriksson et al., 2013) at the lexical or the sentence level inferred from their voice. A further study found that a single word *hello* was enough for the listeners to distinguish speakers of different trustworthiness. The *hello* judged as trustworthy was characterized by a high starting f0 then a marked decrease at mid-utterance to finish on a strong rise (Belin et al., 2017). Additionally, a study asked listeners to judge spoken words of which the pitch contour was manipulated (Ponsot et al., 2018). They showed that sounds rated as trustworthy showed a rapid pitch increase on the second syllable of the word while sounds rated as dominant showed a gradual pitch decrease on both syllables. The modulation of f0 on speakers' intended confidence was consistent with a view that vocal tract length could serve as a functional role in one's socio-communicative ability. Speakers can volitionally modulate vocal parameters to imitate voices of different pitches and preferred to adjust f0 (and vocal fold tension in the vocal tract) downward and upward to imitate lower or higher pitched voices when asked to exaggerate body size during speech (Waters et al., 2021). It is suggested that, to encode socio-pragmatic information such as lack of confidence and credibility at the word level, the speaker could mark their voice with more dynamic pitch (Belin et al., 2017; Goupil et al., 2021b).

Our findings on mean intensity and intensity range were generally consistent with the findings on sentence. On vowels in the current study, the normalized mean intensity was higher when the speaker's intended confidence than lack of confidence or no emotion or confidence. The intensity range was larger when the speaker's intended high confidence than no emotion or confidence was encoded in the voice. Consistent to the previous studies based on the listener's perceived, speakers perceived to be unreliable (i.e., uncertain or dishonest) pronounced words with more variable pitch and speech rate, as well as a reduced intensity at the beginning of the word (Goupil and Aucouturier, 2021a; Goupil et al., 2021b). This means that a less certain speaker typically sounded less louder, which could serve as a possible explanation why the intensity of unconfident expressions was smaller than confident expressions. Compared with the neutral expression, speakers varied their voice intensity to a greater extent under either level of confidence (Jiang and Pell, 2017). Like speaker unreliability which was marked by vocal cues of unstable intensity to encode one's dishonesty and uncertainty, intensity variation can be dramatic to encode speaker levels of confidence.

The pattern on duration showed that speakers were able to use temporal cues to mark to the difference between no intended confidence and intended confidence. Speakers prolonged the pronunciation time when they intended to be confident or unconfident compared with they were refrained from emotions and attitudes. Duration has been associated with communicative meanings (e.g., Speaker persuasiveness: Scherer et al., 1973; Jiang et al., 2017; Speech acts: Hellbernd and Sammler, 2016; Speaker emotion, Banse and Scherer, 1996; Sauter et al., 2010). This finding added novel data to the previous studies on the role of temporal cues on encoding speaker's confidence information in the small unit of vowels.

Role of lexical tone in vocal expression of confidence

Despite pitch and loudness were both essential to the encoding of socio-pragmatic meanings (Jiang and Pell, 2017; Caballero et al., 2018; Pell and Kotz, 2021), they seemed to act in concert with the lexical tone to form complex interactive patterns when encoding speaker confidence. A previous study (Zhang et al., 2021) on weighting patterns of different acoustic parameters in encoding prominence in four mandarin tones showed that, on the syllable of flat tones, the mean, maximal and minimal pitch contributed more for marking prominent syllables than mean intensity; while on the syllable of contour tones, the mean intensity and intensity variation weighed higher than pitch-related features. Consistent with these findings, the speaker modulated their mean pitch to a greater extent in the flat tone than the contour tone and demonstrated a stronger modulation of intensity variation in the contour tone than the flat tone to distinguish between the confident and the neutral-intending vowels. Taken together, the speaker tended to modulate mean f0 and intensity levels on the flat tone whereas they tended to vary f0 and intensity level on the contour tone when encoding different levels of communicative meaning. An ERP study investigating the online processing of tone and intonation in Mandarin sentences showed that native Mandarin listeners can distinguish between question intonation and statement intonation when the intonation is associated with a final Tone 4, but fail to do so when the intonation is associated with a final Tone2, which indicated that the processing of intonation can be rapidly influenced by different lexical tones (Liu et al., 2016).

Studies on the interaction between boundary tone and affective prosody showed two patterns how lexical tone and intonation added up: the simultaneous addition of lexical tone of the boundary syllables and sentence intonation or the successive addition of the sentence intonation to the end of the lexical tones instead of simultaneously to the last syllables (Chao, 1933). A previous study (Li et al., 2011) with monosyllabic utterances showed that speakers used a successive addition pattern to express the speakers' emotion, with the falling successive tone to express disgust and angry and the rising successive tone to express happy and surprise.

According to account of successive addition, the expressive tone was added on the lexical tone by prolonging the duration after the lexical tones are completed. The current findings of longer duration when the speaker expressed confident information compared with the neutral expression suggest that the expressive tone seemed to be successively added to the end of the lexical tones to encode of confidence-related suprasegmental features on different lexical tones. The pattern of successive addition tones in the duration had no difference between the flat tone and the contour tone which indicated the same addition pattern that the expressive tone of confidence was added to both the flat tone and contour tone. Interestingly, the current findings of f0 features suggest that the expressive tone seemed to also affect the f0 contour of the lexical tones. Compared with the neutral-intending expression, the speakers raised mean f0 and had a greater variation of f0 in the unconfident level. Based on the above results, the vocal expression could be added on the lexical tones by a successive addition which was similar to the emotional expressions found in previous studies that were added on the lexical tones by the way of successive addition. Pending more investigations, this finding could expand the successive addition tone account by showing how vocal expression of confidence interacted with lexical tone.

Limitation and future directions

This study focused on the segmental and suprasegmental representation of speakers' intended confidence using vowels in a Chinese dialect with a rich tonal system. Dual-route approach of speech communication has assumed the speaker encodes meaning in vocal cues at both linguistic and social level (Sumner et al., 2014; Sumner, 2015). Considering the listeners can automatically and rapidly map of co-present cues (tone, dialect) in speech to recognize social attributes of speakers (Sumner et al., 2014), the speakers due to this reason encode the confidence expression in the segmental and suprasegmental level of vowels. Therefore, the interaction between vocal expression and lexical tones observed on pitch cues provides ingredients to further investigations on how the addition patterns supra-segmental and segmental cues affect listener perception of speaker socio-communicative meanings.

Most previous researches focused on how speakers encode communicative meanings based on standard languages used typically in a formal setting (e.g., English, Mandarin, etc.,), but few has extended the findings to variations of languages typically used in a less-formal setting (e.g., dialect, accented-speech, Jiang et al., 2018, 2020). Comparing native English speakers and English second-language (L2) learners in the acoustic encoding of persuasiveness, a study showed that the consonantal durations, particularly those of continuants, were significantly longer relative to the vowels that followed them when native speakers intended persuasiveness, while for second language learners, the duration of consonants did not change between the neutral-intending and persuasive speech (Banzina, 2021). Speakers of different accents displayed different pronunciation strategies of using phonetic cues in characterizing socio-communicative meanings. In a machine learning experiment of listeners' perception of confidence and doubt in speakers with different accents, while durational feature contributed to a larger extent in the native accent, the mean and range of intensity contributed more in the foreign and regional accent for the speaker to be perceived with different certainties (Jiang and Pell, 2018). The issue regarding how socio-pragmatic information is encoded in informal dialects and non-standard variations of languages awaits further investigations.

Although the materials were validated by independent listeners, the speakers did not provide their own assessment on the vowels in the current study. In further studies, assessing the self-rated confidence expression after elicitation is necessary to confirm the confidence levels based on speaker's intention to directly compare how listeners and speakers use vocal cues to decode different levels of speaker confidence.

Future researches could enhance the generalizability of the present findings by adding more speakers considering the limited speakers in the present study and taking into consideration different speech acts and attitudes to dialects. Considering the non-spontaneous elicitation of vowels in the laboratory, the logic follow-up is to do a more naturalistic study by using a spontaneous elicitation procedure, for instance, to respond to the conversational partner with certain communicative.

While a possible articulatory mechanism was inferred based on acoustic results of the current study, the acoustic parameters remained indirect clues. Combined with the role of formant cues in differentiating confident from unconfident and neutral-intending speech, the speech-motor mechanism of the larynx and tongue should be validated to explore the internal articulatory mechanism and its vocal movement through physiological measurement.

Conclusion

Employing an expression elicitation paradigm for different vocal expression in Wuxi dialect vowels, this acoustic-phonetic study explored the segmental and suprasegmental acoustic representation of confident, unconfident and neutral-intending speech in vowels. Compared with the neutral-intending expression, the speakers raised F1, F2, mean intensity and had a greater variation of intensity in the confident level, while they raised mean f0 and had a greater variation of f0 in the confident level. Additionally, only F1 can distinguish between the confident and unconfident expressions. More importantly, we showed that lexical tone modulated the acoustic encoding of speaker confidence levels in vowels. Specifically, the speaker modulated the mean f0 and mean intensity to a larger extent on the flat tone than the contour tone to differentiate levels of confidence in voice, while they modulated the range of f0 and intensity more on the contour tone than the flat tone. Tonal cues in the Wuxi dialect have an indispensable role in encoding different levels of confidence.

Data availability statement

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

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Committee in Institute of Lingusitics, Shanghai International Studies University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

YJ wrote the first manuscript, performed experiments, and analyzed data for this manuscript. YH analyzed the data, edited the manuscript, and prepared the figures. XJ supervised the study and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was sponsored by Natural Science Foundation of China (31971037), the Shanghai Planning Office of Philosophy and Social Sciences (2018BYY019), the "Shuguang Program" supported by Shanghai Education Development Foundation, Shanghai

Municipal Education Committee (20SG31), and Major Program of National Social Science Foundation of China (No. 18ZDA293).

Acknowledgments

We are very grateful to Ping Xu, Musheng Xu, Meijun Sun, and Yudu Hua for their production and the members of CBCS Shanghai for their assistance in the study.

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.2022.1028106/full#supplementary-material

References

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

Banzina, E. (2021). Exploring phonetic cues to persuasive oral presentation: a study with British English speakers and English L2 learners. *Lang. Teach.* In press. Available at: https://journals.sagepub.com/doi/abs/10.1177/13621688211037610 (Accessed November 22, 2022). doi: 10.1177/13621688211037610

Barreda, S. (2016). Investigating the use of formant frequencies in listener judgments of speaker size. *J. Phon.* 55, 1–18. doi: 10.1016/j.wocn.2015.11.004

Barreda, S., and Nearey, T. M. (2011). Formant frequencies, vowel identity, and the perceived relative tallness of synthetic speakers. *J. Acoust. Soc. Am.* 130, 2443–2443. doi: 10.1121/1.3654801

Bauerly, K. R. (2018). The effects of emotion on second formant frequency fluctuations in adults who stutter. *Folia Phoniatr. Logop.* 70, 13–23. doi: 10.1159/000488758

Belin, P., Boehme, B., and McAleer, P. (2017). The sound of trustworthiness: acoustic-based modulation of perceived voice personality. *PLoS One* 12, e0185651–e0185612. doi: 10.1371/journal.pone.0185651

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments*. California: University of California Press.

Caballero, J. A., Vergis, N., Jiang, X., and Pell, M. D. (2018). The sound of im/politeness. *Speech Commun.* 102, 39–53.

Caffi, C., and Janney, R. W. (1994). Toward a pragmatics of emotive communication. *J. Pragmat.* 22, 325–373. doi: 10.1016/0378-2166(94)90115-5

Cao, X. (2003). The research about Wuxi dialect. (Thesis, Suzhou University).

Chao, Y. R. (1933). Tone and intonation in Chinese. *Bull. Inst. Hist. Philol. Acad. Sin.* 4, 121–134.

Coates, J. (2012). "The role of epistemic modality in women's talk," in *Modality in Contemporary English*. eds. R. Facchinetti, F. Palmer and M. Krug (Berlin, Boston: De Gruyter Mouton), 331–348.

Cutler, A., and Chen, H.-C. (1997). Lexical tone in Cantonese spoken-word processing. *Percept. Psychophys.* 59, 165–179. doi: 10.3758/BF03211886

Dusan, S. (2007). *Vocal Tract Length During Speech Production*. Interspeech in Antwerp, Belgium. 1366-1369.

Eady, S. J. (1982). Differences in the F0 patterns of speech: tone language versus stress language. *Lang. Speech* 25, 29–42. doi: 10.1177/002383098202500103

Eriksson, A., Barbosa, P. A., and Akesson, J. (2013). *The Acoustics of Word Stress in Swedish: A Function of Stress Level, Speaking Style and Word Accent.* Interspeech in Lyon, France, 778–782.

Fant, G. (1960). *Acoustic Theory of Speech Production.* Mouton & Co, The Hague, Netherlands.

Goudbeek, M., Goldman, J. P., and Scherer, K. R. (2009). Emotion dimensions and formant position, in Tenth Annual Conference of the International Speech Communication Association, Brighton, UK, 6–10 September. Available at: https://www.isca-speech.org/archive_v0/archive_papers/interspeech_2009/papers/i09_1575.pdf (Accessed on November 22, 2022).

Fisher, N. I., and Hall, P. (1991). Bootstrap algorithms for small samples. *Journal of Statistical Planning and Inference* 27, 157–169. doi: 10.1016/0378-3758(91)90013-5

Goupil, L., and Aucouturier, J.-J. (2021a). Distinct signatures of subjective confidence and objective accuracy in speech prosody. *Cognition* 212:104661. doi: 10.1016/j.cognition.2021.104661

Goupil, L., Ponsot, E., Richardson, D., Reyes, G., and Aucouturier, J.-J. (2021b). Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nat. Commun.* 12:861. doi: 10.1038/s41467-020-20649-4

Guyer, J. (2016). Investigating multiple roles of vocal confidence in persuasion. Doctoral dissertation. Available at: http://hdl.handle.net/1974/15054

Guyer, J. J. (2016). The influence of vocally expressed emotions on attitude change. *Mind Pad*, 21–25. Available at: https://www.researchgate.net/profile/Joshua-Guyer/publication/293335329_The_influence_of_vocally_expressed_emotions_on_attitude_change/links/574dcc2608ae82d2c6be242c/The-influence-of-vocally-expressed-emotions-on-attitude-change.pdf#page=21 (Accessed on November 22, 2022).

Hellbernd, N., and Sammler, D. (2016). Prosody conveys speaker's intentions: acoustic cues for speech act perception. *J. Mem. Lang.* 88, 70–86. doi: 10.1016/j.jml.2016.01.001

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

Jansen, N., and Chen, A. (2020). Prosodic encoding of sarcasm at the sentence level in Dutch. *Speech Prosody* 2020, 409–413. doi: 10.21437/SpeechProsody.2020-84

Ji, Y., and Jiang, X. (2021). "A study of confident voices in the Wuxi dialect based on formants" in *Paper presented at the meeting of the 14th Phonetic Association of China*, Lanzhou, China.

Jiang, X., Sanford, R., and Pell, M. D. (2018). Neural architecture underlying person perception from in-group and out-group voices. *NeuroImage* 181, 582–597.

Jiang, X., Gossack-Keenan, K., and Pell, M. D. (2020). To believe or not to believe? How voice and accent information in speech alter listener impressions of trust. *Quarterly Journal of Experimental Psychology* 73, 55–79. doi: 10.1177/1747021819865833

Jiang, X., and Lu, L. (2021). "A study of confident voices based on Stop VOT" in Paper presented at the Meeting of the 14th Phonetic Association of China, Lanzhou, China.

Jiang, X., and Pell, M. D. (2014). "Encoding and decoding confidence information in speech." in *Proceedings of the International Conference on Speech Prosody*, May, 573–576.

Jiang, X., and Pell, M. D. (2015). On how the brain decodes vocal cues about speaker confidence. *Cortex*. 66, 9–34. doi: 10.1016/j.cortex.2015.02.002

Jiang, X., and Pell, M. D. (2016). Neural responses towards a speaker's feeling of (un) knowing. *c* 81, 79–93. doi: 10.1016/j.neuropsychologia.2015.12.008

Jiang, X., Sanford, R., and Pell, M. D. (2017). Neural systems for evaluating speaker (Un) believability. *Hum. Brain Mapp.* 38, 3732–3749. doi: 10.1002/hbm.23630

Jiang, X., and Pell, M. D. (2017). The sound of confidence and doubt. *Speech Comm.* 88, 106–126. doi: 10.1016/j.specom.2017.01.011

Jiang, X., and Pell, M. (2018). "Predicting confidence and doubt in accented speakers: human perception and machine learning experiments." in *Proceedings of the 9th International Conference in Speech Prosody*, 269–273.

Jiang, X., Sanford, R., and Pell, M. D. (2017). Neural systems for evaluating speaker (Un) believability. *Hum. Brain Mapp.* 38, 3732–3749. doi: 10.1002/hbm.23630

Juslin, P. N., and Laukka, P. (2003). Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol. Bull.* 129, 770–814. doi: 10.1037/0033-2909.129.5.770

Kuhlen, A. K., Bogler, C., Swerts, M., and Haynes, J.-D. (2015). Neural coding of assessing another person's knowledge based on nonverbal cues. *Soc. Cogn. Affect. Neurosci.* 10, 729–734. doi: 10.1093/scan/nsu111

Ladefoged, P., Harshman, R., Goldstein, L., and Rice, L. (1978). Generating vocal tract shapes from formant frequencies. *J. Acoust. Soc. Am.* 64, 1027–1035. doi: 10.1121/1.382086

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

Lee, J., Shaiman, S., and Weismer, G. (2016). Relationship between tongue positions and formant frequencies in female speakers. *J. Acoust. Soc. Am.* 139, 426–440. doi: 10.1121/1.4939894

Li, A., Fang, Q., and Dang, J. (2011). "Emotional intonation in a tone language: experimental evidence from Chinese." in 17th international congress of phonetic sciences (ICPhS) XVII regular session, August, 1198–1201.

Li, A., Fang, Q., and Dang, J. (2012). "Emotional expressiveness of successive addition boundary tone in mandarin Chinese" in Proceedings of the 6th International Conference on Speech Prosody, SP 2012, 2, 591–594.

Lieberman, P., and Blumstein, S. E. (1988). S*peech physiology, speech perception, and acoustic phonetics*. Cambridge, MA: Cambridge University Press. doi: 10.1017/cbo9781139165952

Liu, X., and Chen, L. (2018). Anhui Wuhu liulang fangyan yuyin xitong [the homophony syllabary of Wu dialect in Liulang town, Wuhu county in Anhui province]. *Fangyan* 3, 276–286. Available at: http://chinese-thought.ecnu.edu.cn/60/f9/c35454a418041/page.htm (Accessed November 22, 2022)

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

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

London, H., McSeveney, D., and Tropper, R. (1971). Confidence, overconfidence and persuasion. *Hum. Relat.* 24, 359–369. doi: 10.1177/001872677102400502

London, H., Meldman, P. J., and Lanckton, A. (1970a). The jury method: how the persuader persuades. *Public Opin. Q.* 34, 171–183. doi: 10.1086/267787

London, H., Meldman, P. J., and Lanckton, A. V. C. (1970b). The jury method: some correlates of persuading. *Hum. Relat.* 23, 115–121.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing type I error and power in linear mixed models. *J. Mem. Lang.* 94, 305–315. doi: 10.1016/j.jml.2017.01.001

Menn, L., and Boyce, S. (1982). Fundamental frequency and discourse structure. *Lang Speech* 25, 341–383.

McAleer, P., Todorov, A., and Belin, P. (2014). How do you say "hello"? Personality impressions from brief novel voices. *PLoS One* 9, 1–9. doi: 10.1371/journal.pone.0090779

Pell, M. D., and Kotz, S. A. (2021). Comment: the next frontier: prosody research gets interpersonal. *Emot. Rev.* 13, 51–56. doi: 10.1177/1754073920954288

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., and Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: a comparison of four languages. *J. Phon.* 37, 417–435. doi: 10.1016/j.wocn.2009.07.005

Pisanski, K., Mora, E. C., Pisanski, A., Reby, D., Sorokowski, P., Frackowiak, T., et al. (2016). Volitional exaggeration of body size through fundamental and formant frequency modulation in humans. *Sci. Rep.* 6:34389. doi: 10.1038/srep34389

Ponsot, E., Burred, J. J., Belin, P., and Aucouturier, J. J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proc. Natl. Acad. Sci. U. S. A.* 115, 3972–3977. doi: 10.1073/pnas.1716090115

Salais, L., Arias, P., Le Moine, C., Rosi, V., Teytaut, Y., Obin, N., et al. (2022). Production Strategies of Vocal Attitudes. *Interspee Shanxi qianyang fangyan yuyin de tedian he yanbianch in Incheon, Korea*, 4985–4989. https://doi.org/10.21437/Interspeech.2022-10947

Sauter, D. A., Eisner, F., Calder, A. J., and Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *Q. J. Exp. Physiol.* 63, 2251–2272. doi: 10.1080/17470211003721642

Scherer, K. R., London, H., and Wolf, J. J. (1973). The voice of confidence: paralinguistic cues and audience evaluation. *J. Res. Pers.* 7, 31–44. doi: 10.1016/0092-6566(73)90030-5

Singmann, H. (2019). "Afex: analysis of factorial experiments." Available at: https://github.com/singmann/afex/ (Accessed May 22, 2022).

Sumner, M. (2015). The social weight of spoken words. *Trends Cogn. Sci.* 19, 238–239. doi: 10.1016/j.tics.2015.03.007

Sumner, M., Kim, S. K., King, E., and McGowan, K. B. (2014). The socially weighted encoding of spoken words: a dual-route approach to speech perception. *Front. Psychol.* 4, 1–13. doi: 10.3389/fpsyg.2013.01015

Swerts, M., and Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *J. Mem. Lang.* 53, 81–94. doi: 10.1016/j.jml.2005.02.003

Turner, R. E., Walters, T. C., Monaghan, J. J. M., and Patterson, R. D. (2009). A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data. *J. Acoust. Soc. Am.* 125, 2374–2386. doi: 10.1121/1.3079772

Van Zant, A. B., and Berger, J. (2020). How the voice persuades. *J. Pers. Soc. Psychol.* 118, 661–682. doi: 10.1037/pspi0000193

Waters, S., Kanber, E., Lavan, N., Belyk, M., Carey, D., Cartei, V., et al. (2021). Singers show enhanced performance and neural representation of vocal imitation. *Philos. Trans. R. Soc. B* 376:399. doi: 10.1098/rstb.2020.0399

Wen, S. (1996). A brief account of the phonology of Wuxi dialect. *J. Wuxi Educ. Coll.* 04, 36–38.

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

Xu, Y., and Prom-On, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: synthesizing speech melody via model-based stochastic learning. *Speech Comm.* 57, 181–208. doi: 10.1016/j.specom.2013.09.013

Zhang, Y. (2020). Shanxi qianyang fangyan yuyin de tedian he yanbian [phonological features and evolution of Qianyang dialect in Shanxi province]. *Fangyan* 2, 244–256. Available at: https://www.cnki.com.cn/Article/CJFDTotal-FYZA202002017.htm (Accessed on November 22, 2022)

Zhang, W., Clayards, M., and Zhang, J. (2021). "Effects of mandarin tones on acoustic Cue weighting patterns for prominence." in *2021 12th International Symposium on Chinese Spoken Language Processing, ISCSLP 2021*.

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 15 December 2022
DOI 10.3389/fcomm.2022.1029400

OPEN ACCESS

EDITED BY
Sandra Madureira,
PUCSP, Brazil

REVIEWED BY
Julien Meyer,
Centre National de la Recherche Scientifique (CNRS), France
Alexsandro Rodrigues Meireles,
Federal University of Espirito Santo, Brazil

*CORRESPONDENCE
Samuel K. Akinbo
akinbosk@gmail.com

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

RECEIVED 27 August 2022
ACCEPTED 28 November 2022
PUBLISHED 15 December 2022

CITATION
Akinbo SK, Samuel O, Alaga IB and Akingbade O (2022) An acoustic study of vocal expression in two genres of Yoruba oral poetry.
Front. Commun. 7:1029400.
doi: 10.3389/fcomm.2022.1029400

An acoustic study of vocal expression in two genres of Yoruba oral poetry

Samuel K. Akinbo[1]*, Olanrewaju Samuel[2], Iyabode B. Alaga[2] and Olawale Akingbade[2]

[1]Department of Linguistics, University of British Columbia, Vancouver, BC, Canada, [2]Department of Linguistics and African Languages, University of Ibadan, Ibadan, Nigeria

This pilot study proposes an acoustic study of the vocal expressions in Ìjálá and Ẹ̀sà, two genres of Yorùbá oral poetry. For this study, we conducted an experiment, involving the vocalization of an original poem in speech mode, Ìjálá and Ẹ̀sà. The vocalizations were recorded and analyzed acoustically. The results of the study show that cepstral peak prominence (CPP), Hammarberg index and Energy of voiced sound below 500 Hz distinguish comparisons of Ẹ̀sà, Ìjálá and speech but are not as reliable as F0 height and vibrato. By comparing the pitch trajectories of the speech tones and poetic tunes, we show that poetry determines tone-to-tune mapping but can accommodate language when it is feasible. The results of our investigation are not only in line with the previous impressionistic observations about vocal expression in Yorùbá oral poetry but contribute with new findings. Notably, our investigation supports vocal tremor as the historical origin of vibrato in Ìjálá. As a result of this, we strongly recommend the instruments of phonetic science for the study of vocal expression in African oral poetry.

KEYWORDS

oral poetry, tone, vibrato, vocal effort, vocal expression, phonetics

1. Introduction

One of the major challenges with the study of African oral poetry is lack of instrumental analysis, even when the analysis of vocal expression is based on terms with phonetic correlates (Babalọla, 1963; Ajuwon, 1977; Ọlabimtan, 1977; Ọlátúnjí, 1979). In this preliminary study, we address this issue by proposing an acoustic analysis of vocal expression in two genres of Yorùbá oral poetry.

Studies show that vocal expressions, such as pitch raising and increased loudness, communicate emotions (Scherer, 1985; Juslin and Laukka, 2003). For instance, high pitch and increased loudness are associated with high level arousal such as anger, happiness and excitement (Banse and Scherer, 1996; Juslin and Laukka, 2003; Johnstone et al., 2005; Goudbeek and Scherer, 2010; Lindquist et al., 2016; Scherer, 2021). The use of vocal expression is not limited to affective communication. For example, in many African oral traditions, various genres of verbal art are distinguished based on the vocal expressions that are associated with them (e.g., Uzochukwu 1981 on Igbo elegiac poetry of Nigeria; Boadi 1989 on Akan praise poem of Ghana; Brown 1995 on indigenous South African oral poetry Ọlátúnjí 1979 on Yorùbá oral poetry). The acoustic correlates of vocal expressions in affective speech have been extensively investigated (see the reviews in Scherer, 1986; Juslin and Laukka, 2003; Kamiloğlu et al., 2020), but the vocal expression in African oral traditions have mostly been analyzed without the instruments of phonetic science.

The present paper describes an acoustic study of vocal expressions in two genres of Yorùbá oral poetry. The study is based on Ìjálá and Ẹ̀sà, two genres of Yorùbá oral poetry. We will argue that the instrument of acoustic phonetics can offer valuable insights on the verbal aesthetics of African oral poetry.

Before turning to the details of this study, we present the basic sound inventory of Yorùbá in Section 2. The discussion in Section 3 focuses on the features of Ìjálá and Ẹ̀sà in the context of Yorùbá oral poetry. To understand the phonetic correlates of vocal expression in Ìjálá and Ẹ̀sà, we conducted a production experiment. The details of the experimental procedure are presented in Section 4. The results of the experiment are presented in Section 5. The discussion and conclusion are presented in Section 6.

2. Background on Yorùbá sound inventory

Yorùbá is a Volta-Niger language spoken in West Africa and prominently South-Western Nigeria (Blench, 2019). This section presents a description of the relevant sound patterns in Standard Yorùbá, which is the basis of this work.

2.1. Tone in speech

Yorùbá is a tone language, which means pitch contrasts bring about lexical or grammatical distinctions in meaning (Yip, 2002; Hyman, 2018). The language has three contrastive tones, namely H(igh), M(id) and L(ow) (Akinlabi, 1985; Pulleyblank, 1986).

(1) Yorùbá: Tonal minimal set

H	bá	'to meet'
M	ba	'to braid'
L	bà	'to land'

As shown in (1), H tone is marked with an acute accent and L tone with a grave accent. However, M tone is unmarked (Bamgboṣe, 1965; Awóbùlúyì, 1978). Throughout this work, this tone-marking convention in Yorùbá orthography is adhered to.

2.2. Vowels

The vowel inventory of Yorùbá contains seven oral vowels and three nasal vowels, which are presented in (2) (Bamgboṣe, 1966; Awóbùlúyì, 1978; Pulleyblank, 1988).

(2) Yorùbá vowels (Pulleyblank, 2009, p. 868)

		Front	Central	Back
Oral	High	i [i]		u [u]
	Mid ATR	e [e]		o [o]
	Mid RTR	ọ [ɔ]		ẹ [ɛ]
	Low		a [a]	
Nasal	High	in [ĩ]		un [ũ]
	Low		an [ã]	

The phonetic transcription of the vowels is in square brackets and accompanied with the Standard Yorùbá orthography. Except in graphs and tables, the orthographic transcription is used throughout this paper. The low nasal vowel "an" is often pronounced as "ọn" (Bamgboṣe, 1966; Awóbùlúyì, 1978). Considering that the difference between "ọn" and "an" is phonetic, Pulleyblank (1988, p. 237) analyze the free variation between the vowels to be ‘a low-level phonetic effect." This phonetic distinction between "ọn" and "an" is not crucial to the goal of this paper. We now turn to how the tones and vowels are vocalized in Ìjálá and Ẹsà.

3. Basic description of Ìjálá and Ẹ̀sà

Ìjálá and Ẹ̀sà are some of the genres of oral poetry in Yorùbá culture. Most genres of Yorùbá oral poetry are associated with deities and ancestral devotion. Ìjálá is associated with Ògún, the Yorùbá deity of metallurgy and metal-related works (Babalọla, 1963; Ajuwon, 1977; Ọlátúnjí, 1979). Ògún is considered the patron of people who engage in metal works such as blacksmiths, goldsmiths, hunters and professional drivers. The devotees are obliged to pay homage to Ògún. One of the ways of paying homage to Ògún is through the chanting of oríkì, which is the embodiment of the eulogy and epithets about an entity, in this case about Ògún. The mode of performance of this oríkì in chant form is referred to as Ìjálá. In Yorùbá oral history, there are four hypotheses about the origin of Ìjálá. All the hypotheses point to the Yorùbá deity of metallurgy Ògún, but only two trace the origin of the vocal expression in Ìjálá to Ògún's geriatric voice and his alcohol consumption (see Babalọla, 1963, p. 6–12).

Ẹ̀sà, which is also known as Ìwí Egúngún, is associated with Egúngún creed of ancestral veneration (Olajubu, 1974; Adedeji, 1978). Periodically, the ancestral spirits physically manifest as Egúngún masquerade. To pay homage to the spirits of the departed ancestors, the devotees chant praises in Ẹ̀sà poetic mode.

Unlike most genres of music in the culture, most genres of Yorùbá oral poetry are not danceable and may not be accompanied with a drum performance. However, instrumental or vocal music can occur during the intermissions of the poetry performance (Babalọla, 1963; Ajuwon, 1977; Adedeji, 1978; Fámúle, 2018). The instrumental music might involve the representation of Yorùbá phrases with a talking drum (Euba, 1990; Villepastour, 2010; Akinbo, 2019, 2021). Depending on the genre or the mood of the performer, the oral performance may be closer to speech or music, thus they are considered semi-musical verbal arts (Babalọla, 1963; Ọlátúnjí, 1979; Ògúndejì, 1991). Similar to Yorùbá, the genres of oral poetry in other African societies are also semi-musical (Uzochukwu, 1981; Boadi, 1989; Okpewho, 1992; Finnegan, 2007, 2012; Purvis, 2009).

Studies suggest that Yorùbá oral poetry can be identified based on the contents of the poem, the identity of the performer and vocal expression (Gbadamosi and Beier, 1959; Ọlátúnjí, 1979). The traditional contents of Ìjálá and Ẹ̀sà are eulogy and epithets, but the contents may include historical events, personal eulogy of hunters and non-hunters, social commentaries, humor and all aspects of human existence (see Babalọla, 1963; Ọlátúnjí, 1979; Idamoyibo, 2006). Most importantly, the texts of a specific genre (e.g., Ìjálá) can be used for other genres of Yorùbá oral poetry (e.g., Ẹ̀sà) (Gbadamosi and Beier, 1959; Ọlátúnjí, 1979). As a result of this, textual contents are not reliable in distinguishing various genres of Yorùbá oral poetry. For example, the popular Ìjálá chanter, Ògúndáre Fọ́yánmu, is widely known for incorporating contemporary socio-political issues in his poems (Olaniyan, 2013). A noteworthy example is the Ìjálá chant of Fọ́yánmu about the historic Nigerian tax war of 1969 (see Adeniran, 1974)[1]. Other examples come from the syncretic practices of Yorùbá Christians and Muslims. Although Ìjálá is traditionally associated with Ògún, Yorùbá Christians and Muslims often incorporate Ìjálá chants (and analogously other genres of Yorùbá verbal art) in their religious worships (Idamoyibo, 2006, 2011; Ajibade, 2007; Olátúnjí, 2012; Dada, 2014). Thus, the content of a poem and the identity of the performer are not reliable in identifying genres of Yorùbá oral poetry.

There is consensus that various genres of Yorùbá oral poetry are best distinguished or classified based on vocal expression (e.g., Ọlátúnjí, 1979; Yai, 1989, etc.). For example, the contents of Ìjálá are always chanted in vibrato (Babalọla, 1963). Ẹ̀sà does not involve vibrato like Ìjálá, but a pattern of vowel insertion and lengthening (Olajubu, 1974; Adedeji, 1978). Regardless of the subject matter, a triply long vowel [ooo] at the end of the first poetic line is a recurrent characteristics of Ẹ̀sà (Olajubu, 1974; Olajubu and Ojo, 1977; Adedeji, 1978). When Christian and Islamic musicians incorporate Yorùbá oral poetry into their religious worships, the genres of oral poetry are recognized, not through the contents of the poem nor the identity of chanters, but the distinctive vocal expression which is associated with the chant.

Ìjálá shares the same vocal expression with Ìrèmọ̀jé, which is a funeral dirge for hunters (Babalọla, 1963; Ajuwon, 1977). Although Ìjálá can be adapted to suit any content, Ìrèmọ̀jé is restricted to funeral rites. As a result of this, there is an on-going debate as to whether Ìrèmọ̀jé is a distinct genre or a sub-genre of Ìjálá (see Babalọla, 1963; Ajuwon, 1977; Olajubu, 1984; Idamoyibo, 2006). The features of vocal expressions in Ìjálá have also been described in terms of rhythm (Babalọla, 1963; Ọlabimtan, 1977), but we do not consider rhythm in this work. Previous studies suggests that vocal expression in Yorùbá oral poetry involves stress given that it involves loudness and pitch raising (Siertsema, 1959; Babalọla, 1963; Ọlabimtan, 1977; Ọlátúnjí, 1979). The vocal expression in Yorùbá oral poetry and stress have loudness and pitch raising in common, but the phonetic properties are not as a result of stress in the oral poetry. Unlike stress which involves a syllable being prominent than the other in a word (Liberman and Prince, 1977; Halle and Idsardi, 1995; Kager, 2007), all the syllables of the words in Yorùbá oral poetry are produced with loudness and pitch raising (Babalọla, 1963). Most importantly, Yorùbá is a tone language, not a stress-timed language (Akinlabi, 1985; Pulleyblank, 1986; Kenstowicz, 2006).

In this work, we investigate the phonetic correlates of vocal expression in Ìjálá and Ẹ̀sà. Yorùbá, the textual bases of the poem, is a tone language, which means pitch contrasts bring about lexical distinctions. Considering that the melody of verbal arts such as chanting depends on pitch contour, we also investigate how poetic melodies interact with the linguistic demand of tone contrast. To answer the questions, we conducted an experiment. The details of the experiment are presented in the next section.

4. Methodology

4.1. Stimuli, participant and procedure

The stimulus in this work is an original poem which was composed by the third author of this paper. As shown in (3), the poem is written in Standard Yorùbá orthography. Oral performance is usually from memory, so in order to make it easier for the consultant to memorize, we selected the oríkí for its brevity. By selecting an original poem instead of a widely known traditional poem as the stimulus, we were able to control for the effect of content and familiarity.

One male native speaker of Standard Yorùbá was voluntarily recruited for this study. This consultant (age 28) was a fourth-year undergraduate of the Yorùbá study program at the Department of Linguistics and African Languages, University of Ibadan. The consultant had trainings in chanting various genres of Yorùbá oral poetry, including Ìjálá and Ẹ̀sà. A week

[1] An audio of the Ìjálá poem by Fọ́yánmu can be found in this link: https://www.youtube.com/watch?v=KzTBR7VJknQ.

before he was scheduled for a recording session, the consultant memorized the orìkí that was composed for the study. At the recording session, he was asked to recite the poem six times in normal speech mode. After reciting the poem in speech mode, he chanted the poem six times in Ìjálá and six times in Ẹ̀sà.

The renditions of the poem in speech mode and Ìjálá mode were recorded in a quiet room at the sampling rate of 48.1 kHz in wav format. Following Babalọla (1963, p. 121), each stretch "of utterances after a breath pause" is grouped as a line of the poem. In line with the observation in Ọlátúnjí (1979), the utterances within each pair of breath pauses form a meaningful whole. Based on the chanting of the poem in Ìjálá and Ẹ̀sà, the text of the poem were grouped into four lines.

(3) An original Yorùbá poem

Line 1 Adédùntán Àbẹ̀jẹ́ ọmọ Bàbálọ́jà
"Adeduntan, the child of Babaloja"
Line 2 ẹyinjúu Ọmọladùn baríọlá ọmọba Lépolóyin
"the eyeball of Omoladun, the honorable
princess of Lepoloyin"
Line 3 tẹ́ẹ́rẹ́ gbajó ọmọ̀ọ̀dọ̀ àgbà, ìdílẹ̀kẹ̀ ẹlẹ́rìn-ín ẹ̀yẹ
"suitably slim for dance, a wise child with
a bead-befitting waist"
Line 4 dúdú wù mí, dúdú dá'mi l'ọ́rùn máa wolẹ̀
máa rọra olówó tí ń f'owó ṣàánú
"(your) blackness is alluring, walk cautiously
(you) benevolent rich"

The tones and vowels of the text in speech and poetic modes were manually annotated in Praat (Boersma, 2001). For the three tones in the language (i.e., H, L, and M), F0(Hz) values of the pitch contour were extracted at 25%, 50%, and 75% intervals. To replicate the pitch contours as they appear in Praat windows for data visualization and analysis, each tone is labeled in a serial order, in this case T1.1, T1.2, T1.3...T2.1, etc, as shown in Figure 1. For each serially labeled tone, F0(Hz) values were extracted at twelve intervals. We extracted F0 values, intensity, formant and three spectral measurements (namely CPP, Energy below 500 hz and Hammarberg) from the annotations, using the Praat scripts written by Riebold (2013) and Xu (2013). Using the script tremor.praat (Brückl, 2021), we measured vibrato rate (rate of frequency tremor). The praat script only works on segments that are >3 s long, but the duration of all the vowels vocalized with vibrato is <1 s. To make each vibrato vocalization at least 3 s long, each of the vibrato vowels was sextupled by itself. It is from the sextupled form that we extracted vibrato rate. See Brückl (2021), Riebold (2013), and Xu (2013) for more details on the scripts.

In the next section, we discuss the motivation for each of the acoustic measurements that were utilized in this work. Our data and the R code of our statistical analysis are in the Supplementary material that is attached to this article.

4.2. Motivation for acoustic measurements in this work

Nine acoustic parameters were measured for the annotated vowels in order to detect the acoustic correlates of vocal expression. The parameters are fundamental frequency, intensity, cepstral peak prominence, energy below 500 Hz, Hammarberg index, formant 1, formant 2, duration and frequency tremor. The parameters were selected based on the description of Ìjálá and Ẹ̀sà in previous studies. In this section, each of these parameters are described.

4.2.1. Fundamental frequency and intensity

We start the discussion with fundamental frequency (F0) which primarily depends on the vibration rate of the vocal cords. F0(Hz) is measured in hertz (Hz). The perceptual correlate of F0 is pitch (Ladefoged and Johnson, 2015). Pitch contrasts that bring about lexical or grammatical meaning distinctions are called tone (Yip, 2002; Hyman, 2018). As shown in Section 2, Yorùbá has three tones. Given that the vocal expressions in Yorùbá oral poetry involve pitch raising, it is crucial to investigate how pitch contours of speech melodies are mapped to poetic melodies. For this reason, we extracted the F0(Hz) values of the tones in speech and poetic modes. To capture the full-time interval of the pitch contours, we extracted F0(Hz) values at twelve intervals for each tone. Recall that increased loudness is also a property of vocal expression in Yorùbá oral poetry. The perceptual correlate of intensity is loudness, but the relationship is not linear. Consequently, we also measured the intensity of all vowels in speech and poetic modes.

4.2.2. Cepstral peak prominence: CPP (dB)

Cepstral peak prominence (CPP) is the difference between the maximum cepstral peak value occurring within the boundaries of the expected phonational quefrencies and the corresponding value on the regression line fitted on the cepstrum (Hillenbrand et al., 1994; Hillenbrand and Houde, 1996). The degrees of glottal closure and vocal fold tension directly corresponds to the values of CPP(dB) (Kim, 1970; Iverson and Salmons, 1995; Inwald et al., 2011). Given that the glottal closure reduces the level of noise in vocal signal, CPP(dB) measures the level of noise in a vocal signal: the higher the noise, the lower the value of CPP and vice versa. In languages where the degree of glottal opening determines breathiness and aspiration, CPP(dB) is a reliable acoustic parameter for breathy and aspirated sounds (Hillenbrand et al., 1994; Blankenship, 2002; Esposito and Khan, 2012; Khan, 2012; Seyfarth and Garellek, 2018; Berkson, 2019). CPP(dB) was originally developed for measuring breathiness (Hillenbrand et al., 1994), but its usage has been extended to the evaluation of dysphonia. Studies show that the severity of dysphonia correlates with lower CPP values

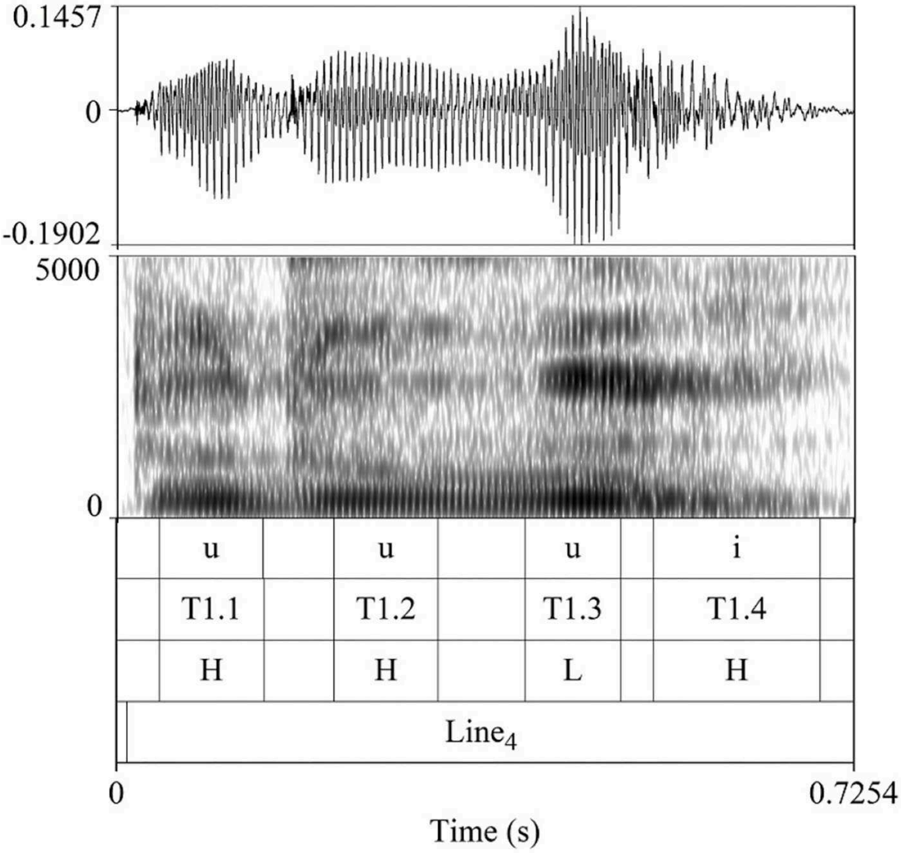

FIGURE 1
Annotations of vowels, tones, tone-sequences, and poetic lines.

when compared to normal voice (Hillenbrand and Houde, 1996; Heman-Ackah et al., 2002, 2003, 2014; Awan and Roy, 2005; Awan et al., 2009; Fraile and Godino-Llorente, 2014, etc.). As reported in Wolfe and Martin (1997), CPP(dB) values in hoarse and breathy voice are lower when compared to strained voice. As a result of the findings in various studies, the American Speech-Language-Hearing Association (ASHA) recommends CPP(dB) as a tool for "measuring the overall level of noise in the vocal signal" and as "a general measure of dysphonia" (Patel et al., 2018).

The use of CPP(dB) values has also been extended to the evaluation of effortful speech production and emotive speech. For example, increased loudness and sustained vowel production in effortful speech production correlate with higher CPP(dB) values when compared to normal speech production (Rosenthal et al., 2014; McKenna and Stepp, 2018; Phadke et al., 2020). Given that chanting involves increased loudness and high vocal demand, we could understand the vocal features of the poetic modes by measuring CPP(dB) values. Considering that nasality can decrease CPP(dB) values (see Madill et al., 2019), we only extracted CPP(dB) values for oral vowels.

4.2.3. Energy below 500 Hz

Another acoustic parameter which is used in this work is the proportion of spectral Energy below 500 Hz (dB). This measurement is often used for evaluating vocal quality in affective speech (Tolkmitt et al., 1982; Johnstone et al., 2005; Scherer et al., 2017). Low values of the Energy below 500 Hz are associated with the tensioning of vocal cords (Tolkmitt et al., 1982; Scherer et al., 2002, 2017; Johnstone et al., 2005). The values of Energy below 500 Hz (dB) were only extracted for oral vowels.

4.2.4. Hammarberg index (dB)

We extracted the values of Hammarberg index for evaluating vocal expression. The Hammarberg index is defined as the difference between the energy maximum in the 0–2,000 Hz frequency band and in the 2,000–5,000 Hz band (Hammarberg et al., 1980). Studies suggest that increase in loudness and F0(Hz) correlates with lower values of Hammarberg index (Scherer et al., 2017; Hakanpää et al., 2021; Sundberg et al., 2021). As an addition measurement for pitch raising and increased loudness, we measured Hammarberg index for all oral vowels in speech and poetic modes.

4.2.5. Formant frequencies

Formants are defined as "a resonating frequency of the air in the vocal tract" (Ladefoged and Johnson, 2015, p. 315). The first two formants, namely formant 1 (F1) and formant 2 (F2) are important in determining the quality of vowels. Specifically, F1(Hz) is mostly determined by vowel height and F2(Hz) is determined by vowel frontness or backness. The values of F1 increases in loud and effortful speech and verbal arts, but the values of F2 is not consistent under the same condition (Huber et al., 1999; Traunmüller and Eriksson, 2000; Huber and Chandrasekaran, 2007; Koenig and Fuchs, 2019, etc.). To understand the effect of vocal expression on the acoustics of vowels, F1(Hz) and F2(Hz) values were extracted for all the oral vowels.

4.2.6. Vibrato rate (frequency tremor)

We also measured the rate of frequency modulation or tremor. When frequency modulation occurs as a result of alcohol withdrawal syndrome (Koller et al., 1985; Anouti and Koller, 1995), aging (Gregory et al., 2012; Martins et al., 2015), or neurological disorders that cause involuntary movement of muscles in the throat, larynx (voice box), and vocal cords, it is called vocal tremor (Deuschl et al., 1998; Hlavnička et al., 2020). When used intentionally in singing, frequency tremor is called vibrato (Seashore, 1938; Dromey et al., 2003; Nix et al., 2016). The typical values of vibrato rate range from 4 to 7 Hz (Seashore, 1938; Dromey et al., 2003; Nix et al., 2016). In neurological diseases, vocal tremor frequencies are categorized into slow (<4 Hz), intermediate (4–7 Hz) or rapid (>7 Hz) (Deuschl et al., 1998; Charles et al., 1999). The slow tremor frequencies are prominent in all neurological disorders, but the intermediate and rapid tremor are mostly found in a subset of neurological disorders (Deuschl et al., 1998; Hlavnička et al., 2020).

4.3. Statistical analysis

The linear-mixed effect model was fitted to each acoustic parameters for each vowel and tone, to determine whether speech and poetic modes have a significant effect. In this case, the fixed effect is the modes (i.e., speech, Ìjálá and Ẹ̀sà), and the random effect is each iteration of the poems in all modes. For the tones, the random effect is the tone-bearing unit, in this case the vowels. This was done using the package "lme4" in R (Bates et al., 2014). We ran *post-hoc* pairwise comparisons for the mixed

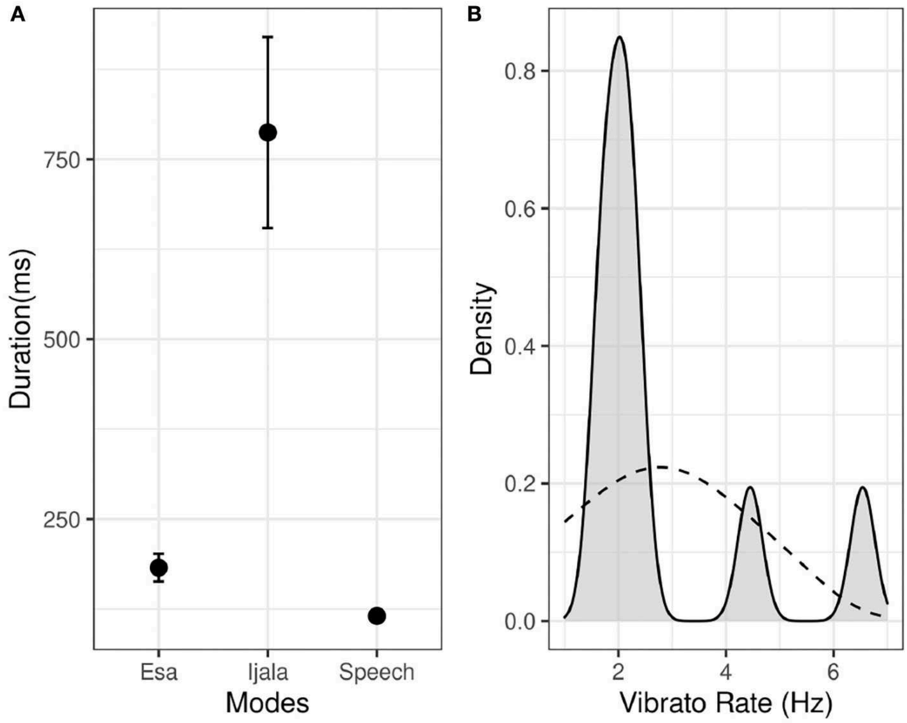

FIGURE 2
(A) Vibrato [a] in Ìjálá and the corresponding non-vibrato [a] in speech and Ẹ̀sà modes; **(B)** Vibrato rate(Hz) of the relevant vowels in Ìjálá mode.

effect model using the package "emmeans" (Lenth and Lenth, 2018).

To calculate the correspondence between pitch trajectories of speech and poetic modes, we used Pearson correlation coefficient *R* which measures the strength and direction of a linear relationship between two variables. The value of *R* is always between +1 and −1. The closer the value of *R* is to +1, the stronger the positive relationship between the two variables. However, the closer the value of *R* is to −1, the stronger the negative relationship between the two variables. If the value of *R* is 0, it means there is no relationship between the two variables (see Rumsey, 2009, for a basic description this statistical measurement). A regression line in a scatter plot describes the strength and direction of the linear relationship between the variables under consideration.

The null hypothesis is that there is no difference between speech and poetic modes for all the acoustic parameters. If the $p \leq 0.05$, there is a statistically significant effect of speech or poetic modes for the acoustic parameters. Therefore, we have a strong evidence against the null hypothesis. A $p > 0.05$ indicates weak evidence against null hypothesis (Rumsey, 2009; Wasserstein and Lazar, 2016). In the next section, we present the results of the acoustic analysis.

5. Results

We discuss the phonological attribute of vocal expression, before turning to the results of the acoustic analysis. In Ẹ̀sà, the triply long vowel [ooo] is inserted at the begining of the first word in line 1 even though the text does not have such vowel. If we recall that this is a recurrent attribute of Ẹ̀sà, we can say the long vowel is an attribute of vocal expression in Ẹ̀sà.

We now turn to the results of the acoustic analysis. One syllable in each of the first three lines were lengthened and

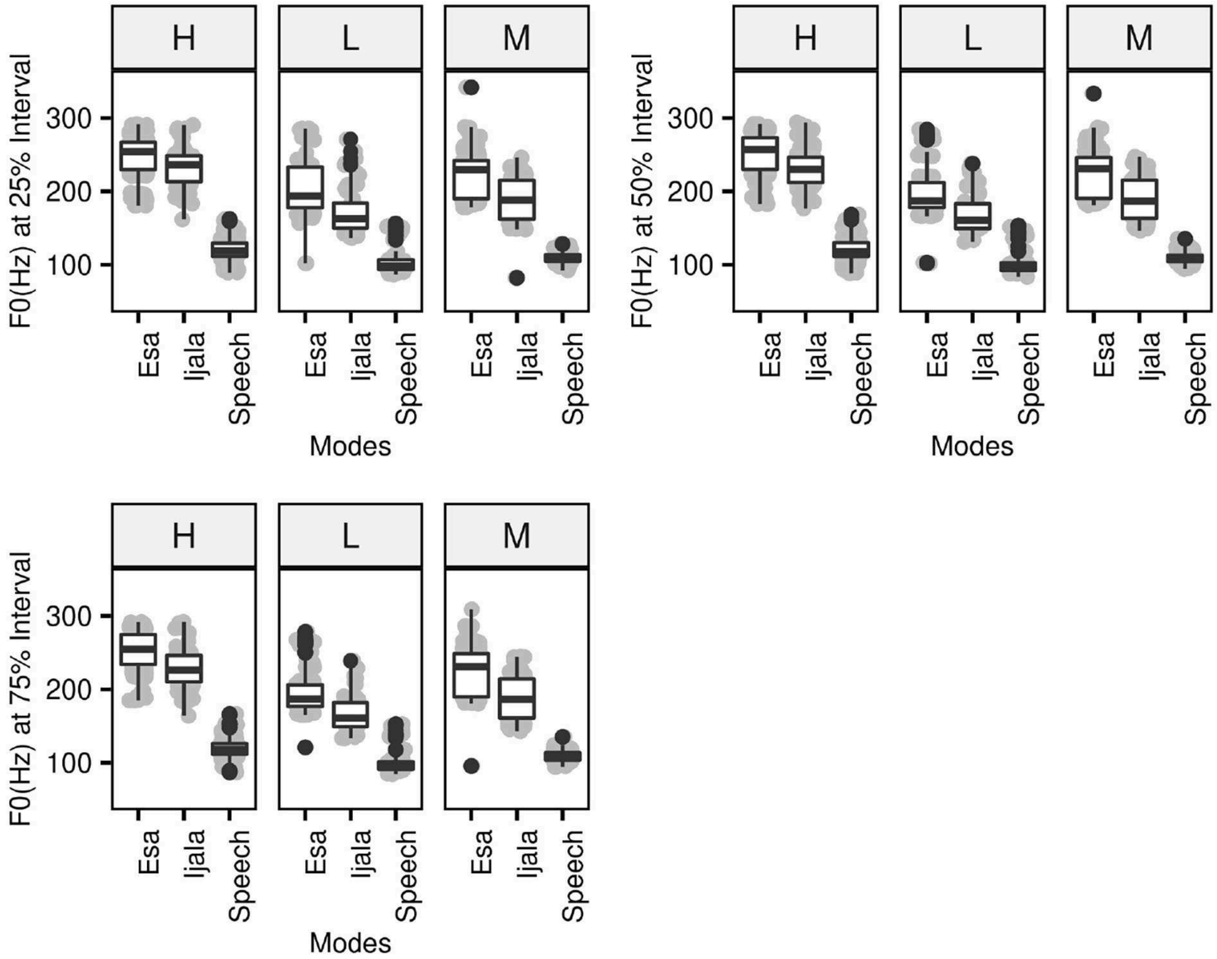

FIGURE 3
F0(Hz) values of H, L, and M tones in speech and poetic modes at 25%, 50%, and 75% intervals.

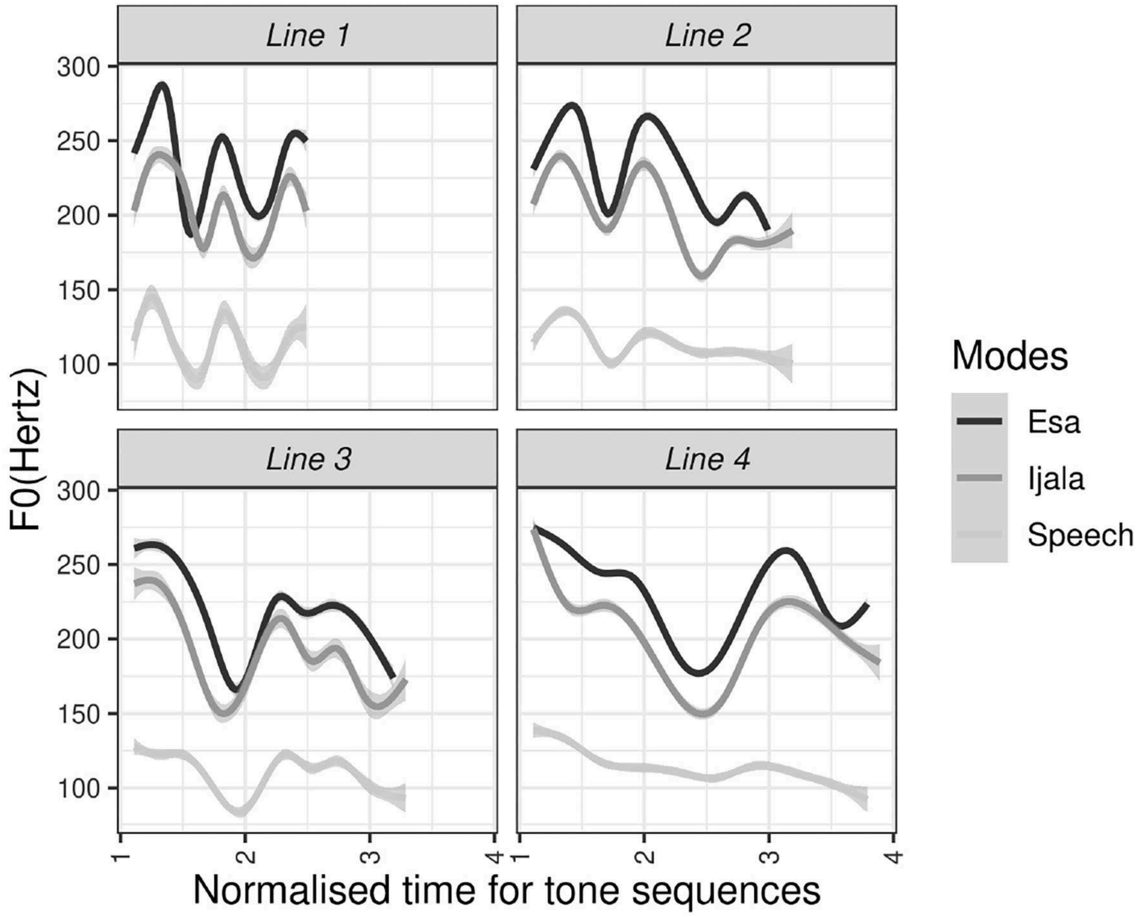

FIGURE 4
Pitch contours of each line in speech and poetic modes.

vocalized with vibrato in the Ìjálá mode, but at no point was Ẹ̀sà vocalized with vibrato. All the syllables targeted for vibrato (except for one) are in the range of the last and penultimate word in each poetic line. The syllable that were consistently targeted for vibrato in each iteration of the poem contains the sequence [ba], but the syllables with the vowel [ɪ, ɛ] were variably vocalized with vibrato. Considering that the other vowels were not consistently targeted for vibrato, we only measured the duration of the vowel that was consistently targeted for vibrato, as shown in Figure 2A. The vowel targeted for vibrato in the Ìjálá mode is significantly longer than the corresponding vowel in Ẹ̀sà ($p < 0.001$), which in turn is longer than the corresponding vowel in speech mode. However, the distinction between Ẹ̀sà and speech modes for the vibrato [a] is not statistically significant.

All the vowels that were produced with vibrato in Ìjálá have a vibrato rate in the range 1.6–2 Hz, as shown in Figure 2B. However, the vowels [ɛ] and [a] have the vibrato rate of 4.45 and 6.5 (Hz) respectively in one of their repetitions. Thus, the variation cannot be attributed to vowel-type. Considering that the vibrato rate of 4.45 (Hz) and 6.5 (Hz) are only found in two tokens, they are considered outliers.

Compared to Ìjálá, F0 values at 25%, 50%, and 75% intervals are higher in Ẹ̀sà for the three tones, as shown in Figure 3. Also at every interval, the values of F0(Hz) for each tone are higher in the poetic modes than the speech mode. The vocalization modes have a significant effect on the values of F0(Hz) for speech, Ìjálá and Ẹ̀sà comparisons (p<0.001). This shows that the vocal expression targets all the tones in the language. The results also show that the three tones in the language have distinct F0(Hz) values regardless of the vocalization mode.

The pitch trajectory of the tone sequences in each poetic line is compared in Figure 4. In the figure, the y-axis contains the acoustic measurement of pitch contour in F0(Hz), and the x-axis contains the proportional time of tone sequences. The dark line is for the pitch contours of Ẹ̀sà, the dark gray for Ìjálá and the light gray line is for the pitch contour of speech. There are four

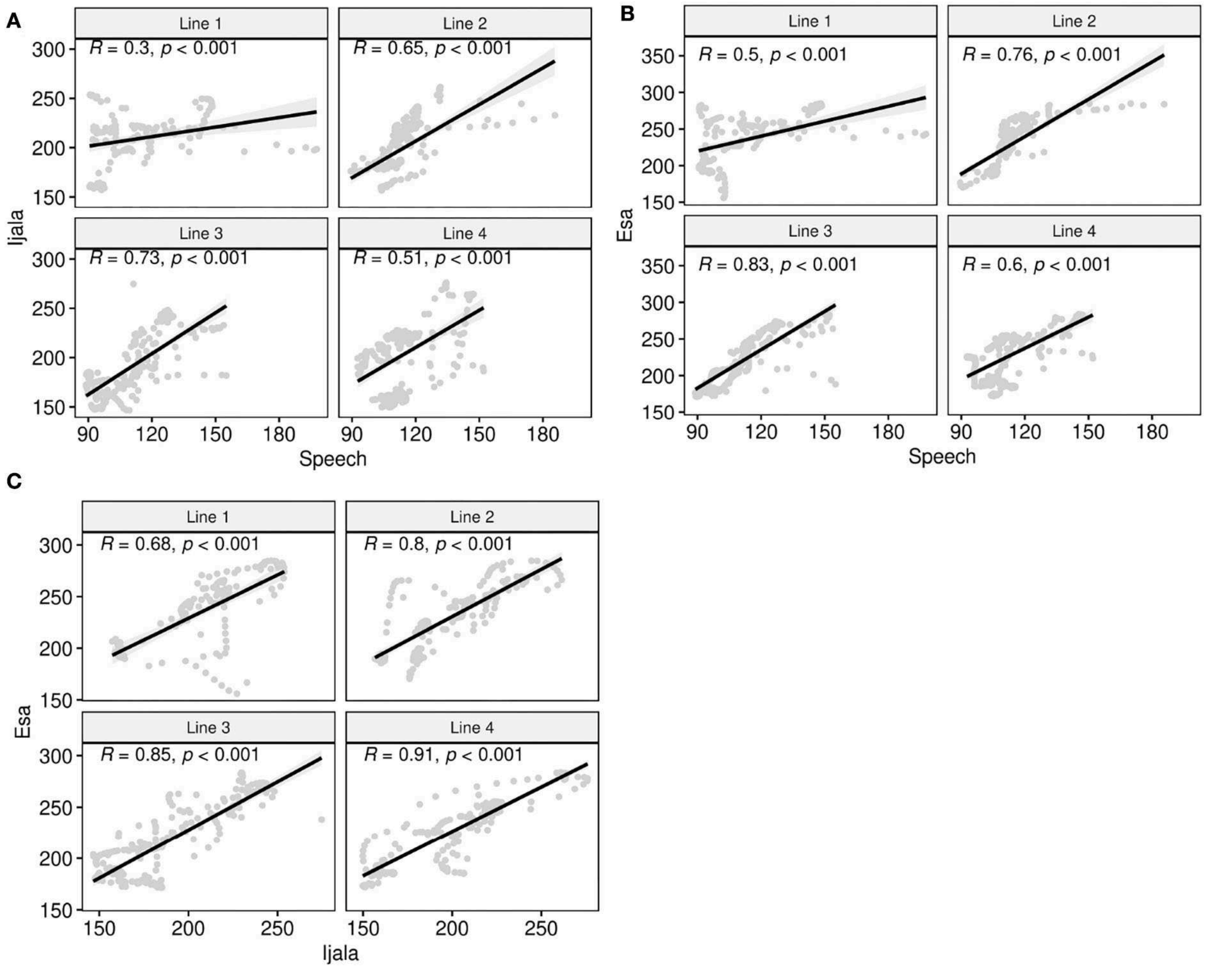

FIGURE 5
Correlation between the pitch trajectories of each line in speech and poetic modes: **(A)** Ìjálá vs. speech; **(B)** Ẹ̀sà vs. speech; **(C)** Ẹ̀sa vs. Ìjálá.

panels in the graph, where each panel is for the sequence of tones in each poetic line.

As shown in Figure 4, the values of F0(Hz) are higher in poetic modes when compared to the speech mode. The values of F0(Hz) are higher in Ẹ̀sà than Ìjálá. Figure 4 also shows that the pitch trajectory of each poetic line in speech mode is similar to that of the corresponding poetic line in Ẹ̀sà and Ìjálá.

To check the degree of similarity between the pitch contours of speech and poetic modes, we applied Pearson correlation coefficient to the pitch contour. To investigate whether linear relationship between speech and poetic modes varies based on poetic lines, the correlation coefficient are applied to each of the four poetic lines for every speech, Ìjálá and Ẹ̀sà comparisons. The results are shown in Figure 5.

Figure 5 shows that there are statistically significant positive correlations between the trajectories of speech and poetic melodies ($p < 0.001$), but the degree of correlation varies by poetic lines and genres. Comparing the lines in Figures 5A,B, we see that the correlation between the pitch contours of Ẹ̀sà and speech is higher when compared to the correlation between the pitch contours of Ìjálá and speech. Figure 5 also shows that the pitch contours of Ìjálá and Ẹ̀sà are closer than they are to the pitch contours of speech.

Similar to the F0(Hz) values, the intensity of oral vowels is significantly higher in poetic modes than in speech mode ($p < 0.001$), as shown in Figure 6. However, the distinction between the intensity Ẹ̀sà and Ìjálá varies depending on vowel-type. As shown in Figure 6, the intensity is higher in Ẹ̀sà than Ìjálá for all vowels, except for the vowel [u]. The difference between the intensity of Ẹ̀sà and Ìjálá is only significant ($p \leq 0.004$) for the vowels [o, ɔ, a]. The values of the Energy below 500 Hz(dB) are significantly higher in speech than poetic modes for all vowels ($p \leq 0.004$), except the vowels [e, o, ɔ]. The distinction between the Energy below 500 Hz(dB) of Ẹ̀sà and Ìjálá is only significant for [o, ɔ]. We now turn to the results of CPP(dB) and Hammarberg index(dB), which are presented Figure 7.

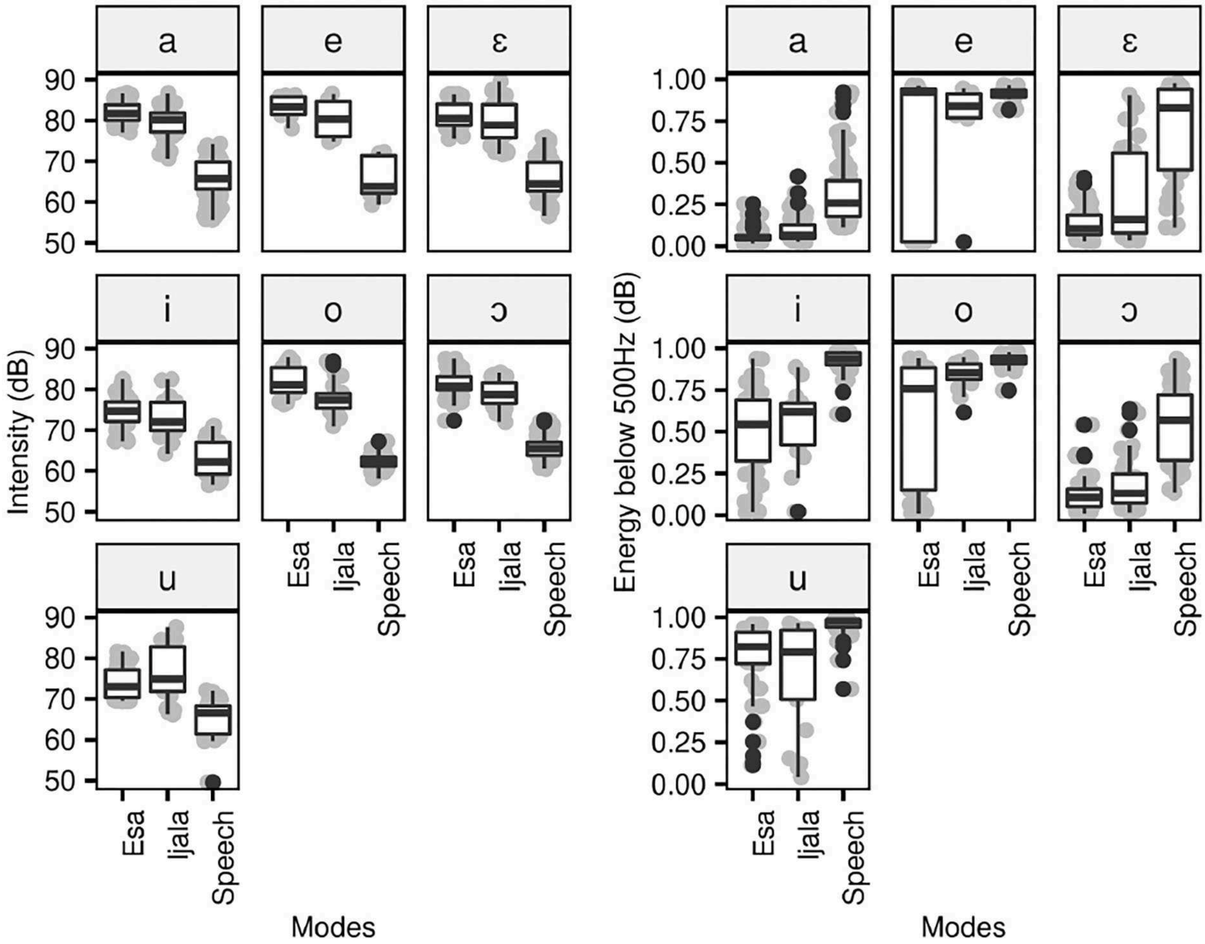

FIGURE 6
Intensity and energy below 500(Hz) of oral vowels in poetic and speech modes.

The results of the statistical analysis indicate that the mean values of CPP(dB) are significantly higher in poetic modes when compared to speech mode ($p < 0.001$). However, the difference between the CPP(dB) values of Ẹ̀sà and Ìjálá is not significant. The values of Hammarberg index are significantly lower in poetic modes than speech mode ($p < 0.001$). The distinction between the Hammarberg index(dB) values of Ẹ̀sà and Ìjálá is only statistically significant for the vowels [i, u, ɔ] ($p \leq 0.013$). The results of the formant values are presented in Figure 8.

There is an effect of vocal expression on vowel formants. As shown in Figure 8, the values of F1 for all oral vowels are higher in poetic modes than speech mode. For the values of F1(Hz), the distinction between poetic modes and speech mode is significant for all vowels ($p \leq 0.019$), except the vowel [u]. The graph in Figure 8 also shows that, for all vowel except [u], the values of F1 are slightly higher in Ẹ̀sà than Ìjálá. However, the F1 distinction between Ẹ̀sà than Ìjálá is only significant ($p \leq 0.026$) for the vowels [o, ɔ, a]. We now turn the values of F2(Hz). There is no obvious distinction between the values of F2(Hz) for speech, Ẹ̀sà and Ìjálá, except for the vowels [e, ɛ] that have lower F2(Hz) in poetic modes. Even in this case, the distinction is only significant for the vowel [ɛ].

In the next section, we discuss the results and their implications for the analysis of vocal expression in Ẹ̀sà and Ìjálá.

6. Discussion and conclusion

We set out in this work to understand the acoustic correlates of vocal expression in Ìjálá and Ẹ̀sà, under an experimental condition. The results of our investigation show that Hammarberg index(dB), Energy below 500 Hz(dB) and F1(Hz) distinguish the speech mode from each of the poetic modes for some vowels but are not as reliable as vibrato, F0(Hz), CPP(dB) and intensity(dB) which distinguish the poetic modes from each of the poetic modes for all vowels. For Ẹ̀sà vs. Ìjálá, Ẹ̀sà

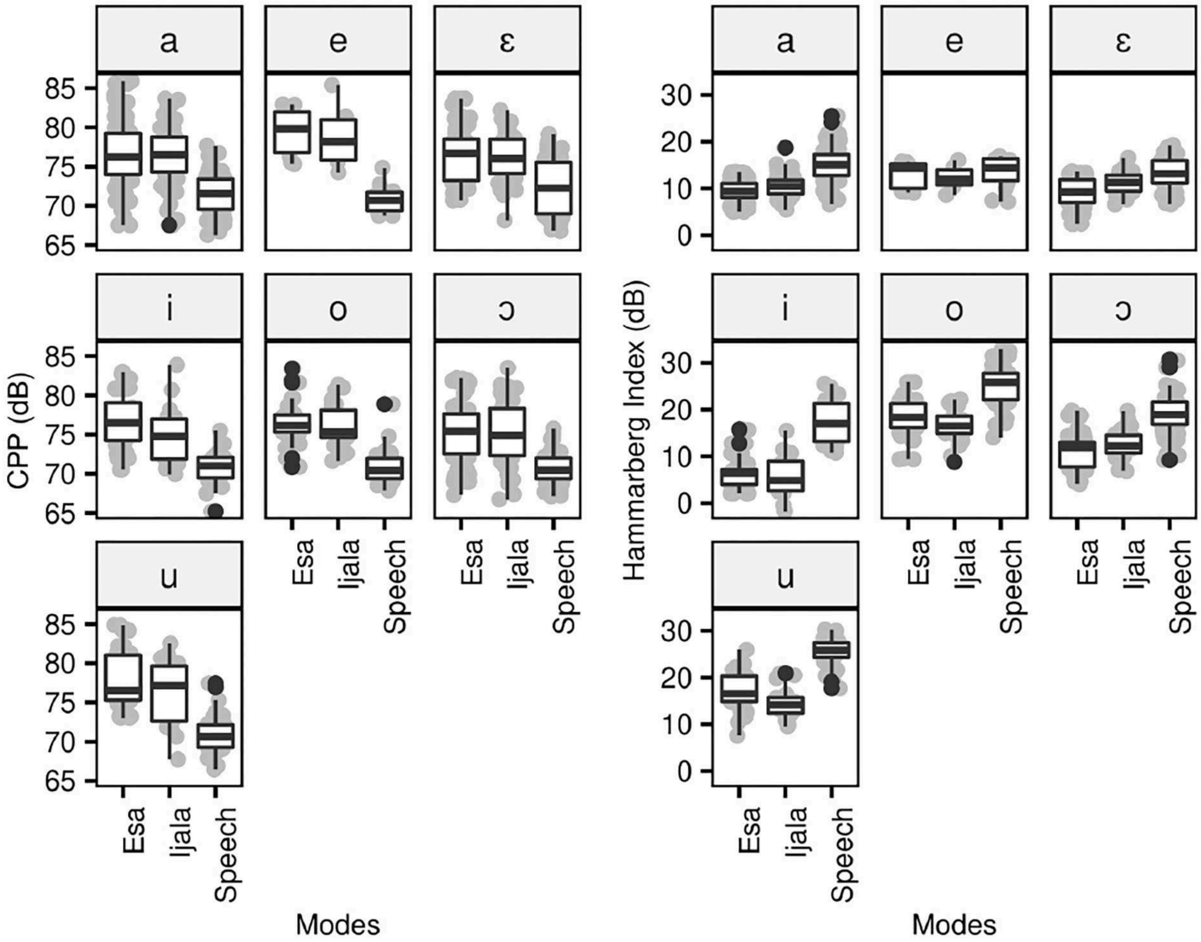

FIGURE 7
CPP(dB) and Hammarberg index(dB) of oral vowels in poetic and speech modes.

vs. speech and Ìjálá vs. speech comparisons, The most reliable acoustic parameters are vibrato and F0(Hz), given that vibrato only features in Ìjálá and that there is a significant effect of vocalization modes on F0(Hz), regardless of tone and poetic line. The results also show that there is a correspondence between the pitch trajectories of speech tones and poetic tunes, but the degree of correspondence varies by poetic lines and genres. Another distinctive feature that distinguishes Ẹ̀sà from Ìjálá is the epenthesis and lengthening of the vowel [o] in the first poetic line.

The acoustic correlates of vocal expressions in Ẹ̀sà and Ìjálá are consistent with increased vocal effort, given that higher values of F1(Hz), intensity and CPP are associated with increased vocal effort (e.g., Jessen et al., 2005; Rosenthal et al., 2014; McKenna and Stepp, 2018). The lower values of Energy below 500 Hz in poetic modes are also consistent with vocal tensing found in effortful speech. An increased vocal effort is expected as a feature of both genres, considering that vocal performance in a large space requires high vocal effort (Sundberg, 1977; Beechey et al., 2018) and that Ìjálá and Ẹ̀sà are typically performed to a large audience in an open space (Babalọla, 1963; Adedeji, 1978; Yai, 1989). It is probably vocal effort that previous research mischaracterised as stress.

The range of vibrato rate (1.6–2 Hz) in Ìjálá is atypical of the vibrato rate (4–7 Hz) in singing but consistent with vocal tremor as the historical origin of vocal expression in Ìjálá. Although the range of the vibrato rate reported in this work is prominent in all neurological diseases, it is difficult to tell whether the vibrato in Ìjálá historically developed from the vocal symptoms of alcohol withdrawal, aging or both. The pitch-height distinction between Ìjálá and Ẹ̀sà cannot be attributed to vibrato considering that the vibrato and non-vibrato sections of Ìjálá have lower pitch height than Ẹ̀sà, as shown in Figure 4.

Another notable finding of this study is tone-tune mapping. Studies on tone-tune mapping indicate that song melodies in a tone language are not determined by language, but music

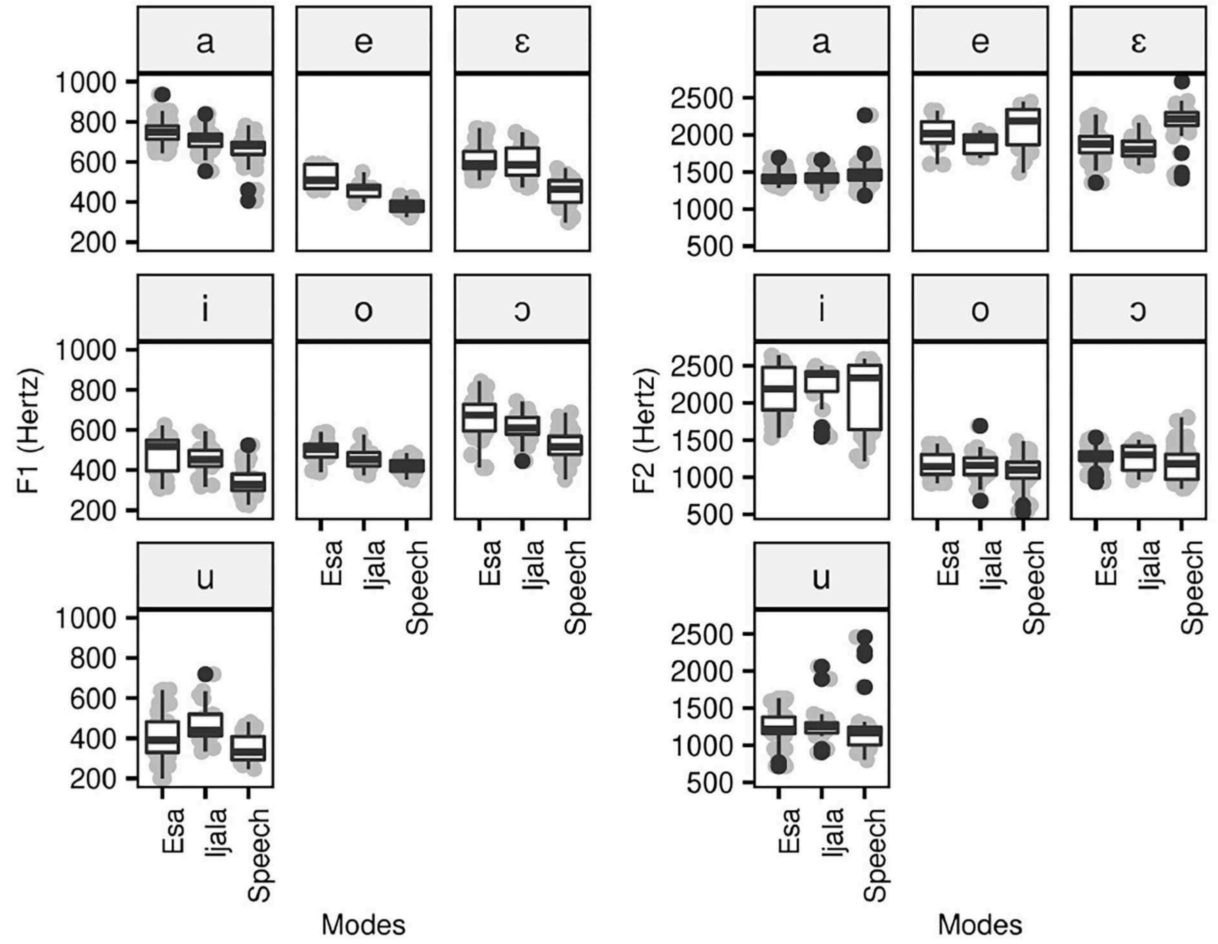

FIGURE 8
Formant plots of oral vowels in poetic and speech modes.

can accommodate language when it is musically feasible (Ho, 2006; Schellenberg, 2009, 2013; McPherson and Ryan, 2018). The results of our study is in line with the findings of studies on tone-tune mapping in singing, given that the correspondence relations between the pitch contours of speech tones and poetic tunes varies based on genres and poetic lines. As shown in the results of the correlation coeffieccient in Figure 5, the tune of Ẹ̀sà is closer to speech-tone melody than the tune of Ìjálá. This indicates that Ẹ̀sà is closer to speech than Ìjálá, in terms of ton-tune mapping. It remains to be seen whether this makes the chants of Ẹ̀sà more intelligible than Ìjálá.

Studies on affective use of vocal expression find that pitch raising and increased loudness are the most reliable cues for high level of arousal, such a excitement, fear and anger (Banse and Scherer, 1996; Juslin and Laukka, 2003; Johnstone et al., 2005; Goudbeek and Scherer, 2010; Lindquist et al., 2016; Scherer, 2021). It remains to be seen whether the pitch raising and increased loudness in Ìjálá and Ẹ̀sà are also associated with high level arousal such as excitement and happiness. Considering that Ìrèmọ̀jé is a dirge with similar vocal expressions as Ìjálá, future research involving more participants should compare the acoustic cues of vocal expression in Ìjálá and Ìrèmọ̀jé.

The major limitation of this work is that it is based on data from one participant. As a result of this, we cannot tell whether the acoustic correlates of vocal expression in this work are specific to the single participant or applies to other Yorùbá chanters. Thus, future research should replicate the present study on a larger population of Yorùbá chanters. Another limitation of this research is that we did not specifically look at the effect of vibrato on each lexical tone. To the best of our knowledge, the effect of vibrato on tone has not been studied, future research on singing or chanting in a tone language should investigate the interaction of tone and vibrato.

In sum, our study supports the observation in previous studies that vocal expressions, such as pitch raising, vowel epenthesis and lengthening, distinguish Ìjálá, Ẹ̀sà and speech. Contrary to the previous impressionistic observations, increased loudness as vocal expression does not distinguish Ìjálá from

Ẹ̀ṣà but the poetic modes from speech. In addition, we have shown that the vocal expression in Yorùbá oral poetry might be attributed to high vocal effort. Our analysis of tone-tune mapping in the poetic modes indicates that the poetic tunes correspond to the melody of speech tones, but the degree of correspondence varies based on poetic lines and genres. In addition to the analytical importance, the present study also supports vocal tremor as the historical origin of vocal expression in Ìjálá. It is important to note that the properties of vocal expression reported in this work were not possible to capture through older impressionistic observation methods. Given that properties of vocal expression in oral poetry are better captured in phonetic terms, we strongly recommend the instruments of phonetic science as valuable tools for the study of African verbal arts.

Author's note

This research was carried out when the first author was a student working under an SSHRC insight grant (435-2016-0369) awarded to Douglas Pulleyblank at the University of British Columbia. For comments and suggestions on various aspects of this work, we thank Douglas Pulleyblank, Rose-Marie Dechaine, Adélékè Adéèkó, Tolu Odebunmi, Akosua Addo, and the reviewers. Errors of fact or explanation are our own responsibility.

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

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

SKA designed the experiment, analyzed the data, write up the background, results, and discussion. OS collected the data, analyzed the data, and edited the manuscript. IBA set up the stimulus and write up the background. OA collected the data and edited the manuscript. All authors contributed to the article and approved the submitted version.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.1029400/full#supplementary-material

References

Adedeji, J. (1978). The poetry of the Yoruba masque theatre. *Afr. Arts* 11, 62–100. doi: 10.2307/3335415

Adeniran, T. (1974). The dynamics of peasant revolt: A conceptual analysis of the Agbekoya Parapo uprising in the western state of Nigeria. *J. Black Stud.* 4, 363–375. doi: 10.1177/002193477400400401

Ajibade, G. O. (2007). New wine in old cups: Postcolonial performance of Christian music in Yorùbá land. *Stud. World Christian.* 13, 105–126. doi: 10.3366/swc.2007.13.2.105

Ajuwon, B. (1977). *Funeral Dirges of Yoruba Hunters* (Ph.D. Thesis). Indiana University.

Akinbo, S. K. (2019). Representation of Yorùbá tones by a talking drum: an acoustic analysis. *Linguist. Lang. Afr.* 5, 11–23. doi: 10.4000/lla.347

Akinbo, S. K. (2021). The language of gángan, a Yorùbá talking drum. *Front. Commun.* 6, 650382. doi: 10.3389/fcomm.2021.650382

Akinlabi, A. (1985). *Tonal Underspecification and Yoruba Tone* (Ph.D. Thesis). University of Ibadan.

Anouti, A., and Koller, W. C. (1995). Tremor disorders, diagnosis and management. *Western J. Med.* 162, 510.

Awan, S. N., and Roy, N. (2005). Acoustic prediction of voice type in women with functional dysphonia. *J. Voice* 19, 268–282. doi: 10.1016/j.jvoice.2004.03.005

Awan, S. N., Roy, N., and Dromey, C. (2009). Estimating dysphonia severity in continuous speech: application of a multi-parameter spectral/cepstral model. *Clin. Linguist. Phonet.* 23, 825–841. doi: 10.3109/02699200903242988

Awóbùlúyì, Ọ. (1978). *Essentials of Yorùbá Grammar*. Ibadan: Oxford University Press Nigeria.

Babalọla, S. A. (1963). *The Content and Form of Yoruba Ijala* (Ph.D. Thesis). SOAS University of London.

Bamgboṣe, A. (1965). *Yoruba Orthography: A Linguistic Appraisal with Suggestions for Reform*. Ibadan: Ibadan University Press.

Bamgboṣe, A. (1966). *A Grammar of Yoruba, Vol. 5*. Cambridge: Cambridge University Press.

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

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *J. Stat. Softw*. 67, 1. doi: 10.18637/jss.v067.i01

Beechey, T., Buchholz, J. M., and Keidser, G. (2018). Measuring communication difficulty through effortful speech production during conversation. *Speech Commun*. 100, 18–29. doi: 10.1016/j.specom.2018.04.007

Berkson, K. H. (2019). Acoustic correlates of breathy sonorants in Marathi. *J. Phon*. 73, 70–90. doi: 10.1016/j.wocn.2018.12.006

Blankenship, B. (2002). The timing of nonmodal phonation in vowels. *J. Phon*. 30, 163–191. doi: 10.1006/jpho.2001.0155

Blench, R. (2019). Niger-congo: An alternative view. Manuscript. Available online at: http://www.rogerblench.info/Language/Niger-Congo/General/Niger-Congo%20an%20alternative%20view.pdf (accessed January 10, 2021).

Boadi, L. A. (1989). Praise poetry in Akan. *Research in African Literatures*, 20, 181–193.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot Int*. 5, 341–345.

Brown, D. J. B. (1995). *Orality, Textuality and History: Issues in South African Oral Poetry and Performance* (Ph.D. Thesis). Rutgers University.

Brückl, M. (2021). tremor3.05. praat-script for the computation of 18 measures of (vocal) tremor. doi: 10.13140/RG.2.2.13850.57287

Charles, P. D., Esper, G. J., Davis, T. J., Maciunas, R. J., and Robertson, D. (1999). Classification of tremor and update on treatment. *Am. Fam. Physician* 59, 1565.

Dada, A. O. (2014). Old wine in new bottle: elements of Yoruba culture in Aladura Christianity. *Black Theol*. 12, 19–32. doi: 10.1179/1476994813Z.00000000017

Deuschl, G., Bain, P., Brin, M., and Committee, A. H. S. (1998). Consensus statement of the movement disorder society on tremor. *Mov. Disord*. 13, 2–23. doi: 10.1002/mds.870131303

Dromey, C., Carter, N., and Hopkin, A. (2003). Vibrato rate adjustment. *J. Voice* 17, 168–178. doi: 10.1016/S0892-1997(03)00039-0

Esposito, C. M., and Khan, S. U. D. (2012). Contrastive breathiness across consonants and vowels: a comparative study of Gujarati and White Hmong. *J. Int. Phon. Assoc*. 42, 123–143. doi: 10.1017/S0025100312000047

Euba, A. (1990). *Yorùbá Drumming: The Dùndún Tradition*. Bayreuth: Bayreuth African Studies.

Fámúle, O. (2018). *Èdè* àyàn: The language of Àyàn in Yorùbá art and ritual of Egúngún. *Yoruba Studies Review*, 2, 1–50.

Finnegan, R. (2007). *The Oral and Beyond: Doing Things with Words in Africa*. London: James Currey; University of Chicago Press.

Finnegan, R. (2012). *Oral Literature in Africa*. Cambridge, UK: Open Book Publishers.

Fraile, R., and Godino-Llorente, J. I. (2014). Cepstral peak prominence: acomprehensive analysis. *Biomed. Signal Process. Control* 14, 42–54. doi: 10.1016/j.bspc.2014.07.001

Gbadamosi, B., and Beier, U. (1959). *Yoruba Poetry*. Ibadan: Government Printers.

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

Gregory, N. D., Chandran, S., Lurie, D., and Sataloff, R. T. (2012). Voice disorders in the elderly. *J. Voice* 26, 254–258. doi: 10.1016/j.jvoice.2010.10.024

Hakanpää, T., Waaramaa, T., and Laukkanen, A.-M. (2021). Training the vocal expression of emotions in singing: effects of including acoustic research-based elements in the regular singing training of acting students. *J. Voice*. doi: 10.1016/j.jvoice.2020.12.032

Halle, M. and Idsardi, W. (1995). "General properties of stress and metrical structure," in *The handbook of phonological theory,* ed J. Goldsmith, (Cambridge, MA: Blackwell), 403–443

Hammarberg, B., Fritzell, B., Gaufin, J., Sundberg, J., and Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngol*. 90, 441–451. doi: 10.3109/00016488009131746

Heman-Ackah, Y. D., Michael, D. D., Baroody, M. M., Ostrowski, R., Hillenbrand, J., Heuer, R. J., et al. (2003). Cepstral peak prominence:a more reliable measure of dysphonia. *Ann. Otol. Rhinol. Laryngol*. 112, 324–333. doi: 10.1177/000348940311200406

Heman-Ackah, Y. D., Michael, D. D., and Goding Jr, G. S. (2002). The relationship between cepstral peak prominence and selected parameters of dysphonia. *J. Voice* 16, 20–27. doi: 10.1016/S0892-1997(02)00067-X

Heman-Ackah, Y. D., Sataloff, R. T., Laureyns, G., Lurie, D., Michael, D. D., Heuer, R., et al. (2014). Quantifying the cepstral peak prominence, a measure of dysphonia. *J. Voice* 28, 783–788. doi: 10.1016/j.jvoice.2014.05.005

Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *J. Speech Lang. Hear. Res*. 37, 769–778. doi: 10.1044/jshr.3704.769

Hillenbrand, J., and Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. *J. Speech Lang. Hear. Res*. 39, 311–321. doi: 10.1044/jshr.3902.311

Hlavnička, J., Tykalová, T., Ulmanová, O., Dušek, P., Horáková, D., Růžička, E., et al. (2020). Characterizing vocal tremor in progressive neurological diseases via automated acoustic analyses. *Clin. Neurophysiol*. 131, 1155–1165. doi: 10.1016/j.clinph.2020.02.005

Ho, W. S. V. (2006). "The tone-melody interface of popular songs written in tone languages," in *Proceedings of the 9th International Conference on Music Perception and Cognition*, eds M. Baroni, A. R. Addessi, R. Caterina, and M. Costa (Bologna: Bononia University Press), 1414–1422.

Huber, J. E., and Chandrasekaran, B. (2007). Effects of increasing sound pressure level on lip and jaw movement parameters and consistency in young adults. *J. Speech Lang. Hear. Res*. 2, 1368–1379. doi: 10.1044/1092-4388(2006/098)

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T. A., and Johnson, K. (1999). Formants of children, women, and men: the effects of vocal intensity variation. *J. Acoust. Soc. Am*. 106, 1532–1542. doi: 10.1121/1.427150

Hyman, L. M. (2018). What tone teaches us about language. *Language* 94, 698–709. doi: 10.1353/lan.2018.0040

Idamoyibo, A. A. (2006). *Yoruba Traditional Music in Christian Worship: A Case Study of Ijala Musical Genre* (Ph.D. Thesis). University of Ibadan.

Idamoyibo, A. A. (2011). Esa music and the struggle for relevance in the 21st century. *Pakistan J. Soc. Sci*. 8, 234–239. doi: 10.3923/pjssci.2011.234.239

Inwald, E. C., Döllinger, M., Schuster, M., Eysholdt, U., and Bohr, C. (2011). Multiparametric analysis of vocal fold vibrations in healthy and disordered voices in high-speed imaging. *J. Voice* 25, 576–590. doi: 10.1016/j.jvoice.2010.04.004

Iverson, G. K., and Salmons, J. C. (1995). Aspiration and laryngeal representation in Germanic. *Phonology* 12, 369–396. doi: 10.1017/S0952675700002566

Jessen, M., Koster, O., and Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *Int. J. Speech Lang. Law* 12, 174–213. doi: 10.1558/sll.2005.12.2.174

Johnstone, T., van Reekum, C. M., Hird, K., Kirsner, K., and Scherer, K. R. (2005). Affective speech elicited with a computer game. *Emotion* 5, 513. doi: 10.1037/1528-3542.5.4.513

Juslin, P. N., and Laukka, P. (2003). Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol. Bull*. 129, 770. doi: 10.1037/0033-2909.129.5.770

Kager, R. (2007). "Feet and metrical stress," in *The Cambridge Handbook of Phonology*, ed P. de Lacy (Cambridge: Cambridge University Press), 195–227.

Kamiloğlu, R. G., Fischer, A. H., and Sauter, D. A. (2020). Good vibrations: a review of vocal expressions of positive emotions. *Psychon. Bull*. 27, 237–265. doi: 10.3758/s13423-019-01701-x

Kenstowicz, M. (2006). "Tone loans: the adaptation of English loanwords into Yoruba," in *Selected Proceedings of the 35th Annual Conference on African Linguistics*, eds J. M. Mugane, J. P. Hutchison., and D. A. Worman (Somerville, MA: Cascadilla Proceedings Project), 136–146.

Khan, S., and u. D. (2012). The phonetics of contrastive phonation in Gujarati. *J. Phon*. 40, 780–795. doi: 10.1016/j.wocn.2012.07.001

Kim, C.-W. (1970). A theory of aspiration. *Phonetica* 21, 107–116. doi: 10.1159/000259293

Koenig, L. L., and Fuchs, S. (2019). Vowel formants in normal and loud speech. *J. Speech Lang.Hear. Res*. 62, 1278–1295. doi: 10.1044/2018_JSLHR-S-18-0043

Koller, W., O'Hara, R., Dorus, W., and Bauer, J. (1985). Tremor in chronic alcoholism. *Neurology* 35, 1660–1660. doi: 10.1212/WNL.35.11.1660

Ladefoged, P., and Johnson, K. (2015). *A Course in Phonetics*. Wadsworth, OH: Cengage Learning.

Lenth, R., and Lenth, M. R. (2018). Package 'lsmeans'. *Am Stat*. 34, 216–221.

Liberman, M., and Prince, A. (1977). On stress and linguistic rhythm. *Linguist. Inquiry* 8, 249–336.

Lindquist, K. A., Gendron, M., Satpute, A. B., and Lindquist, K. (2016). "Language and emotion," in *Handbook of Emotions*, eds M. Lewis, J. M. Haviland-Jones, and L. F. Barrett, *4th Edn* (New York, NY: The Guilford Press), 579–594.

Madill, C., Nguyen, D. D., Yick-Ning Cham, K., Novakovic, D., and McCabe, P. (2019). The impact of nasalance on cepstral peak prominence and harmonics-to-noise ratio. *Laryngoscope* 129, E299-E304. doi: 10.1002/lary.27685

Martins, R. H. G., Benito Pessin, A. B., Nassib, D. J., Branco, A., Rodrigues, S. A., and Matheus, S. M. M. (2015). Aging voice and the laryngeal muscle atrophy. *Laryngoscope* 125, 2518–2521. doi: 10.1002/lary.25398

McKenna, V. S., and Stepp, C. E. (2018). The relationship between acoustical and perceptual measures of vocal effort. *J. Acoust. Soc. Am*. 144, 1643–1658. doi: 10.1121/1.5055234

McPherson, L., and Ryan, K. M. (2018). Tone-tune association in Tommo So (Dogon) folk songs. *Language* 94, 119–156. doi: 10.1353/lan.2018.0003

Nix, J., Perna, N., James, K., and Allen, S. (2016). Vibrato rate and extent in college music majors: a multicenter study. *J. Voice* 30, 756-e31. doi: 10.1016/j.jvoice.2015.09.006

Ògúndejì, P. A. (1991). *Introduction to Yoruba Oral Literature*. Ibadan: Ibadan Center for External Studies; University of Ibadan.

Okpewho, I. (1992). *African Oral Literature: Backgrounds, Character, and Continuity, Vol. 710*. Bloomington, IN: Indiana University Press.

Ọlabimtan, A. (1977). Rhythm in Yoruba poetry: the example of orin-arungbe. *Res. Afr. Literat*. 8, 201–218.

Olajubu, C. O., and Ojo, J. R. O. (1977). Some aspects of Ọ̀yó Yorùbá masquerades. *Africa* 47, 253–275. doi: 10.2307/1158862

Olajubu, O. (1974). Iwì Egúngún chants: an introduction. *Research in African Literatures*, 31–51.

Olajubu, O. (1984). Funeral dirges of Yoruba hunters. *Res. Afr. Literat*. 14, 592–596.

Olaniyan, S. O. (2013). An ecocritical reading of Ijala chant: an example of Ogundare Foyanmu's selected Ijala chant. *J. Literat. Art Stud*. 3, 692–701. doi: 10.17265/2159-5836/2013.11.004

Ọlátúnjí, D. O. (1979). The Yoruba oral poet and his society. *Res. Afr. Literat*. 10, 179–207.

Olátúnjí, M. O. (2012). Modern trends in the Islamized music of the traditional Yorùbá concept, origin, and development. *Matatu* 40, 447–455. doi: 10.1163/18757421-040001029

Patel, R. R., Awan, S. N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., et al. (2018). Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function. *Am. J. Speech Lang. Pathol*. 27, 887–905. doi: 10.1044/2018_AJSLP-17-0009

Phadke, K. V., Laukkanen, A.-M., Ilomäki, I., Kankare, E., Geneid, A., and Švec, J. G. (2020). Cepstral and perceptual investigations in female teachers with functionally healthy voice. *J. Voice* 34, 485-e33. doi: 10.1016/j.jvoice.2018.09.010

Pulleyblank, D. (1986). *Tone in Lexical Phonology*. Dordrecht: D. Reidel Publishing Company.

Pulleyblank, D. (1988). Vocalic underspecification in Yoruba. *Linguist. Inquiry* 19, 233–270.

Pulleyblank, D. (2009). "Yoruba," in *The World's Major Languages*, ed B. Comrie (Oxford: Taylor & Francis), 866–882.

Purvis, T. M. (2009). Speech rhythm in Akan oral praise poetry. *Text Talk* 29, 201–218. doi: 10.1515/TEXT.2009.009

Riebold, J. (2013). *Vowel Analyzer*. Available online at: https://github.com/jmriebold

Rosenthal, A. L., Lowell, S. Y., and Colton, R. H. (2014). Aerodynamic and acoustic features of vocal effort. *J. Voice* 28, 144–153. doi: 10.1016/j.jvoice.2013.09.007

Rumsey, D. J. (2009). *Statistics II for Dummies*. Hoboken, NJ: Wiley Publishing Inc.

Schellenberg, M. (2009). "Singing in a tone language: Shona," in *Selected Proceedings of the 39th Annual Conference on African Linguistics*, eds A. Ojo and L. Moshi (Somerville, MA: Cascadilla Proceedings Project), 137–144.

Schellenberg, M. H. (2013). *The Realization of Tone in Singing in Cantonese and Mandarin* (Ph.D. Thesis). University of British Columbia.

Scherer, K. R. (1985). "Vocal affect signaling: a comparative approach," in *Advances in the Study of Behavior, Vol. 15*, eds K. Scherer, J. Rosenblatt, C. Beer, M. Busnel, and P. Slater (New York, NY: Academic Press), 189–244.

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

Scherer, K. R. (2021). Comment: advances in studying the vocal expression of emotion: current contributions and further options. *Emot. Rev*. 13, 57–59. doi: 10.1177/1754073920949671

Scherer, K. R., Grandjean, D., Johnstone, T., Klasmeyer, G., and Bänziger, T. (2002). "Acoustic correlates of task load and stress," in *7th International Conference on Spoken Language Processing (Interspeech 2002)*, eds J. H. L. Hansen and B. Pellom (Adelaide, SA: Causal Productions), 2017–2020.

Scherer, K. R., Sundberg, J., Fantini, B., Trznadel, S., and Eyben, F. (2017). The expression of emotion in the singing voice: acoustic patterns in vocal performance. *J. Acoust. Soc. Am*. 142, 1805–1815. doi: 10.1121/1.5002886

Seashore, C. E. (1938). *The Psychology of Music*. New York. NY: McGraw-Hill.

Seyfarth, S., and Garellek, M. (2018). Plosive voicing acoustics and voice quality in Yerevan Armenian. *J. Phon*. 71, 425–450. doi: 10.1016/j.wocn.2018.09.001

Siertsema, B. (1959). Stress and tone in Yoruba word composition. *Lingua* 8, 385–402. doi: 10.1016/0024-3841(59)90037-3

Sundberg, J. (1977). The acoustics of the singing voice. *Sci. Am*. 236, 82–91. doi: 10.1038/scientificamerican0377-82

Sundberg, J., Salomão, G. L., and Scherer, K. R. (2021). Analyzing emotion expression in singing via flow glottograms, long-term-average spectra, and expert listener evaluation. *J. Voice* 35, 52–60. doi: 10.1016/j.jvoice.2019.08.007

Tolkmitt, F., Helfrich, H., Standke, R., and Scherer, K. R. (1982). Vocal indicators of psychiatric treatment effects in depressives and schizophrenics. *J. Commun. Disord*. 15, 209–222. doi: 10.1016/0021-9924(82)90034-X

Traunmüller, H., and Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *J. Acoust. Soc. Am*. 107, 3438–3451. doi: 10.1121/1.429414

Uzochukwu, S. (1981). *Traditional Elegiac Poetry of the Igbo: A Study of the Major Types* (Ph.D. Thesis). University of Lagos.

Villepastour, A. (2010). *Ancient Text Messages of the Yorùbá Bàtá Drum: Cracking the Code*. Surrey: Ashgate Publishing Limited.

Wasserstein, R. L., and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *Am. Statist*. 70, 129–133. doi: 10.1080/00031305.2016.1154108

Wolfe, V., and Martin, D. (1997). Acoustic correlates of dysphonia: type and severity. *J. Commun. Disord*. 30, 403–416. doi: 10.1016/S0021-9924(96)00112-8

Xu, Y. (2013). "Prosodypro–A Tool for Large-Scale Systematic Prosody Analysis", in *Tools and Resources for the Analysis of Speech Prosody* eds B. Bigi and D. Hirst (Aix-en-Provence: Laboratoire Parole et Langage), 7–10.

Yai, O. (1989). "Issues in oral poetry: criticism, teaching, and translation," in *Georgetown University Round Table on Languages and Linguistics 1986*, eds Battestini and P. X. Simon (Washington, DC: Georgetown University Pres), 91–106.

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

frontiers | Frontiers in Communication

TYPE Correction
PUBLISHED 06 January 2023
DOI 10.3389/fcomm.2022.1127007

Corrigendum: An acoustic study of vocal expression in two genres of Yoruba oral poetry

Samuel K. Akinbo[1]*, Olanrewaju Samuel[2], Iyabode B. Alaga[2] and Olawale Akingbade[2]

[1] Department of Linguistics, University of British Columbia, Vancouver, BC, Canada, [2] Department of Linguistics and African Languages, University of Ibadan, Ibadan, Nigeria

KEYWORDS

oral poetry, tone, vibrato, vocal effort, vocal expression, phonetics

OPEN ACCESS

EDITED AND REVIEWED BY
Sandra Madureira,
PUCSP, Brazil

*CORRESPONDENCE
Samuel K. Akinbo
✉ akinbosk@gmail.com

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

RECEIVED 19 December 2022
ACCEPTED 28 December 2022
PUBLISHED 06 January 2023

CITATION
Akinbo SK, Samuel O, Alaga IB and Akingbade O (2023) Corrigendum: An acoustic study of vocal expression in two genres of Yoruba oral poetry. *Front. Commun.* 7:1127007. doi: 10.3389/fcomm.2022.1127007

A corrigendum on
An acoustic study of vocal expression in two genres of Yoruba oral poetry

by Akinbo, S. K., Samuel, O., Alaga, I. B., and Akingbade, O. (2022). Front. Commun. 7:1029400. doi: 10.3389/fcomm.2022.1029400

In the published article, there were a few text errors. These have been outlined below.

A correction has been made to the **Abstract**. This sentence previously stated:

"For this study, we conducted an experiment, involving the vocalization of an original poem in speech mode, Ìjálá and Èsá."

The corrected sentence appears below:

"For this study, we conducted an experiment, involving the vocalization of an original poem in speech mode, Ìjálá and Ẹ̀sà."

A correction has been made to **Methodology**, *Stimuli, participant and procedure*, paragraph 4. The sentences previously stated:

"To make each vibrato vocalization at least 3 s long, each of the vibrato vowel were tripled by itself. It is from the tripled form that we extracted vibrato rate."

The corrected sentences appear below:

"To make each vibrato vocalization at least 3 s long, each of the vibrato vowels was sextupled by itself. It is from the sextupled form that we extracted vibrato rate."

A correction has been made to **Discussion and conclusion**, paragraph 2. The sentence previously stated:

"…that Ìjálá and Ẹ̀sa are typically performed to a large audience in an open space…"

The corrected sentence appears below:

"…that Ìjálá and Ẹ̀sà are typically performed to a large audience in an open space…"

A correction has been made to **Discussion and conclusion**, paragraph 7. The sentence previously stated:

"…vocal expression such as pitch raising distinguishes genres of Yorùbá oral poetry from speech."

The corrected sentence appears below:

"…vocal expressions, such as pitch raising, vowel epenthesis and lengthening, distinguish Ìjálá, Ẹ̀sà and speech."

The authors apologize for these errors and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

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

TYPE Original Research
PUBLISHED 09 January 2023
DOI 10.3389/fcomm.2022.1088170

An acoustic study on character voices of dominators and subordinates: A case study on male characters in *Empresses in the Palace*

Wen Liu[†], Xinyi Zhang[†] and Changwei Liang*

Center for Language Sciences, School of Literature, Shandong University, Jinan, China

OPEN ACCESS

EDITED BY
Plinio Almeida Barbosa,
State University of Campinas, Brazil

REVIEWED BY
Sandra Madureira,
PUCSP, Brazil
Philippe Boula De Mareüil,
Université Paris-Saclay, France

*CORRESPONDENCE
Changwei Liang
✉ liangchangwei@sdu.edu.cn

[†]These authors have contributed equally to this work and share first authorship

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

RECEIVED 03 November 2022
ACCEPTED 07 December 2022
PUBLISHED 09 January 2023

CITATION
Liu W, Zhang X and Liang C (2023) An acoustic study on character voices of dominators and subordinates: A case study on male characters in *Empresses in the Palace*.
Front. Commun. 7:1088170.
doi: 10.3389/fcomm.2022.1088170

Introduction: Voice has been used to project identity in dubbing, in order to auditory portray appropriate role images in TV dramas. This study investigates the character voices of leading male characters in *Empresses in the Palace.*

Methods: Different acoustic characteristics of character voices and matching relation between acoustics and role images are explored by comparing F0, CPP, harmonic amplitude differences of speech spectrum.

Results: The voice quality of characters is related to their relative social status. The subordinates usually adopt a higher pitch or breathy voice, while the dominators use a lower pitch or modal/creaky voice. In addition, CPP, F0, and H1-A3 are the key acoustic indicators to distinguish character voices.

Discussion: These results reveal the acoustic characteristics of character voices of certain types, as well as provide guidance for dubbing vividly.

KEYWORDS

character voice, male character, social status, voice quality, acoustic analysis

1. Introduction

1.1. Voice, voice quality, and projection of identity

The terms "voice" and "voice quality" have not been defined with a broad agreement, since the researches in these fields are transdisciplinary. These terms are usually defined in both a broad sense and a narrow sense. Narrowly, "voice" refers to how the vocal folds vibrate, namely, vibratory patterns of the vocal folds; and "voice quality" represents the voicing produced at the glottis, which is also termed as "phonation type" or "phonation quality" (Esling et al., 2019, p. 2–8). In a broad sense, "voice" is essentially synonymous with "speech", while "voice quality" refers to the auditory characteristics of the speaker's voice (Abercrombie, 1967, p. 91; Esling et al., 2019, p. 123; Liu, 2021). The relationship between voice and voice quality can be simply explained as follows: "voice" has a physical and physiological base that refers to the acoustic signal, while "voice quality" refers to the perceptual impression that occurs as a result of that signal (Kreiman and Sidtis, 2011, p. 5). American National Standards Institute also defines "voice quality" as the attribute of auditory sensation (ANSI et al., 1960, p. 45). In the following part, "voice" refers to sound produced by the vocal folds. In contrast, "voice quality" indicates the perceptual ramifications caused by different vocal fold configurations.

Voice quality is one of the primary means by which speakers project their identity, i.e., their physical, psychological, and social characteristics (Laver, 1980, p. 2; Kreiman and Sidtis, 2011, p. 1). Information such as gender, age, and mood can be easily perceived by the listener, even without seeing the speaker. Voice allows us to recognize individuals and emotional states, also called "auditory face" (Belin et al., 2004). The listener can form an impression of the speaker based on the voice. It specifically includes body type, lifestyle, mental state (Kreiman et al., 2005), and even morality (Teshigawara, 2003; Podesva and Callier, 2015). Although voice identification is not always a certainty (Bonastre et al., 2003), we can still recognize a person instantly through voice individuality (Dolar, 2006, p. 22), especially for the recognition of familiar voices (Van Lancker et al., 1985; Eriksson, 2005; Kreiman and Sidtis, 2011; Hansen and Hasan, 2015). For example, even the greeting "hello" can convey information about the speaker his/herself to the listener, allowing the listener to form a judgement of the speaker's personality; based on this, a two-dimensional model on personality judgement was constructed (Wu et al., 2021).

As an essential factor influencing a speaker's identity perception, behavioral research has found that pitch is related to the listener's perception of dominance. Morton (1977) found that birds and mammals use harsh and relatively low-frequency sounds when hostile, while higher-frequency, more pure tone-like sounds are adopted when frightened, appeasing, or approaching in a friendly manner. The association between particular images and certain sounds across many languages is also known as sound symbolism (e.g., Sapir, 1929; Hinton et al., 1994; Fónagy, 2001; among many others). Ohala (1984, 1994) proposed the frequency code hypothesis to represent this sound pattern, indicating that these sound symbolic patterns have phonetic bases. To be specific, for both humans and animals, compared to a voice with higher fundamental frequency (F0), a voice with lower F0 is commonly perceived as having a larger vocalizer and larger body, and are thus perceived more dominant, aggressive, and threatening. Cao and Kong (2016) demonstrated that the length and volume of the human pharyngeal cavity, as well as the length and volume of the vocal tract, are all significantly positively correlated with body height. In recent years, the frequency code hypothesis and sound symbolism have also been used to investigate how human voices match perceptual impression of femininity, vulnerability, submissiveness, politeness, friendliness, insecurity, uncertainty or charisma, and so on (e.g., Grawunder and Winter, 2010; Noble and Xu, 2011; Signorello et al., 2012; Mixdorff et al., 2018; Cartei et al., 2019; Yang et al., 2020; Rallabandi et al., 2021; Weiss et al., 2021), as well as the naming of animation characters when dubbing in Walt Disney cartoons (Lippi-Green, 2012) or *Pokémon* names (Kawahara et al., 2018). Since high vowels tend to have a higher intrinsic F0 than low vowels (Chen et al., 2021), the names of characters with initial high vowels tend to be smaller and lighter in *Pokémon* (Kawahara et al., 2018). In addition, Puts et al. (2006) suggested that the pitch of the male voice may reflect his perceptions of his dominance. Specifically, men who believe they are physically dominant to their competitors lower their voice pitch, whereas men who believe they are less dominant raise their pitch. Stern et al. (2021) also indicated a significant negative correlation between voice pitch and self-reported sociosexuality, dominance, and extraversion; moreover, lower voice pitch is perceived as being more attractive in men. All the studies mentioned above demonstrate the importance of pitch in the construction of a speaker's identity.

1.2. Voice, voice quality, and voice acting

The definition of dubbing also includes both broad and narrow perspectives. Broadly, dubbing includes all sound elements in a film or TV drama. In the narrow sense, dubbing refers to the creative activity of adding voice to characters by voice actors, adding extra voices and narration, or replacing the dialogue in the original film with another language (Liu, 1994, p. 4). In order to convey identity and to portray a character actively, dubbing requires the conscious use of his/her voice's ability to convey the speaker's message, and to add to films or TV dramas a voice that fits the character's image. In *Animation Sound Design*, Zhao et al. (2015) mentioned that "sound can also have the same narrative effects as images, and is even, to some extent, precedent over visual expression". Dubbing can make characters more vivid, bringing new vitality to the film and animation industry.

The method for dubbing to portray characters in film and TV drama is mainly to match the perceptual impression produced by the dubbed voice to the character's image. Apart from characters who use the contrast between appearance and voice to reach a comedy effect (e.g., Lina Lamont in *Singin'in the Rain*, whose beautiful, gentle appearance contrasts sharply with her shrill, harsh voice, Donen et al., 1952), dubbing, in general, needs to match and highlight the uniqueness of the character auditorily, thereby enhancing the audience's recognition of the character's image. For example, in the film *My Fair Lady*, despite being a flower girl, the leading female is considered an actual princess because of speaking a fluent upper-class style accent (Shaw and Fisher, 1963; Cukor, 1964). In Japanese animations, the epilaryngeal settings, which played a major role in distinguishing four heroic and villainous voice types, were identified as the auditorily critical vocal components that differentiate good and evil characters (Teshigawara, 2003). Moreover, to portray unique and distinctive characters, voice actors often use several stereotypical phonation types or actively raise pitch significantly to affect the listener's perception of the character's voice quality, further reinforcing these stereotypes. For example, popular

American media usually use harsh voice and exaggerated emotional states associated with yelling/shouting modes of expression to portray the racial stereotype of black people (Moisik, 2012). Moreover, voice with a Yiddish accent once used to make the stereotype of wolf complete in Disney cartoon (Lippi-Green, 2012, p. 107). Stereotypes on homosexuality also influence dubbing. When acting for homosexual males, voice actors often choose a higher F0, falsetto, or higher formant frequencies to present effeminacy (Podesva, 2007; Cartei and Reby, 2012). Similarly, using creaky voice in Chinese films and TV dramas may be considered a sign of promiscuity (Callier, 2010).

According to the broader definition of voice quality, the perceptual impressions produced by dubbed voices are their voice quality, and the perceptual differences are caused by physiological and acoustic changes reaching a certain level. Thus, there is an acoustic basis for voice quality changes. Sweet voice is often used in Japanese animation for mature and traditional female characters, such as mothers, elder sisters, and teachers. Moreover, they have a lower pitch than the voice of the leading females. As described by Klatt and Klatt (1990) and Pépiot (2014), breathy voice is considered to be feminine; similarly, sweet voice also shows breathy voice (Starr, 2015). In addition, in order to discriminate between protagonists and antagonists acoustically, Teshigawara (2003) investigated the acoustic characteristics of voice quality using the voice of heroes and villains in Japanese animation, indicating that voices of villains usually have a lower second formant (F2) and more high-frequency energy. Tong and Moisik (2021) analyzed the voice of protagonists and antagonists in American cartoons using long-term average spectrum (LTAS), and found that the protagonists and antagonists exhibit high- and low-frequency dominance in their spectral profiles, respectively. Meanwhile, harsh voice is also often used to portray lazy and brutal villains in Chinese films and TV dramas (Callier, 2012).

1.3. Research question and purpose

With the great demand for voice acting in film and TV dramas, the academic community has also paid more attention to acoustic research on voice acting. The commonly used methods include harmonic analysis, LTAS, principal component analysis, etc. By reviewing theoretical and empirical studies on voice acting in the literature, it is not difficult to find that most existing studies try to explore the acoustic characteristics that distinguish the voices of characters of different types through single-parameter acoustic analysis. However, the nature of the voice quality is multi-dimensional. Therefore, it is only by introducing more acoustic parameters into the study of voice acting that we can discuss in-depth which acoustic parameters are the key indicators that distinguish different characters. In other words, the distinctive features across characters need to be sought by means of acoustic analysis to characterize different styles. Furthermore, with the increase in the quality of film and TV dramas, the demand for personalized dubbing continues to grow, placing greater demands on voice actors' expertise. Unfortunately, the existing dubbing guidance is usually subjective and lacks practicality and operability. For example, maintaining a "sense of drama" and "elasticity of voice", achieving a "richness of both voice and emotion" and improving "characterization" (Yang L., 2021; Yang Z. S., 2021) are too general, and remain at the level of subjective descriptions, not providing a scientific explanation for acoustic characteristics of personalized dubbing, making it difficult to grasp in practice.

Considering these problems mentioned above, this study focuses on the acoustic characteristics of the dubbed voices and their matching with the characters' image. The following two major questions need to be addressed. One is whether there are acoustic differences between different characters' dubbings, as well as the matching between acoustic parameters and the character's image. The other is what kind of reference we can provide to voice actors for improving the matching between the characters and their dubbing based on the acoustic analysis results.

2. Materials and methods

The speech data used in this study was taken from the entire TV series *Empresses in the Palace* (甄嬛传). The reason for choosing this TV series is that this TV series has a high reputation and is more prototypical among Chinese costume dramas. As of August 2022, it has accumulated 12.81 billion views on LeTV alone, which is a much wider audience. In addition, the TV series has generated a vibrant secondary creation, attracting a larger audience.

2.1. Information about the selected characters

In this study, five major male characters were selected from *Empresses in the Palace*, and their "character status" was determined according to their character image (see Table 1).

2.2. Speech materials

Adobe Premiere Pro 2020 was used to extract audio files from MP4-format video files. The sampling rate of speech signals is 44 kHz, with 16-bit sampling resolution, and the recording is monophonic. The speech material consists of monosyllabic and sentence files.

TABLE 1 Information of the selected characters.

No.	Character name	Character identity	Character status: dominator/subordinate	Relative social status	Age of the voice actor
M01	AISIN-GIORO Yinzhen	Monarch, an emperor in power	Dominator	High	41
M02	AISIN-GIORO Yunli	Prince, a noble with little power	Both	Low	37
M03	WEN Shichu	Imperial physician of the Court	Subordinate	Low	33
M04	SU Peisheng	Chief eunuch of the court	Subordinate	Low	60
M05	ZHANG Tingyu	Minister, loyal and high-ranking	Both	Low	45

"Both dominator and subordinate" refers to the role's having different relative status in relation to other roles.

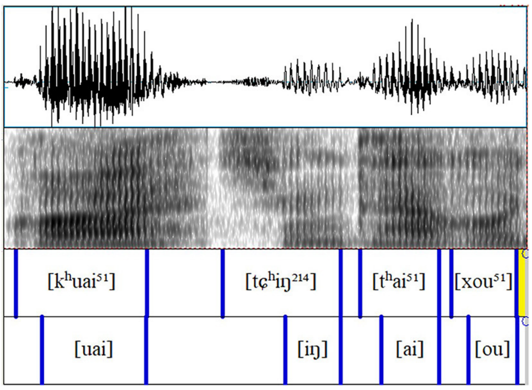

FIGURE 1
An example of annotated monosyllabic tokens "快kuài请qǐng太tài后hòu (Go and invite the Empress Dowager!)".

TABLE 2 Duration of the sentence samples of each character (two decimal places).

Character	Total duration (s)	Average duration of each sample (s)	S.D. (s)
M01 (Monarch)	241.05	40.18	1.24
M02 (Prince)	245.42	40.90	1.42
M03 (Physician)	248.47	41.41	4.66
M04 (Eunuch)	254.91	42.49	2.85
M05 (Minister)	266.72	44.45	2.80

2.2.1. Monosyllabic tokens

The obtained audio files were divided by character, with 480 monosyllabic tokens for each character, with a total of 2,400 tokens for all five characters. The final (rhyme) part of each token was then annotated in *Praat 6.1.37* (Boersma and Weenink, 2021), with the initials, finals, and tones annotated in the first tier, and the finals also separately annotated in the second layer, as shown in Figure 1.

2.2.2. Sentence samples

To make sure that the duration of sentence samples for each character was long enough and similar, sentences of varying lengths were joined into a total of 30 sentence samples of around 40 s in *Praat* software, with a total duration of 1,256 s, while avoiding strong emotions, noise, background music, and sound effects segments as much as possible. Details of sentence samples are given in Table 2.

2.3. Acoustic parameters and analytical procedure

Based on previous studies discussed in the introduction section (e.g., Ohala, 1984; Starr, 2015; Stern et al., 2021; Tong and Moisik, 2021), this study mainly focuses on acoustic parameters that are closely related to voice quality, and that are mainly used to represent pitch, harmonic-to-noise ratio, and spectral energy intensity.

One measure is F0. It refers to the frequency at which the vocal folds vibrate, i.e., the first harmonic (H1) in the spectrum, which is a physical quantity in acoustics. Pitch, on the other hand, is the perception of the height of a sound, and is a psychological concept (Kong, 2015). The primary physical quantity that carries pitch is F0 (Liu, 1924). F0 plays a vital role in voice quality perception, and can determine the pitch of the voice quality. Generally speaking, the higher the F0, the higher the perceived pitch.

The other measure is cepstral peak prominence (CPP), which is defined as the amplitude of the cepstral peak, measured based on normalized overall amplitude in the spectrum (Hillenbrand et al., 1994). And a speech signal whose spectrum shows a well-defined harmonic structure will show a very prominent cepstral peak (Hillenbrand and Houde, 1996;

Miramont et al., 2020). Thus, CPP can be used to quantify the periodicity and harmonic-to-noise ratio of the speech signal. In general, the more periodic the speech signal, the weaker the noise, and the greater value of the CPP. On the contrary, the smaller the CPP, the stronger the noise. Therefore, CPP is also often used to quantify differences among phonation types, and is considered a reliable acoustic parameter for discriminating breathy voice from non-breathy voice (Blankenship, 2002). Hartl et al. (2003) also indicated a negative correlation between breathy voice and CPP.

In addition, the commonly used harmonic amplitude parameters (H1-H2, H2-H4, etc.) and long-term average spectrum (LTAS) are also selected to investigate the energy distribution in different frequency ranges in the spectrum. The selection of harmonic amplitude parameters is based on Kreiman et al. (2014), who proposed the following four harmonic components to be particularly important in the simulation of voice source spectrum, namely, H1-H2 (the amplitude difference between the first harmonic and the second harmonic), H2-H4 (the amplitude difference between the second harmonic and the fourth harmonic), H4-H2k (the amplitude difference between the fourth harmonic and the harmonic nearest 2,000 Hz), H2k-H5k (the amplitude difference between the harmonic nearest 2,000 Hz and the harmonic nearest 5,000 Hz). In addition, H1-A3 (the amplitude difference between the first harmonic and the harmonic nearest to the third formant) is also a useful parameter in studying phonation types (Iseli et al., 2007). The amplitude difference reflects the strength of the spectral energy attenuation among different frequency ranges. The larger the amplitude difference, the stronger the spectral energy attenuation in that range, and the greater the spectral tilt. Specifically, H1-H2 is proportional to the open quotient (OQ), which reflects the duration of the open phase of the vocal folds within a glottal cycle. The larger the OQ, the less tightly closed the vocal folds, the more the airflow leak, the stronger the spectral energy attenuation, and the more prominent the breathy voice (Ladefoged, 2003, p. 178–181; DiCanio, 2009). H2-H4 is the auxiliary measuring parameter to determine phonation types, and has been used among cross-linguistic studies to compare voicing between different phonetic systems (Esposito, 2006). H4-H2k, H1-A3, and H2k-H5k represent the degree of spectral energy attenuation and spectral tilt at low, low-mid, and mid-high frequency ranges respectively. On the other hand, LTAS is described by Leino (2009) as "a means of viewing the average frequency distribution of the sound energy in a continuous speech sample", reflecting the distribution of spectral energy across frequency ranges. Note that the prerequisite for using LTAS is that the duration of the speech sample is long enough so that the linguistic content of the speech sample can be ignored, and the interference of non-speech components can be avoided, to focus on the personal characteristics of the speaker's voice (Pittam, 1987; Mendoza et al., 1996). On the basis, Li et al. (1969) stated that the duration of the speech sample should be at least 30 s, while Fritzell et al. (1974) stated that LTAS results are more stable and reliable when the speech signal is ∼40 s.

In this study, F0, CPP, and harmonic amplitude parameters were extracted using *VoiceSauce* (Shue et al., 2009), and LTAS was extracted using *Wavesurfer* (Sjölander and Beskow, 2000). On this basis, *Z*-scores were calculated using *SPSS Statistics 26.0*, and data with *Z*-scores >2 or <-2 (∼5%) were considered outliers and removed. The data were then tested for normality using *Origin 2021*. A *t*-test was conducted for data that follow a normal distribution, otherwise a Kruskal-Wallis non-parametric test was conducted. Finally, multi-dimensional scaling analysis (MDS) was carried out using *SPSS Statistics 26.0*.

3. Results

3.1. Pitch

The pitch measurements for different character voices are analyzed first. Figure 2 shows a boxplot of the F0 data, where data beyond two standard deviations are set as outliers. The value of mean F0 demonstrates that: M02 (Prince) > M05 (Minister) > M04 (Eunuch) > M01 (Monarch) > M03 (Physician). Specifically, M03 (Physician), and M01 (Monarch) have lower mean F0, which are 96.83 and 107.78 Hz, respectively. M02 (Prince) and M05 (Minister) have higher mean F0, 128.57 and 127.02 Hz respectively. M04 (Eunuch) has an intermediate mean F0, which is 113.51 Hz. M02 (Prince), M04 (Eunuch), and M05 (Minister) have larger F0 variation range (outliers included), and larger standard deviations, which are 48–278 Hz (31.01), 46–221 Hz (20.25), and 47–244 Hz (25.87), respectively. However, the F0 variation and standard deviation of M01 (Monarch) and M03 (Physician) are relatively lower, which are 46–191 Hz (15.50) and 71–154 Hz (9.66), respectively.

The following two conclusions can be drawn from the F0 mentioned above. First, based on the relationship between F0 and pitch, it is not difficult to find that M01 (Monarch) and M03 (Physician) have lower mean F0, which suggests that they have a lower pitch, with a deep voice. On the contrary, M02 (Prince) and M05 (Minister) have larger mean F0, higher pitch, and relatively brighter voice quality. Combining Figure 2 with Table 1, we found a clear difference in the social status of the five characters, creating an identity dichotomy between dominators and subordinates. M01 (Monarch), the highest social status dominant of them all, has a lower pitch. The other characters are in the position of subordinate relative to M01 (Monarch). M02 (Prince), M04 (Eunuch), and M05 (Minister) have a relatively higher pitch, but M03 (Physician) has a lower pitch, which will be discussed in detail in section 4. Second, the range of F0 variation, i.e., the pitch range, affects the intonation of the dubbed characters. Chao (1968, p. 39) used the analogy of "large waves" and "small ripples" to illustrate the relationship

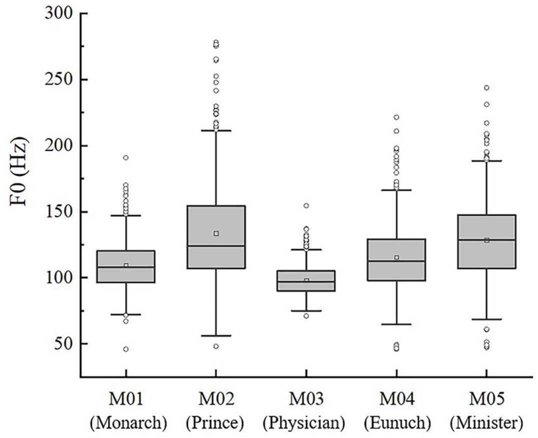

FIGURE 2
F0 distribution of the five characters.

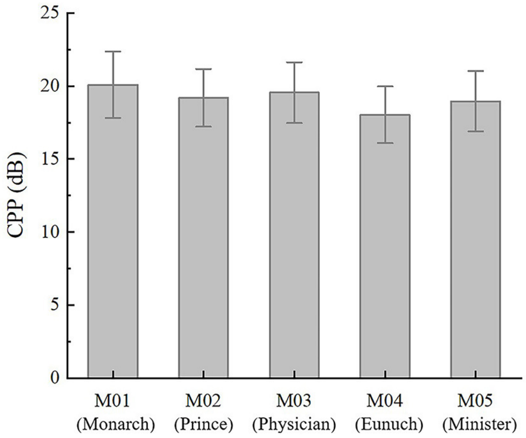

FIGURE 3
CPP of the five characters.

between intonation and tone in Chinese, in which the tone is superimposed on intonation. "Intonation is the pattern of the pitch movement of an utterance", and pitch is one of its constituent elements (Cao, 2002; Ding, 2005). Among the male characters, M02 (Prince) and M05 (Minister) have the most extensive F0 range and standard deviation, indicating that they have large F0 fluctuations during phonation. Their voices are with a lilt, with a sense of rhythm and rhyme, and full of emotion. Compared to the voices of the other characters, M01 (Monarch) and M03 (Physician) have a small range of F0 variation, and their voices lack fluctuation. The flatness of the intonation makes the voice sound more calm, or suppressed in emotion.

3.2. Harmonic-to-noise ratio

CPP can be used to measure the periodicity and harmonic-to-noise ratio of the speech signal in the previous literature, which is an essential acoustic parameter for quantifying phonation types (e.g., Hillenbrand et al., 1994; Hillenbrand and Houde, 1996; Blankenship, 2002; Hartl et al., 2003; Miramont et al., 2020). Figure 3 shows a boxplot based on the CPP for each character's dubbed voice. The results for mean CPP indicate that: M01 (Monarch) > M03 (Physician) > M02 (Prince) > M05 (Minister) > M04 (Eunuch). To be specific, M01 (Monarch) has the highest CPP (20.12 dB). M04 (Eunuch) has the lowest CPP (18.06 dB). CPP of the other characters are in between. CPP of the five characters have small standard deviations, which are between 1.7 and 2.2, showing a concentrated distribution pattern. On the whole, M01 (Monarch) has the largest CPP, indicating that this voice has the least noise component in its speech signal. The CPP of M04 (Eunuch) is significantly lower than that of the other characters, indicating a significant noise component in his speech signal. So the turbulent noise can be perceived in his voice.

3.3. Harmonic measures and long-term average spectrum

As discussed in section 2.3, harmonic parameters can be used to analyze spectral energy distribution in the speech signal. These measures represent the attenuation of spectral energy, in which higher values indicate a more substantial energy attenuation in that frequency range. The five harmonic parameters are shown in Table 3 and Figure 4 for each character. Figure 5 is the LTAS of each character.

H1-H2 and H2-H4 indicate the strength of energy attenuation in low frequency, which are important indicators for measuring breathy voice. The larger their values, the more serious the airflow leak through the glottis. Table 3 and Figure 4 show that M04 (Eunuch) has the largest H1-H2 and a relatively large H2-H4, suggesting a strong attenuation of its spectral energy in low frequency. M02 (Prince) has a relatively small H1-H2 and the smallest H2-H4, indicating a weak attenuation of its spectral energy in low frequency. The remaining three characters have an overall intermediate range of energy attenuation in low frequency. H4-H2k and H1-A3 indicate the strength of energy attenuation in low-mid frequency. M03 (Physician) and M04 (Eunuch) both have a larger or the largest H4-H2k and H1-A3, both of which have strong spectral energy attenuation in low-mid frequency. M02 (Prince) and M05 (Minister) have the smallest values, with weak spectral energy attenuation in low-mid frequency. M01 (Monarch) has a smaller H4-H2k and a larger H1-A3, with intermediate energy attenuation in low-mid frequency. H2k-H5k indicates the strength of energy

TABLE 3 Harmonic measures of the five characters (mean ± S.D., in dB, two decimal places).

Character	H1-H2	H2-H4	H4-H2k	H1-A3	H2k-H5k
M01 (Monarch)	5.01 (5.91)	5.27 (5.51)	2.08 (7.38)	13.67 (9.08)	28.84 (7.31)
M02 (Prince)	5.05 (3.87)	3.35 (6.09)	1.96 (7.54)	8.62 (8.81)	29.88 (8.21)
M03 (Physician)	6.30 (3.32)	3.58 (6.98)	5.75 (8.42)	13.45 (10.18)	28.41 (11.54)
M04 (Eunuch)	7.00 (4.08)	5.23 (6.43)	4.30 (7.78)	16.01 (9.70)	25.90 (10.46)
M05 (Minister)	5.35 (3.54)	4.24 (5.45)	1.48 (7.97)	7.96 (10.26)	27.67 (9.16)

The harmonic measures were extracted with vowel formants corrections (Iseli et al., 2007), except for H5k.

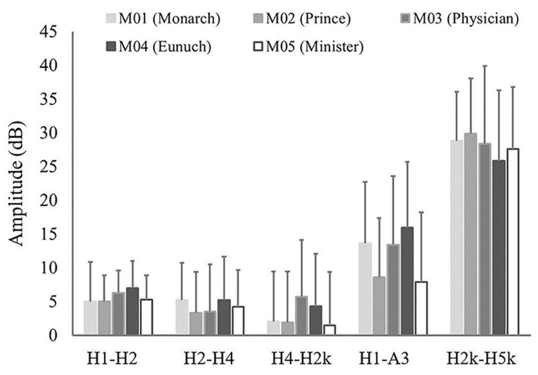

FIGURE 4
Harmonic measures of the five characters.

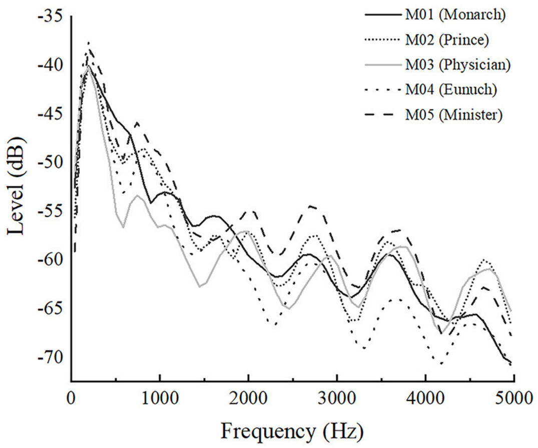

FIGURE 5
LTAS of the five characters.

attenuation in mid-high frequency. M02 (Prince) has the largest H2k-H5k, and its spectral energy attenuation is the strongest in mid-high frequency. M04 (Eunuch) has the smallest H2k-H5k, and its spectral energy attenuation is the weakest in mid-high frequency. The remaining three have intermediate energy decay in mid-high frequency.

3.4. Voice quality

The distinctive features of the voice of the five characters can be obtained when combining the acoustic characteristics of all parameters in each character's voice (see Table 4).

F0 shows the pitch level of the character's voice, while CPP and the strength of spectral energy attenuation (i.e., spectral tilt) together show whether the character's voice has breathy voice. Usually, the lower the CPP, and the stronger the overall attenuation of the spectral energy, the higher the degree of breathy voice. The voice quality of the five characters in Table 4 has the following characteristics. To be specific, in terms of F0, the dominant M01 (Monarch), who has the highest social status, has a lower F0 and a lower voice pitch, while the subordinate M05 (Minister) and M02 (Prince), who have a relatively low social status, both have a higher F0 and a higher voice pitch. Regarding phonation types, M01 (Monarch) has a neutral level of all harmonic parameters and the largest CPP, with no obvious non-modal phonation characteristics, which can be considered a modal voice among the five characters. Within the other characters, the higher social status dominators M02 (Prince) and M05 (Minister) have a weaker energy attenuation than the other characters, a weaker noise component in the speech signal, and no apparent breathy voice. The lower social status subordinate characters M03 (Physician) and M04 (Eunuch) have an overall stronger attenuation in spectral energy, a stronger noise component in the speech signal, and obvious breathy voice. It can be seen that, when dubbing for a subordinate character, the voice actors tend to choose to raise pitch or use breathy voice; however, when dubbing for a dominant character, the actor tends to lower pitch and does not adopt breathy voice.

3.5. Statistical tests and multi-dimensional scaling

The five characters can be easily distinguished from the perspective of subjective auditory perception. In terms of acoustic performance, the voices of the five characters also

TABLE 4 Distinctive features of the character voice of the five characters.

	M01 (Monarch)	M02 (Prince)	M03 (Physician)	M04 (Eunuch)	M05 (Minister)
F0	–	+	–	±	+
CPP	+	±	±	-	±
Spectral tilt	±	–	+	+	–

"+" Indicates that the character's voice has this acoustic characteristic, "–" indicates that it does not, and "±" indicates that the level is intermediate. Determination of the spectral tilt combines the low, low-mid, and mid-high frequency energy attenuation in section 3.3.

TABLE 5 Results of statistical tests.

Pairs	F0	CPP	H1-H2	H2-H4	H4-H2k	H1-A3	H2k-H5k
M01-M02	***	***	**	***	n.s.	***	*
M01-M03	***	***	n.s.	***	***	*	n.s.
M01-M04	***	***	***	n.s.	***	***	***
M01-M05	***	***	n.s.	**	n.s.	***	n.s.
M02-M03	***	**	***	n.s.	***	***	n.s.
M02-M04	***	***	***	***	***	***	***
M02-M05	n.s.	*	n.s.	**	n.s.	n.s.	**
M03-M04	***	***	***	***	**	***	***
M03-M05	***	***	**	**	***	***	n.s.
M04-M05	***	***	***	*	***	***	**
Discrimination rate	90%	100%	70%	80%	70%	90%	60%

The statistical tests are conducted using *t*-test or Kruskal-Wallis test as mentioned in section 2.3. Specifically, n.s., $p > 0.05$, not significant. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

differ acoustically by comparing F0, CPP, and harmonic parameters. These phenomena make us wonder whether voices that differ significantly in speech perception are also acoustically significantly different. Which acoustic parameters are the key acoustic indicators that distinguish character voices? Which characters have more similar voices than others, and which have more different voices? Based on the questions mentioned above, we combine the five characters in pairs to form a total of 10 pairs. According to the normality test results, parametric or non-parametric tests are conducted on the F0, CPP, and harmonic parameters of the voices. For each parameter, the ratio of significant differences in the 10 pairs is calculated as the discrimination rate of that parameter. The results show only significant differences among some parameters for some characters (see Table 5).

As can be seen from Table 5, CPP shows significant or highly significant differences between all 10 pairs ($p < 0.05$ or $p < 0.01$), suggesting CPP is significantly different among all five characters. F0 shows highly significant differences ($p < 0.01$ or $p < 0.001$) for each pair, except between M02 (Prince) and M05 (Minister). That is, with the exception of M02 (Prince)-M05 (Minister), the rest of the character pairs can be discriminated from each other by F0. For H1-A3, with the exception of M02 (Prince)-M05 (Minister), there are significant or highly significant differences between all pairs ($p < 0.05$ or $p < 0.001$). For H2-H4, with the exception of the pairs M01 (Monarch)-M04 (Eunuch) and M02 (Prince)-M03 (Physician), there are significant or highly significant differences between all pairs ($p < 0.05$ or $p < 0.01$). There is no significant difference for H1-H2 between the pairs M01 (Monarch)-M03 (Physician), M01 (Monarch)-M05 (Minister), and M02 (Prince)-M05 (Minister), and no significant difference for H4-H2k between the pairs M01 (Monarch)-M02 (Prince), M01 (Monarch)-M05 (Minister), and M02 (Prince)-M05 (Minister). For H2k-H5k, only six of 10 pairs shows significant difference ($p < 0.05$ or $p < 0.01$), while the differences between other pairs are not statistically significant.

The statistical tests reveal whether the voices of the five characters differ significantly on each of the seven acoustic parameters. The multi-dimensional scaling (MDS) analysis is adopted to show the distribution of the voices of the five characters in low dimensions more clearly. MDS is used to reduce the high-dimensional space of the distance between the characters' voices, measuring seven parameters including F0, CPP, H1-H2, etc., into a lower-dimensional space. The degree of dissimilarity of the five character voices in two dimensions is explored using the distance among characters (Torgerson, 1952; Cox and Cox, 2008). The results are shown in Figure 6.

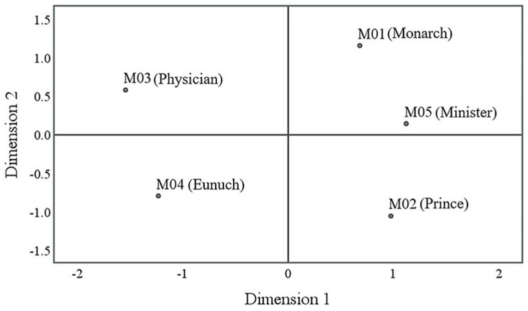

FIGURE 6
Multi-dimensional scaling of the character voices.

TABLE 6 Correlation coefficient between acoustic parameters and MDS coordinates (three decimal places).

Acoustic measures	Correlation with dimension 1	Correlation with dimension 2
F0	0.755	−0.612
CPP	0.317	0.753
H1-H2	−0.898	−0.258
H2-H4	−0.079	0.286
H4-H2k	−0.979	0.0559
H2k-H5k	0.524	0.154
H1-A3	−0.775	0.199

Correlation coefficient between each acoustic parameter (averaged measurement for each character) and the coordinates of these characters in each MDS dimension.

Figure 6 demonstrates the distribution of the five voices in a two-dimensional space, with distances that reflect their mutual dissimilarity. According to the statistical result (Stress = 0.17, RSQ = 0.69), the goodness-of-fit indexes are fairly good. In terms of the degree of dissimilarity among the character voices, an intuitive interpretation of Figure 6 is that, the further apart the character voices in space, the more significant the difference in voice quality among the character voices; and conversely, the closer the distance, the more similar the voice qualities of the character voices. According to Figure 6, M02 (Prince) and M05 (Minister) are closer in the space, indicating that they are acoustically similar. Moreover, Table 5 also shows that in 57% of cases, the acoustic parameters do not discriminate between these two voices. M01 (Monarch) is also closer to M05 (Minister), and there is no significant difference between their acoustic parameters in 43% of the cases. The distances between M01 (Monarch), M02 (Prince), M03 (Physician), and M04 (Eunuch) are all farther apart from each other, indicating that they are acoustically different. Table 5 shows a significant difference between the acoustic parameters of all these four voices in 86% of cases. In terms of phonation types, this space can be divided into three parts depending on the degree of breathy voice, namely, obvious breathy voice (M03 Physician and M04 Eunuch), modal voice (M01 Monarch), and no obvious breathy voice (M02 Prince and M05 Minister). Dimension 1 can form a continuum from no breathy voice (right) to breathy voice (left). In terms of the relative social status of the characters, Dimension 2 can similarly divide the character voices into two parts, that is, higher relative social status (M01 Monarch) and lower relative social status (M02 Prince, M03 Physician, M04 Eunuch, and M05 Minister), forming a continuum concerning the relative social status of the characters.

The result of MDS in Figure 6 is constructed based on seven acoustic measures, while we need to further explore the extent to which each of these parameters is related to these two dimensions. Therefore, a correlation analysis was conducted between each acoustic parameter (averaged measurement for each character) and the coordinates of these characters in each MDS dimension. The results are shown in Table 6. The correlation coefficient between each acoustic parameter and each dimension represents the extent to which the parameter can explain the variation of the corresponding dimension. Specifically, the larger the absolute value of the correlation coefficient, the better this acoustic parameter reflects the corresponding dimension. In Table 6, H4-H2k, H1-H2, H1-A3, and F0 are strongly correlated with Dimension 1, suggesting that these four acoustic parameters can better explain the "degree of breathy voice". The harmonic parameters are negatively correlated with Dimension 1, and F0 is positively correlated with Dimension 1, which is in line with the cross-linguistic studies indicating that "breathy voice may lead to a reduction in F0" (Liu et al., 2020; Liu, 2021). CPP and F0 correlate strongly with Dimension 2, suggesting that these two acoustic parameters better explain the "relative social status of the characters". CPP is positively correlated with Dimension 2, and F0 is negatively correlated with Dimension 2, which is in line with previous research indicating that "dominant characters tend to use creaky voice" (e.g., Puts et al., 2006; Stern et al., 2021). As the "dominant" and "subordinate" roles in this study are analyzed in two dimensions: "degree of breathy voice" and "relative social status of the character", the interpretation and prediction of role types using acoustic parameters also requires the two separate dimensions.

The following two conclusions can be drawn from these results. First, only one of the seven parameters (i.e., CPP) can 100% discriminate the characters' voices, suggesting that it is difficult to completely distinguish the voices using a single acoustic parameter. Nonetheless, some consistent patterns can also be found in the data. For example, CPP, F0, and H1-A3 are the key acoustic indicators for distinguishing character voices. CPP has the highest discrimination rate (100%), F0 and H1-A3 have a higher discrimination rate (90%), followed by H2-H4 (80%), H1-H2 and H4-H2k (70%), and H2k-H5k the lowest (60%). Second, in the two-dimensional space, Dimension 1 can be interpreted as the degree of breathy voice, and Dimension 2 can be interpreted as the characters' social status level. In terms of similarity among character voices, the pair M02 (Prince)-M05 (Minister) and the pair M01 (Monarch)-M05 (Minister) are closer in space distance, and the discrimination rate between the acoustic parameters of each pair is mostly at a lower level, with relatively similar acoustic performances. M01 (Monarch), M02 (Prince), M03 (Physician), and M04 (Eunuch) are distant from each other in this space, and there are significant differences between a vast majority of their acoustic parameters, with the acoustic performance of the four being more distinctly different. Thus, the statistical test results for the acoustic parameters are consistent with the result of the MDS for acoustic distance, namely, character voices with significant differences between acoustic parameters being more acoustically distant, and conversely, character voices with smaller differences between acoustic parameters being less acoustically distant. Finally, for the interpretation and prediction of role types using acoustic parameters, these two separate dimensions are both required. H4-H2k, H1-H2, H1-A3, and F0 can better explain Dimension 1—"degree of breathy voice"; and CPP and F0 can better explain Dimension 2—"relative social status of the characters".

4. Discussion

It is well-known that voice acting aims to increase the audience's recognition of the character's image auditorily, with the ultimate goal of making a perfect combination of the re-creation using audible speech and the original character (Li, 2007). Based on the acoustic experiment, this section discusses the matching between acoustic parameters and character images, and the implications for guiding dubbing practice.

Firstly, in terms of pitch, adult males' larynx and pharyngeal cavities grow fast due to gender differentiation at puberty, causing an increase in the length and thickness of the vocal folds, leading to a significantly lower frequency of vocal fold vibration than in females. F0 has already been proven to be a reliable cue for distinguishing the voices of adult males and females in many previous studies (e.g., Murry and Singh, 1980; Honorof and Whalen, 2010). In general, low pitch is considered a distinctive feature of the adult male voice. In addition, cross-linguistic studies have found that the pitch range for normal adult male speech is between 80 and 180 Hz, and the range usually does not exceed 100 Hz, with a median of ca. 140 Hz (e.g., Baken and Orlikoff, 2000; Keating and Kuo, 2012; Kuang, 2013; Liu, 2019). According to section 3.1, the average F0s of all characters have a maximum of 128.57 Hz and a minimum of 96.83 Hz, all below 140 Hz, which is at the lower level of normal male pitch. Thus, the voices of all five characters are typically of heterosexual masculine temperament in pitch (Podesva, 2007; Cartei and Reby, 2012). Since the low pitch is somewhat in accord with the male voice's stereotype, it helps to further cast the character's image, making use of the audiences' perception of the lower pitch to position the character quickly. In terms of pitch range, except for M03 (Physician), the pitch range of all other characters' voices reached at least 145 Hz, with M02 (Prince) even reaching 230 Hz, making the pitch range much larger than normal male speech (80–180 Hz as mentioned above). As the F0 reflects not only the tonal information but also the intonational information, it is an important acoustic indicator for the expression of emotional intonation (Zhang et al., 2008). Therefore, intonation is generally changed through changing frequency (Zhang et al., 2021). Usually, there is a greater range of F0 variation for intense emotions such as cheerfulness and anger, and a smaller range of F0 variation for inhibited emotions such as calmness or sadness (Gao et al., 2005; Jia, 2017). Regarding the five characters, M02 (Prince) and M05 (Minister) have a higher pitch range with more dramatic change, making voices with a lilt, creating the image of gentle and affectionate literati who likes poetry, and a loyal minister who is impassioned and forthright in his advice, respectively. Thus, voice actors can exaggerate the pitch variation when dubbing a character, which helps to emphasize the character's feelings in the scene, and to enhance the actors' expressiveness and the audience's sense of immersion. At the same time, voice actors can also make use of the effect of pitch on perception, creating prototypical heterosexual male figures by lowering F0, as well as intensifying or suppressing emotions by increasing or decreasing F0 variation, in order to influence the audiences' perception of the plots and the character images, facilitating a better auditory portrayal of the characters.

Secondly, similar to the sound pattern of "frequency code", numerous previous studies have shown that creaky voice is often associated with higher social status or greater dominance, whereas whispery voice and harsh voice are often associated with lower social status or role of subordinate (Esling, 1978; Ohala, 1984; Yuasa, 2010; Hornibrook et al., 2018; Tavi et al., 2019). This viewpoint is consistent with the results in section 3.4. The dominant characters, i.e., M01 (Monarch), M02 (Prince), and M05 (Minister), have no significant breathy voice compared to the other characters. On the contrary, the subordinate characters, i.e., M03 (Physician) and M04 (Eunuch), significantly use breathy voice. On this basis, in voice training or practice, voice actors can use modal or creaky voice to cast the dominant roles. The subordinate roles should be built using breathy voice. Based on the relationship between physiology, acoustics, and perception, a relationship between a type of character image and a specific voice quality can be established, so that the dubbed voice can be in accord with the character image, to match the audience's perception, and to enhance the credibility and impact of the dubbed characters. Moreover, cross-linguistic studies have shown that breathy voice is usually accompanied by a lower F0 compared to modal voice (Liu et al., 2020; Liu, 2021). M03 (Physician) and M04 (Eunuch) have stronger energy attenuation in low and low-mid frequencies, and there is a significant glottal leak during phonation, which causes lower F0, and pitch is lower than expected. Therefore, when dubbing for subordinate characters who require the use of breathy voice, the voice actor may choose to appropriately lower the pitch, in order to facilitate a smoother and sustained production of breathy voice, thus reducing the difficulty of the voicing.

Finally, voice plasticity is the prerequisite for voice actors to dub for different roles; however, since dubbing needs to fit the personal character, age, and other factors, the possible range of the voice actor's performance is limited by the quality of his/her voice quality, which is determined by the physiological conditions of the voice actor (Kreiman and Sidtis, 2011; Gao, 2013). The reconcilement of the character's image with the physical condition of the voice actor, and bringing out the uniqueness of the actor's voice, are vital considerations. For example, M04 (Eunuch) is a low-ranked and hard-working character, and the choice of an older voice actor (60-year-old, see Table 1) to dub his voice improved the perceived suitability between the voice and the character. From a physiological point of view, aging leads to atrophy of the vocal folds, which are part of the thyroarytenoid muscle, and a significant increase in glottal width, resulting in severe airflow leak through the vocal folds during phonation. This leads to a stronger attenuation in the voice's overall energy, and the appearance of obvious breathy voice (Fischer-Jørgensen, 1967; Dave, 1968; Winkler and Sendlmeier, 2006; Kreiman and Sidtis, 2011, p. 117; Gregory et al., 2012). In terms of acoustics, M04 (Eunuch) has a lower or the lowest spectral energy in most of the frequency ranges, and a strong energy attenuation. As a result, the dubbed voice of M04 (Eunuch) has a sense of fatigue due to its low energy, and a sense of humbleness due to the breathy voice. This voice quality also fits the character's image of being usually unable to speak loudly due to his low social status, which facilitates the portrayal of this character perfectly. Due to the physical limitations, all voice actors' voices are limited to some extent, so measuring the range of roles suitable for the physiological conditions of an actor can contribute, both to the producers' selection of voice actors, and to the personal career planning of a voice actor.

5. Conclusion

This study came to the following three conclusions utilizing acoustic analysis based on the dubbed voice of male characters from *Empresses in the Palace*.

First, the voice quality of the five characters has the following characteristics. Regarding F0, characters with higher social status use a low pitch, while those with relatively lower status adopt a high pitch. In terms of phonation type, the voice of male characters with higher social status do not use breathy voice notably, compared to the characters with lower status, who use breathy voice frequently. Thus, when dubbing for subordinate characters in *Empresses in the Palace*, voice actors tend to raise their pitch or to use breathy voice. On the contrary, low pitch and modal voice or creaky voice are applied to dominating characters in this TV series.

Second, although the dubbed voice of the five characters can be well-discriminated in auditory sensation, CPP is the only single acoustic parameter that can discriminate all five characters, followed by F0 and H1-A3. The three parameters mentioned above are the key acoustic indicators for discriminating character voices. Thus, multiple parameters are of great importance in discriminating character voices. Furthermore, existing parameters may not be enough for this purpose, so more fine-grained acoustic parameters shall be found to effectively discriminate character voices in the further study. Moreover, the results of statistical tests are consistent with the MDS. Characters with significant acoustic differences have greater distances in MDS, while those with fewer acoustic differences have shorter distances in MDS. M02 (Prince) and M05 (Minister) have similar character voices acoustically, and so do M01 (Monarch) and M05 (Minister). Character voices of M01 (Monarch), M02 (Prince), M03 (Physician), and M04 (Eunuch) have apparent differences between each pair.

Third, the findings of this study can also provide some guidance for the practice of voice acting. When dubbing, the voice actors need to quickly get in the scene, and to become one with their characters, which places high demands on the voice actors. In addition to experiencing the emotions of the characters and dubbing immersively, voice actors can improve their extent of fitting with the character in three ways. Firstly, they can exaggerate and typify some acoustic characteristics, such as pitch, to emphasize the character's image. Secondly, they can establish a relationship between a type of character image and a specific type of voice quality, linking character types to voice quality, in order to match the audience's expectations in perception, and to improve the expressiveness of the dubbed voices. Thirdly, they may judge the range of roles suitable for their physiological conditions, in order to improve the match between voice acting and character image, and also to reduce the difficulty of dubbing.

Data availability statement

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

Author contributions

WL, XZ, and CL conceived and designed the study, participated in the statistical analysis, interpreted the data, and wrote the first draft of the manuscript. XZ collected the data. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by National Social Science Foundation (No. 22CYY022) and Future Plan for Young Scholars of Shandong University.

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.

References

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

ANSI, United States of America Standards Institute, and Acoustical Society of America (eds.). (1960). *Acoustical Terminology: ANSI S1.1-1960*. New York, NY: American National Standards Institute.

Baken, R. J., and Orlikoff, R. F. (2000). *Clinical Measurement of Speech and Voice, 2nd Edn*. San Diego, CA: Singular Thomson Learning.

Belin, P., Fecteau, S., and Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends Cogn. Sci.* 8, 129–135. doi: 10.1016/j.tics.2004.01.008

Blankenship, B. (2002). The timing of nonmodal phonation in vowels. *J. Phon.* 30, 163–191. doi: 10.1006/jpho.2001.0155

Boersma, P., and Weenink, D. (2021). *Praat: Doing Phonetics by Computer. Version 6.1.37*. Available online at: http://www.praat.org/ (accessed January 23, 2022).

Bonastre, J. F., Bimbot, F., Boë, L. J., Campbell, J. P., Reynolds, D. A., and Magrin-Chagnolleau, I. (2003). "Person authentication by voice: a need for caution," in *8th European Conference on Speech Communication and Technology* (Geneva), 33–36.

Callier, P. (2010). "Voice quality, rhythm and valorized femininities," in *Poster Session Presented at Sociolinguistics Symposium* (Southampton), 18.

Callier, P. (2012). "Variation in phonation type. Distributions and meanings in a mass-mediated context," in *Poster Session Presented at New Ways of Analyzing Variation* (Bloomington), 41.

Cao, H. L., and Kong, J. P. (2016). Correlations between vocal tract parameters and body heights in adult humans. *J. Tsinghua Univ.* 56, 1184–1189. doi: 10.16511/j.cnki.qhdxxb.2016.26.009

Cao, J. F. (2002). The relationship between tone and intonation in Mandarin Chinese. *Stud. Chin. Lang.* 3, 195–202+286.

Cartei, V., Garnham, A., Oakhill, J., Banerjee, R., Roberts, L., and Reby, D. (2019). Children can control the expression of masculinity and femininity through the voice. *R. Soc. Open Sci.* 6, 190656. doi: 10.1098/rsos.190656

Cartei, V., and Reby, D. (2012). Acting gay: male actors shift the frequency components of their voices towards female values when playing homosexual characters. *J. Nonverbal Behav.* 36, 79–93. doi: 10.1007/s10919-011-0123-4

Chao, Y. R. (1968). *A Grammar of Spoken Chinese*. Berkeley, CA: University of California Press.

Chen, W. R., Whalen, D. H., and Tiede, M. K. (2021). A dual mechanism for intrinsic F0. *J. Phon.* 87, 101063. doi: 10.1016/j.wocn.2021.101063

Cox, M. A., and Cox, T. F. (2008). "Multidimensional scaling," in *Handbook of Data Visualization*, eds C. H. Chen, W. K. Härdle, and A. Unwin (Berlin, Heidelberg: Springer), 315–347.

Cukor, G. (1964). *May Fair Lady*. Burbank, CA: Warner Bros.

Dave, R. (1968). A formant analysis of the clear, nasalized, and murmured vowels in Gujarati. *Ann. Rep. Inst. Phonet. Univ. Copenh.* 2, 119–132. doi: 10.7146/aripuc.v2i.130678

DiCanio, C. T. (2009). The phonetics of register in Takhian Thong Chong. *J. Int. Phon. Assoc.* 39, 162–188. doi: 10.1017/S0025100309003879

Ding, L. (2005). The composition and function of intonation in Putonghau. *J. Shaanxi Univ. Technol.* 3, 59–63.

Dolar, M. (2006). *A Voice and Nothing More*. London: MIT Press.

Donen, S., Kelly, G., and Freed, A. (1952). *Singing in the Rain*. New York, NY: MGM/Pathe Home Video.

Eriksson, A. (2005). "Tutorial on forensic speech science," in *Proceedings of the 9th European Conference on Speech Communication and Technology* (Lisbon), 4–8.

Esling, J. H. (1978). The identification of features of voice quality in social groups. *J. Int. Phon. Assoc.* 8, 18–23. doi: 10.1017/S0025100300001699

Esling, J. H., Moisik, S. R., Benner, A., and Crevier-Buchman, L. (2019). *Voice Quality The Laryngeal Articulator Model*. London: Cambridge University Press.

Esposito, C. M. (2006). *The Effects of Linguistic Experience on the Perception of Phonation* (Dissertation's thesis). University of California, Los Angeles, CA, United States.

Fischer-Jørgensen, E. (1967). Phonetic analysis of breathy (murmured) vowels in Gujarati. *Ann. Rep. Inst. Phonet. Univ. Copenh.* 2, 35–85. doi: 10.7146/aripuc.v2i.130674

Fónagy, I. (2001). *Languages within Language. An Evolutive Approach*. Amsterdam; Philadelphia, PA: John Benjamins.

Fritzell, B., Hallén, O., and Sundberg, J. (1974). Evaluation of Teflon injection procedures for paralytic dysphonia. *Folia Phoniatr. Logopaed.* 26, 414–421. doi: 10.1159/000263803

Gao, H., Su, G. C., and Chen, S. G. (2005). Acoustic features analysis of mandarin speech under various emotional status. *Space Med. Med. Eng.* 5, 350–354.

Gao, P. (2013). *Study of Emotion in Film and Television Dubbing* (Master's thesis). Henan University, Kaifeng, China.

Grawunder, S., and Winter, B. (2010). "Acoustic correlates of politeness: prosodic and voice quality measures in polite and informal speech of Korean and German speakers," in *International Conference for Speech Prosody 5* (Chicago, IL), 10–14.

Gregory, N. D., Chandran, S., Lurie, D., and Sataloff, R. T. (2012). Voice disorders in the elderly. *J. Voice* 26, 254–258. doi: 10.1016/j.jvoice.2010.10.024

Hansen, J. H., and Hasan, T. (2015). Speaker recognition by machines and humans: a tutorial review. *Inst. Elect. Electron. Eng. Signal Process. Mag.* 32, 74–99. doi: 10.1109/MSP.2015.2462851

Hartl, D. M., Hans, S., Vaissière, J., and Brasnu, D. F. (2003). Objective acoustic and aerodynamic measures of breathiness in paralytic dysphonia. *Eur. Arch. Otorhinolaryngol.* 260, 175–182. doi: 10.1007/s00405-002-0542-2

Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *J. Speech Lang. Hear. Res.* 37, 769–778. doi: 10.1044/jshr.3704.769

Hillenbrand, J., and Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *J. Speech Lang. Hear. Res.* 39, 311–321. doi: 10.1044/jshr.3902.311

Hinton, L., Nichols, J., and Ohala, J. J. (eds.). (1994). *Sound Symbolism*. New York, NY: Cambridge University Press.

Honorof, D. N., and Whalen, D. H. (2010). Identification of speaker sex from one vowel across a range of fundamental frequencies. *J. Acoust. Soc. Am.* 128, 3095–3104. doi: 10.1121/1.3488347

Hornibrook, J., Ormond, T., and Maclagan, M. (2018). Creaky voice or extreme vocal fry in young women. *N. Zeal. Med. J.* 131, 36–40.

Iseli, M., Shue, Y. L., and Alwan, A. (2007). Age, sex, and vowel dependencies of acoustic measures related to the voice source. *J. Acoust. Soc. Am.* 121, 2283–2295. doi: 10.1121/1.2697522

Jia, H. M. (2017). *[[Inline Image]][[Inline Image]][[Inline Image]]The Analysis of Chinese Emotional Intonation Based on the Discourse of Chinese Movies and Televisions* (Master's thesis). Jinan University, Guangzhou, China.

Kawahara, S., Noto, A., and Kumagai, G. (2018). Sound symbolic patterns in Pokémon names. *Phonetica* 75, 219–244. doi: 10.1159/000484938

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

Klatt, D. H., and Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.* 87, 820–857. doi: 10.1121/1.398894

Kong, J. P. (2015). *A Basic Course in Experimental Phonetics*. Beijing: Peking University Press.

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., and Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens* 1, e009. doi: 10.3989/loquens.2014.009

Kreiman, J., and Sidtis, D. (2011). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. Oxford: Wiley-Blackwell.

Kreiman, J., Vanlancker-Sidtis, D., and Gerratt, B. R. (2005). "Perception of voice quality," in *The Handbook of Speech Perception*, eds D. B. Pisoni and R. E. Remez (Oxford: Blackwell Publishing), 338-361.

Kuang, J. (2013). The tonal space of contrastive five level tones. *Phonetica* 70, 1–23. doi: 10.1159/000353853

Ladefoged, P. (2003). *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques*. Oxford: Wiley-Blackwell.

Laver, J. (1980). The phonetic description of voice quality. *Camb. Stud. Linguist. London* 31, 1–186.

Leino, T. (2009). Long-term average spectrum in screening of voice quality in speech: untrained male university students. *J. Voice* 23, 671–676. doi: 10.1016/j.jvoice.2008.03.008

Li, K. P., Hughes, G.W., and House, A. S. (1969). Correlation characteristics and dimensionality of speech spectra. *J. Acoust. Soc. Am.* 46, 1019–1025. doi: 10.1121/1.1911794

Li, L. H. (2007). On the Creation of Dubbing. *Contemp. Cinema* 6, 95–98.

Lippi-Green, R. (2012). *English with an Accent: Language, Ideology, and Discrimination in the United States*. London: Routledge.

Liu, F. (1924). *Record of Experiments on the Four Tones*. Shanghai: Qunyi Press.

Liu, W. (2019). An acoustic and perceptual study of the five level tones in Hmu (Xinzhai Variety). *Chin. J.* 1, 79–87. doi: 10.21437/Interspeech.2020-0056

Liu, W. (2021). Physiological and physical basis of phonation types and its linguistic value. *Essays Linguist.* 1, 204–233.

Liu, W., Lin, Y. J., Yang, Z., and Kong, J. P. (2020). Hmu (Xinzhai variety). *J. Int. Phon. Assoc.* 50, 240–257. doi: 10.1017/S0025100318000336

Liu, W. N. (1994). *Film and Television Acoustics*. Nanjing: Nanjing University Press.

Mendoza, E., Valencia, N., Muñoz, J., and Trujillo, H. (1996). Differences in voice quality between men and women: use of the long-term average spectrum (LTAS). *J. Voice* 10, 59–66. doi: 10.1016/S0892-1997(96)80019-1

Miramont, J. M., Restrepo, J. F., Codino, J., Jackson-Menaldi, C., and Schlotthauer, G. (2020). Voice signal typing using a pattern recognition approach. *J. Voice* 36, 34–42. doi: 10.1016/j.jvoice.2020.03.006

Mixdorff, H., Niebuhr, O., and Hönemann, A. (2018). "Model-based prosodic analysis of charismatic speech," in *Proceedings of 9th International Conference of Speech Prosody* (Poznan), 814–818.

Moisik, S. R. (2012). Harsh voice quality and its association with blackness in popular American media. *Phonetica* 69, 193–215. doi: 10.1159/000351059

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *Am. Nat.* 111, 855–869. doi: 10.1086/283219

Murry, T., and Singh, S. (1980). Multidimensional analysis of male and female voices. *J. Acoust. Soc. Am.* 68, 1294–1300. doi: 10.1121/1.385122

Noble, L., and Xu, Y. (2011). "Friendly speech and happy speech-are they the same?," in *17th International Congress of Phonetic Sciences* (Hong Kong), 1502–1505.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica* 41, 1–16. doi: 10.1159/000261706

Ohala, J. J. (1994). "The frequency code underlies the sound-symbolic use of voice pitch," in *Sound Symbolism*, eds L. Hinton, J. Nichols, and J. J. Ohala (New York, NY: Cambridge University Press), 325–347.

Pépiot, E. (2014). Male and female speech: a study of mean F0, F0 range, phonation type and speech rate in Parisian French and American English speakers. *In Speech Prosody* 7, 305–309. doi: 10.21437/SpeechProsody.2014-49

Pittam, J. (1987). The long-term spectral measurement of voice quality as a social and personality marker: a review. *Lang. Speech* 30, 1–12. doi: 10.1177/002383098703000101

Podesva, R. J. (2007). Phonation type as a stylistic variable: the use of falsetto in constructing a persona. *J. Sociolinguist.* 11, 478–504. doi: 10.1111/j.1467-9841.2007.00334.x

Podesva, R. J., and Callier, P. (2015). Voice quality and identity. *Annu. Rev. Appl. Linguist.* 35, 173–194. doi: 10.1017/S0267190514000270

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

Rallabandi, S. S., Naderi, B., and Möller, S. (2021). "Identifying the vocal cues of likeability, friendliness and skilfulness in synthetic speech," in *Proceedings of 11th International Speech Communication Association Speech Synthesis Workshop* (Budapest), 1–6.

Sapir, E. (1929). A study in phonetic symbolism. *J. Exp. Psychol.* 12, 225–239. doi: 10.1037/h0070931

Shaw, B., and Fisher, J. (1963). *Pygmalion*. Melbourne, VIC: Royal Victorian Institute for the Blind Educational Centre.

Shue, Y. L., Keating, P., Vicenik, C., and Yu, K. (2009). VoiceSauce: a program for voice analysis. *J. Acoust. Soc. Am.* 126, 2221. doi: 10.1121/1.3248865

Signorello, R., Derrico, F., Poggi, I., and Demolin, D. (2012). "How charisma is perceived from speech: a multidimensional approach," in 2012 *International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing* (Amsterdam), 435–440.

Sjölander, K., and Beskow, J. (2000). "Wavesurfer-An open source speech tool," in *6th International Conference on Spoken Language Processing* (Beijing), 464–467.

Starr, R. L. (2015). Sweet voice: the role of voice quality in a Japanese feminine style. *Lang. Soc.* 44, 1–34. doi: 10.1017/S0047404514000724

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

Tavi, L., Alumäe, T., and Werner, S. (2019). "Recognition of creaky voice from emergency calls," in *20th Interspeech Conference* (Graz), 1990-1994.

Teshigawara, M. (2003). *Voices in Japanese Animation: A Phonetic Study of Vocal Stereotypes of Heroes and Villains in Japanese Culture* (Dissertation's thesis). University of Victoria, Victoria, BC, Canada.

Tong, K. H., and Moisik, S. R. (2021). Detecting protagonists and antagonists in the voice quality of American cartoon characters: a quantitative LTAS-based analysis. *Phonetica* 78, 345–384. doi: 10.1515/phon-2021-2009

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika* 17, 401–419. doi: 10.1007/BF02288916

Van Lancker, D., Kreiman, J., and Emmorey, K. (1985). Familiar voice recognition: patterns and parameters part I: recognition of backward voices. *J. Phon.* 13, 19–38. doi: 10.1016/S0095-4470(19)30723-5

Weiss, B., Trouvain, J., Barkat-Defradas, M., and Ohala, J. J. (2021). *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers*. Singapore: Springer.

Winkler, R., and Sendlmeier, W. (2006). EGG open quotient in aging voices—Changes with increasing chronological age and its perception. *Logoped. Phoniatr. Vocol.* 31, 51–56. doi: 10.1080/14015430500445534

Wu, Q., Liu, Y., Li, D., Leng, H., Iqbal, Z., and Jiang, Z. (2021). Understanding one's character through the voice: dimensions of personality perception from Chinese greeting word "Ni Hao". *J. Soc. Psychol.* 161, 653–663. doi: 10.1080/00224545.2020.1856026

Yang, L. (2021). Artistic creation and diversified expression of film and television dubbing. *Media Forum* 4, 113–114.

Yang, Z., Huynh, J., Tabata, R., Cestero, N., Aharoni, T., and Hirschberg, J. (2020). "What makes a speaker charismatic? Producing and perceiving charismatic speech," in *Proceedings of 10th International Conference on Speech Prosody* (Tokyo), 685–689.

Yang, Z. S. (2021). On the ability of artistic creation in film and television dubbing in the era of intelligent media. *J. Commun.* 22, 101–102.

Yuasa, I. P. (2010). Creaky voice: a new feminine voice quality for young urban-oriented upwardly mobile American women?. *Am. Speech* 85, 315–337. doi: 10.1215/00031283-2010-018

Zhang, F., Huang, L., Chao, X., Shi, Y., and Qu, C. Y. (2021). Acoustic characteristics of vocal emotion sound of hearing-impaired children and normal hearing children aged 3~5. *J. Audiol. Speech Pathol.* 29, 146–150.

Zhang, P., Wang, L. H., and Liu, S. (2008). "On fundamental frequency contour synthesis and control method for Chinese speech synthesis," in *Proceedings of the 27th Chinese Control Conference* (Kunming), 3211–3214.

Zhao, Q., Huang, P., and Zhai, J. B. (2015). *Designing SoundforAnimation.* Beijing: Renmin University of China Press.

frontiers | Frontiers in Communication

TYPE Brief Research Report
PUBLISHED 17 April 2023
DOI 10.3389/fcomm.2023.1089577

EDITED BY
Plinio Almeida Barbosa,
State University of Campinas, Brazil

REVIEWED BY
Luma Miranda,
Eötvös Loránd University, Hungary
Jacek Kudera,
University of Trier, Germany

*CORRESPONDENCE
Hille Pajupuu
✉ hille.pajupuu@eki.ee

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

RECEIVED 04 November 2022
ACCEPTED 28 March 2023
PUBLISHED 17 April 2023

CITATION
Pajupuu H, Pajupuu J, Altrov R and Kiissel I (2023) Robot reads ads: likability of calm and energetic audio advertising styles transferred to synthesized voices.
Front. Commun. 8:1089577.
doi: 10.3389/fcomm.2023.1089577

Robot reads ads: likability of calm and energetic audio advertising styles transferred to synthesized voices

Hille Pajupuu*, Jaan Pajupuu, Rene Altrov and Indrek Kiissel

Department of Speech Research and Technology, Institute of the Estonian Language, Tallinn, Estonia

The increasing prevalence of audio advertising has provided a challenge to find out more about voices and performance styles used in advertisements. In this study, we were interested in the listeners' preferences when a synthesizer performs the advertisements. As training an advertisement style synthesizer requires big corpora, the creation of which is time-consuming and expensive, we have chosen to use less resource-intensive style transfer on already extant synthesized voices trained on neutral speech. We used a corpus of advertisements created out of 120 male and 120 female voices reading one text in both an energetic and calm advertisement style, the styles most commonly provided by advertising agencies, to train four style transfer models: energetic and calm for both male and female voices. These were used to convert two synthesized female and two male voices that had been created using a Merlin-based speech synthesizer for Estonian. Each converted voice performed three short advertisements. Adult listeners rated the likability of the performances on a 7-point Likert scale. The results showed that the calm performance style was overwhelmingly preferred. We also ascertained the acoustic features of the calm and energetic performances using the open-source toolkit openSMILE to calculate the 88 parameters of the extended Geneva Minimalistic Acoustic Parameter Set. The calm style differed from the energetic in acoustic features that are related to a lower, quieter, and more sonorous voice and a more neutral speaking style. Considering the difference in style ratings, it is worth taking the target audiences' style preferences into account.

KEYWORDS

speech styles, advertisements, synthesized speech, style transfer, acoustic features, GeMAPS

1. Introduction

Every day we are exposed to advertisements on the radio, Internet, and in urban open spaces. To present their products, businesses select suitable voices from voicebanks of media companies, which offer recordings of diverse speakers presenting various speech styles. Although advertisements intended for audio media are widely produced, there are few studies on voices used in advertising and this leads to voices being chosen by intuition (Westermann, 2008; Rodero and Larrea, 2021). More consideration given to the choice of voice might increase the effectiveness of an advertisement. Without such consideration, there is the risk of making a choice, like having a celebrity as the spokesperson, who may draw all the attention onto themselves and the audience remembers them instead of the product (Kuvita and Karlíček, 2014; Erfgen et al., 2015; Grigaliunaite and Pileliene, 2015). The effectiveness of an advertisement may also be affected by intracultural experience.

Desmarais and Vignolles (2019) have shown that vocal preferences are not universal, but are influenced by the dominant sales strategy of the culture. If the culture is dominated by a hard sell strategy, which aims to get the client to make a purchase quickly, the commercial messages are direct, explicit, rational, and based on information. The style of performance in this case is rapid, emphatic, and loud. Male voices used in such advertisements are masculine, dramatic, aggressive, or very enthusiastic; female voices do not emphasize femininity and use a happy or casual tone of voice. In the case of a soft sell strategy, where the client is influenced emotionally into making the purchase, the advertising voices are seductive and softer, the speech is slower, and the intonation is kept flat. Consumers prefer voices and advertising styles which are familiar in their cultural environment (Desmarais, 2000; Desmarais and Vignolles, 2019). Rodero et al. (2013) and Martín-Santana et al. (2015, 2017) have noted the overuse of male voices in advertisements in the belief that deep male voices sound more convincing, believable, and persuasive. Their studies have shown that the gender of the spokesperson had no impact on the efficacy of advertisements of non-gender imaged products. However, in the selection of the spokesperson the likability of the spokesperson's voice should be taken into account, which in turn is impacted by both culture and speech style (Altrov et al., 2018; Baus et al., 2019; Pajupuu et al., 2019; Weiss et al., 2021).

The speech style of advertisements is clearly recognizable due to sounding unnatural in both speech tempo and its exaggerated way of speaking with strong emphases (Chattopadhyay et al., 2003; Rodero, 2020; Rodero and Potter, 2021). Listeners have learned to automatically associate this speech style with advertisements and in order to resist the influence of the advert, have learned to tune it out (Michelon et al., 2020). To make advertisements more acceptable to the listener, it is important to choose a suitable voice for the specific product or goal, but to also consider intracultural preferences (Desmarais and Vignolles, 2019).

Predictions forecast an increase in ad spending on both traditional radio advertisements as well as digital audio advertisements (Statista, 2022). Therefore, the need for effective and engaging audio advertising remains. Studies have shown that emotions fostered in listeners by speech features influence their willingness to buy, which is why the importance of choice of voice type cannot be underestimated (Nagano et al., 2021; Rodero and Larrea, 2021). To save time and money in ad production, it is sensible to search for alternatives. One option is to use text-to-speech synthesis instead of a human voice. Rodero (2017) has shown the effectiveness, attention, and recall of synthesized and human voices in a narrative advertising story. The study came to the conclusion that although current synthesizers are now able to produce a relatively natural and intelligible sound, the importance of expressiveness in advertisements and its absence in synthesized speech gave rise to negative ratings for advertisements performed by non-human voices. Synthesized voices are expected to exhibit the same characteristics as human voices which means the acoustic variety of synthesized voices needs to be further improved. Rodero's (2017) findings also showed gender differences. Most participants in the listening test preferred advertisements performed by male voices. There was also a crossover effect – women gave higher ratings to male voices and men lower. The preference for male voices may be explained by the wider use of them in Spanish radio advertising, leading to listeners being accustomed to such a voice environment (see Rodero et al., 2013).

The speaking style of synthesized speech results from the type of speech data used for training the text-to-speech systems. While there are large-scale neutral speech training corpora available for synthesizing neutral speech, collecting or recording similar quantities of expressive speech in all its variety of emotions, attitudes, and styles is time-consuming and expensive, which is why other methods are being explored (see Zhu and Xue, 2020; Schnell and Garner, 2022). Lately a solution has been sought in speech style transfer, which means transferring the style from one signal to another while preserving the latter's content and speaker's identity. A small expressive speech corpus, possibly one with multiple speakers, can be used to train a model of the desired speech style which can then be applied to synthesized neutral speech (see Gao et al., 2019; Kulkarni et al., 2021; Pan and He, 2021; Li et al., 2022; Ribeiro et al., 2022).

In our own study, we set the goal of determining which performance style of advertisement is preferred in Estonia when the advertisement is performed by a synthesized voice. For human voices, we know that Estonians prefer speech styles that do not require a loud voice (Altrov et al., 2018). It is not known what kinds of voices Estonians prefer in advertisements, nor whether the dominant sales strategy in Estonia is hard or soft sell, but the voicebanks of media companies offer voices in two styles – energetic and calm, and advertisements of both types can be heard on the radio. In our study, we used human-like female and male synthesized voices on which we applied two style models trained on a multi-speaker corpus – an energetic and calm one.

We formulated the following research questions:

Q1: Do Estonian listeners prefer synthetic voices speaking in an energetic or calm synthesized advertising style?

Q2: What speech features differentiate the calm and energetic synthesized advertising style?

2. Method

2.1. Text-to-speech synthesis

The speech models were trained using Merlin (Wu et al., 2016). A neural network based speech synthesis system developed at the Center for Speech Technology Research (CSTR), University of Edinburgh. The system relies on the Theano numerical computation library. To convert text to full-context labels, a front-end text processor developed locally in the Institute of the Estonian Language was used (Kiissel, 2022a). We converted generated parameters to signal using the WORLD vocoder (see samples Kiissel, 2022b).

Four voice models were trained on existing available Estonian corpora of emotionally neutral sentences: Female 1 (365 min, 4,701 sentences), Female 2 (449 min, 5,172 sentences), Male 1 (376 min, 4,165 sentences), and Male 2 (330 min, 2,838 sentences). For the purpose of this study, one sample sentence and three short advertising texts were synthesized with every voice.

2.2. Speech style transfer

For voice conversion we used a Tensorflow based style transfer tool (Gao, 2019; Gao et al., 2019). This is a nonparallel emotional speech conversion tool, one that does not require any paired data, transcripts, or time alignment. It enables the transfer of style-related speech characteristics, while preserving the speaker's identity and linguistic content. It is capable of producing conversions of acceptable quality from relatively small corpora. We have tested it, applying models trained on various corpora to Estonian speech (Pajupuu, 2022).

In this study, to train the style transfer models we used voice samples of 120 female actors and 120 male actors reading the same Estonian pretend-advertisement in two styles – calm and energetic – which were sourced from the database of the audio-visual post-production studio Orbital Vox Studios (wav 44.1 kHz, 16 bit, stereo, average length of advertisement 20 sec). We trained the style transfer models (CycleGAN) between energetic and calm female voices and between energetic and calm male voices. Then we applied these models to neutral style synthesized advertisements.

2.3. Listening test

A web based listening test was created. The test consisted of two parts. In the first part, the likability of two male synthesized voices (M1 and M2) and two female synthesized voices (F1 and F2) had to be rated on a 7-point Likert scale, where 1 = not likable at all ... 7 = very likable, before style transfer had been applied. For all voices, this sentence was provided for listening to: *Olen kõnerobot (nimi) ja õpin reklaame lugema.* [I am speech robot (name) and I am learning to read advertisements]. In the second part, an energetic and calm advertising style had been applied to the synthesized voices and every voice read aloud the following three advertisement fragments in both styles. The fragments were selected to avoid their Estonian semantic content from clashing with either performance style:

- *Hullud päevad kolmapäevast pühapäevani Vesiku kaubakeskuses.* [Crazy days Wednesday to Sunday at Vesiku shopping center.]
- *Sinu tegemiste õnnestumised saavad alguse heast ideest. Laenumarket – kõik tarbimislaenud ühest kohast.* [The success of your endeavors starts from a good idea. Loan market – all consumer loans from one source.]
- *Tule Diili ja vaheta vana uue vastu!* [Come to Deal and swap the old one for a new one!]

Both performers and sentences were listened to in a random order. The rating had to be assigned to the likability of the performance on a 7-point Likert scale, where 1 = not likable at all ... 7 = very likable. The instruction text was as follows:

> We are teaching speech robots to read advertisements. We need your help to find out which advertisement style you prefer as performed by a robot. Please listen to the advertisement fragments and give a rating to the performance style of the advertisement. First get to know the robots that are learning to read advertisements and let us know how you like their voices. You do not have to listen to all of the performances at once, you can save and return later. You can also change previous ratings.

The test was designed to have a duration of ∼10 min. The audio files for the listening test and the data of the listening test are included in the dataset (https://figshare.com/projects/Robot_reads_ads/151404).

2.4. Data analysis

The participating raters of the listening test were adults with tertiary education, 8 women (aged 39–56, $M = 45.4$ years, $SD = 6.4$) and 10 men (aged 36–53, $M = 45.9$ years, $SD = 7.0$). Raters participated voluntarily. The listening test caused no harm to any participant, the identity of the participants has been kept confidential, and no conflict of interest can be identified.

All scores for each rater were normalized using the formula

$$\mathbf{y} = \frac{\mathbf{x} - \mathbf{X}}{\mathbf{s}},$$

where, x is the score, X is the mean of the rater's scores, and s is the standard deviation of the rater's scores. We classified performances with scores above zero as likable, and those below zero as unlikable.

To find out the degree of agreement among the raters (inter-rater reliability), the intra-class correlation coefficient (ICC2k) was calculated using the "psych" package in R (Revelle, 2021). A Welch Two Sample t-test was used to determine whether the advertisement style affected likability ratings (R Core Team, 2022).

2.5. Acoustic analysis

For the acoustic analysis of the calm and energetic synthesized advertising styles we used the open-source toolkit openSMILE (Eyben et al., 2010, 2013). The parameters of the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) were calculated for each performance. The use of eGeMAPS is promising due to the set of acoustic parameters extracted from speech having been specially developed for paralinguistic speech analysis (Eyben et al., 2016). These 88 parameters include statistical properties (arithmetic mean, coefficient of variation, percentiles, etc.) calculated for a set of time-varying low-level acoustic features, including frequency-related, energy-/amplitude-related, spectral, and temporal features (Eyben et al., 2016). All parameters were normalized with the R function *scale*. To identify the acoustic features that distinguish the calm and energetic advertising styles, the Kruskal–Wallis Test was used, and the statistically significant parameters were ordered by the test statistic (R Core Team, 2022).

3. Results

A moderate to good reliability was found within rater measurements. The average measure ICC2k was 0.81 with a 95% confidence interval from 0.70 to 0.90 $F_{(27,513)} = 5.3$, $p < 0.0001$.

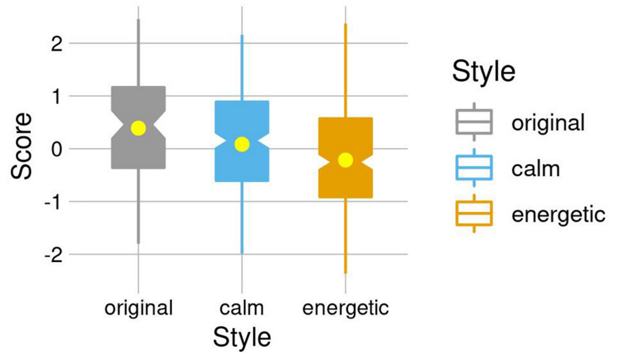

FIGURE 1
The normalized likability scores for the original synthesized speech and the style-transferred calm and energetic advertising styles.

The results of the listening test revealed that the original neutral synthesized voices were considered more likable than voices with advertising styles transferred on them [$M_{\text{original}} = 0.39$ vs. $M_{\text{calm}} = 0.09$, $t_{(130)} = 2.48$, $p = 0.014$; $M_{\text{original}} = 0.39$ vs. $M_{\text{energetic}} = -0.21$, $t_{(141)} = 4.81$, $p < 0.001$]. Of the advertising styles transferred on the original voices, the listeners overwhelmingly preferred the calm advertisement style [$M_{\text{calm}} = 0.09$ vs. $M_{\text{energetic}} = -0.21$, $t_{(474)} = 3.4$, $p < 0.001$], see Figure 1.

Looking at the synthesized voices separately, the original neutral synthesized voices ranked best to worst as follows: F2 ($M = 0.55$), M2 ($M = 0.41$), F1 ($M = 0.39$), M1 ($M = 0.20$), but there was no significant difference in their scores. The calm advertising style was rated significantly higher than the energetic style in the cases of F1 and M2 [$M_{\text{F1\_calm}} = 0.39$ vs. $M_{\text{F1\_energetic}} = -0.29$, $t_{(113)} = 2.42$, $p = 0.017$; $M_{\text{M2\_calm}} = 0.02$ vs. $M_{\text{M2\_energetic}} = -0.69$, $t_{(117)} = 4.21$, $p < 0.0001$]. There was no significant difference between the two styles for F2 and M1 [$M_{\text{F2\_calm}} = 0.03$ vs. $M_{\text{F2\_energetic}} = 0.14$, $t_{(117)} = -0.58$, $p = 0.527$; $M_{\text{M1\_calm}} = 0.19$ vs. $M_{\text{M1\_energetic}} = -0.01$, $t_{(118)} = 1.15$, $p = 0.252$], see Figure 2.

The calm and energetic advertisement performance styles were acoustically quite different: out of the 88 eGeMAPS parameters, 37 significantly differentiated these styles, of which 7 were frequency related, 10 energy/amplitude related, and 20 were spectral parameters. Tempo parameters did not belong among the differentiating features. The amount of significant parameters differentiating the styles reveals that styles might not be characterizable by a single feature, but rather a combination of features, some of which might lack a specific perceptible counterpart. The acoustic eGeMAPS parameters that differentiate the styles are presented with descriptions in the Supplementary material (see List of eGeMAPS parameter abbreviations and Supplementary Table 1).

The parameters that offer a more evident interpretation indicate that a calm advertisement style was characterized by lower pitch (lower f_0), quieter voice (lower loudness, smaller mean alpha ratio), with no abrupt changes in loudness (smaller rising and falling slopes of loudness), and a sonorous voice (less steep spectral slope). The calm advertising style was also characterized by a more neutral rather than an emotional performance (see Liu and Xu, 2014; Pralus et al., 2019; Voße et al., 2022; lower spectral flux, lower f_0, bigger rising slope f_0, and smaller falling slope of f_0, higher Hammarberg index), and the calm style is also differentiated from the energetic by parameters related to timbre (higher MFCC$_2$, MFCC$_3$, MFCC$_4$, see Nordström, 2019, p. 27). See Figure 3.

4. Discussion

In this study we attempted to answer the question of which performance style Estonians prefer in audio advertisements, if the advertisement is performed by a synthesizer. We were also interested in which acoustic features characterize the performance styles. The audio advertisements we obtained from media companies were described by them as calm or energetic based on performance style. A similar division has arisen in intercultural studies, where speech styles in advertising have been associated with the dominant sales strategy in the culture: a calmer style corresponds to the soft sell strategy while the energetic style corresponds to the hard sell strategy (see Desmarais, 2000; Desmarais and Vignolles, 2019).

We trained transferrable style models on 120 advertisements in the calm style and 120 in the energetic style, performed by female and male actors. These were then transferred on already extant voices trained on a neutral speech text-to-speech synthesizer corpus.

In the listening test, the participants rated all the original neutral synthesized voices without style transferred speech styles as more likable than the average (see Figure 2). Of the voices that had advertisement performance styles transferred on them, the calm style was considered likable and the energetic style unlikable (Figures 1, 2). Acoustically, the calm performance style was differentiated from the energetic by 37 eGeMAPS spectral, frequency, and energy/amplitude related properties, which summarily showed that compared to the energetic, the calm performance style is more neutral, the voice lower and quieter without rapid changes in loudness, but also more sonorous rather than breathy (see Figure 3, Supplementary Table 1). Using eGeMAPS makes the results from the various studies that use it easy to compare, but it is not always obvious how these parameters link to acoustically perceived speech characteristics. The use of eGeMAPS would also provide a framework for comparing advertisement styles performed by synthetic voices and humans.

The calm style preferred by listeners aligns with the soft sell sales strategy (Desmarais, 2000; Desmarais and Vignolles, 2019), but we cannot conclude that the soft sell strategy dominates Estonia. Earlier studies on the human voice conducted in Estonia have shown that for other speech styles, Estonians also prefer a quieter and more neutral voice (Altrov et al., 2018). As far as we know, different speech styles in advertisement performances have not been compared in other cultures; rather, the advertising style described in studies seems to correspond to the energetic style that studies found is both unnatural in speech tempo and excessive emphaticism (Chattopadhyay et al., 2003; Rodero, 2020; Rodero and Potter, 2021). Therefore, it is not known whether advertisements performed in the calm style may be a better fit for listeners in other cultures as well, regardless of the sales strategy.

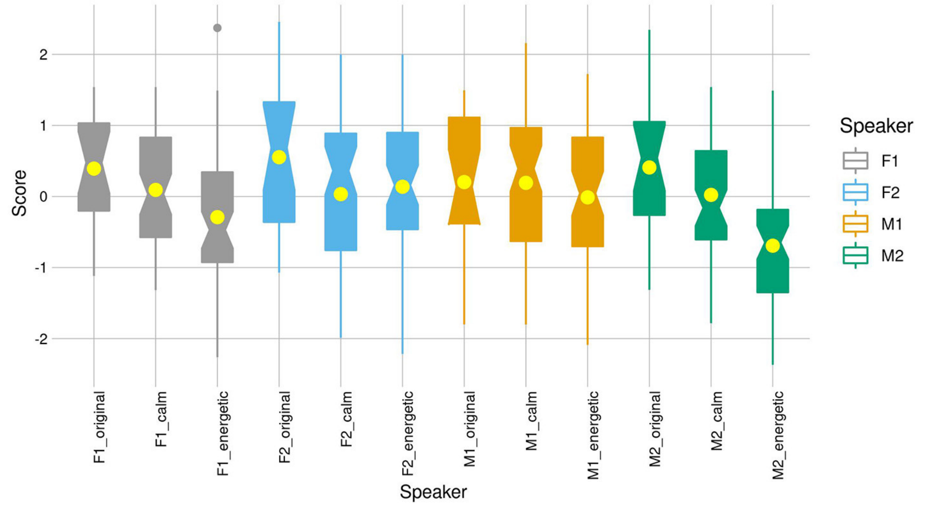

FIGURE 2
The normalized likability scores for the original synthesized speech and the style-transferred calm and energetic advertising styles for each voice separately.

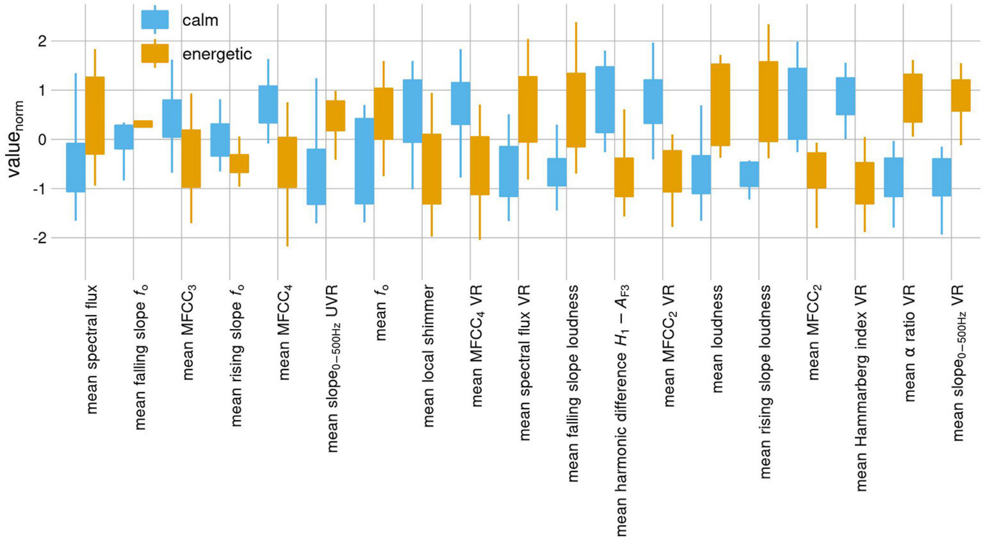

FIGURE 3
The acoustic parameters differentiating the advertising styles. *Spectral parameters:* slope 0–500 Hz = linear regression slope of the logarithmic power spectrum for 0–500 Hz region, alpha ratio = ratio of the summed energy from 50–1000 Hz, Hammarberg index = the ratio of the strongest energy peaks in the 0–2 kHz vs 2–5 kHz regions, MFCC$_{2,3,4}$ = second, third, and fourth Mel-frequency cepstral coefficient; harmonic difference H_1-A_{F3} = ratio of energy of the first f_0 harmonic (H_1) to the energy of the highest harmonic in the third formant range (A_3), spectral flux = difference of the spectra of two consecutive frames; *Energy/Amplitude related parameters:* loudness = estimate of perceived signal intensity from an auditory spectrum, rising/falling slope loudness = slope of rising/falling signal parts of loudness, shimmer = difference of the peak amplitudes of consecutive f_0 periods; *Frequency related parameters:* f_0 = logarithmic fundamental frequency on a semitone frequency scale, starting at 27.5 Hz (semitone 0), rising/falling slope f_0 = slope of rising/falling signal parts of f_0; VR, voiced regions; UVR, unvoiced regions (see descriptions in Eyben et al., 2016).

Concerning the gender of the performer of the advertisement, unlike Rodero's (2017) study on synthesized advertisements, our research did not show a gender preference in synthetic voices (see Figure 2). Whether female or male voices are preferred in advertisements is likely tied to cultural habits: if there are more male voices heard in advertisements in a culture, as Rodero et al. (2013) and Rodero (2017) have noted about Spain, then the preference is for male voices.

Our study reaffirmed the necessity of studying style preferences intraculturally. Considering the results, advertisers can make audio advertisements more palatable for the listeners of the corresponding cultural environment and thereby more effective. The impact of language alongside culture could also be further researched by studying the style preferences of listeners from different language groups within one culture.

From a technical standpoint, our study has shown that style transfer has potential and an effect can be achieved even when training style models on small corpora. Yet, the styles were rated significantly lower than the original neutral synthesized voices and the effect of style transfer was different on every voice (Figure 2). Going forward, the scale of the training corpus could be increased to see if that would improve the quality of the transfer. For a better comparison, the original voices should be trained on an equal amount of emotionally neutral sentences, so that style characteristics would not be amplified to different degrees. Furthermore, the advertisement fragments that are synthesized for rating should be longer than a sentence or two, so as to highlight the applied style model better.

5. Conclusion

Using style transfer on already existing text-to-speech synthetic voices, we discovered that Estonian listeners prefer a synthesized voice performing in a calm advertising style over an energetic one. We conclude that when using synthesizers for voicing advertisements, they could benefit from using a calm style when advertising in Estonia. Further research could show if there are differences between listener preferences of advertisement styles as performed by synthesized voices and humans.

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 below: https://figshare.com/projects/Robot_reads_ads/151404.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

HP, JP, and RA contributed to the conception of the study. HP, JP, and IK wrote sections of the manuscript. IK and JP were responsible for the speech stimuli generation. HP and RA compiled the listening test and conducted it. HP and JP did the data analysis and interpretation. All authors contributed to the manuscript revision, approved the submitted version, and had full access to all the data in the study.

Funding

This work was financed by the European Union through the European Regional Development Fund (Center of Excellence in Estonian Studies), by the National Programme for Estonian Language Technology 2018–2027, and by basic governmental financing of the Institute of the Estonian Language from the Estonian Ministry of Education and Research.

Acknowledgments

The authors thank Orbital Vox Studios, an audio-visual post-production studio, for the opportunity to use their voice database in this study. The authors wish to express their gratitude to the participants of the listening test.

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/fcomm.2023.1089577/full#supplementary-material

References

Altrov, R., Pajupuu, H., and Pajupuu, J. (2018). Phonogenre affecting voice likability. *Proc. Int. Conf. Speech Prosody* 2018, 177–181 doi: 10.21437/SpeechProsody.2018-36

Baus, C., McAleer, P., Marcoux, K., Belin, P., and Costa, A. (2019). Forming social impressions from voices in native and foreign languages. *Sci. Rep.* 9, 1–14. doi: 10.1038/s41598-018-36518-6

Chattopadhyay, A., Dahl, D. W., Ritchie, R. J. B., and Shahin, K. N. (2003). Hearing voices: the impact of announcer speech characteristics on consumer response to broadcast advertising. *J. Consum. Psychol.* 13, 198–204. doi: 10.1207/S15327663JCP1303_02

Desmarais, F. (2000). Authority versus seduction: the use of voice-overs in New Zealand and French television advertising. *Media Int. Austr. Cult. Policy* 96, 135–152. doi: 10.1177/1329878X0009600116

Desmarais, F., and Vignolles, A. (2019). Customer engagement through the vocal touchpoint: an exploratory cross-cultural study. *Adv. Adv. Res.* 2019, 67–78. doi: 10.1007/978-3-658-24878-9_6

Erfgen, C., Zenker, S., and Sattler, H. (2015). The vampire effect: when do celebrity endorsers harm brand recall? *Int. J. Res. Market.* 32, 155–163. doi: 10.1016/j.ijresmar.2014.12.002

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., Andre, E., Busso, C., et al. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Aff. Comput.* 7, 190–202. doi: 10.1109/TAFFC.2015.2457417

Eyben, F., Weninger, F., Gro,ß, F., and Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. *Proc. ACM Int. Conf. Multimedia.* 2013, 835–838. doi: 10.1145/2502081.2502224

Eyben, F., Wöllmer, M., and Schuller, B. (2010). openSMILE: The Munich versatile and fast open-source audio feature extractor. *Proc. ACM Int. Conf. Multimedia.* 2010, 1459–1462. doi: 10.1145/1873951.1874246

Gao, J. (2019). *Emotional Speech Conversion Using Nonparallel Data*. Available online at: https://github.com/bottlecapper/EmoCycleGAN (accessed November 3, 2018).

Gao, J., Chakraborty, D., Tembine, H., and Olaleye, O. (2019). Nonparallel emotional speech conversion. *Proc. Interspeech* 2019, 2858–62. doi: 10.21437/Interspeech.2019-2878

Grigaliunaite, V., and Pileliene, L. (2015). Determination of the impact of spokesperson on advertising effectiveness. *Int. J. Manage. Account. Econ.* 2, 810–822.

Kiissel, I. (2022a). *Merlinil põhinev eesti keele kõnesüntesaator [Merlin based Estonian speech synthesizer]*. Available online at: https://github.com/ikiissel/mrln_et (accessed April 3, 2023).

Kiissel, I. (2022b). *Merlinil põhinevad sünteeshääled [Merlin-based synthetic voices for Estonian]*. Available online at: https://www.eki.ee/~indrek/mrln_et/ (accessed April 3, 2023).

Kulkarni, A., Colotte, V., and Jouvet, D. (2021). "Improving transfer of expressivity for end-to-end multispeaker text-to-speech synthesis," in *Proeedings of 29th European Signal Processing Conference* (Dublin), 31–35.

Kuvita, T., and Karlíček, M. (2014). The risk of vampire effect in advertisements using celebrity endorsement. *Central Eur. Bus. Rev.* 3, 16–22. doi: 10.18267/j.cebr.89

Li, X., Song, C., Wei, X., Wu, Z., Jia, J., Meng, H., et al. (2022). Towards cross-speaker reading style transfer on audiobook dataset. *Proc. Interspeech* 2022, 5528–5532. doi: 10.21437/Interspeech.2022-11223

Liu, X., and Xu, Y. (2014). "Body size projection by voice quality in emotional speech—Evidence from Mandarin Chinese," in *Social and Linguistic Speech Prosody: Proceedings of the 7th International Conference on Speech Prosody,* Dublin, 974–977.

Martín-Santana, J. D., Muela-Molina, C., Reinares-Lara, E., and Rodríguez-Guerra, M. (2015). Effectiveness of radio spokesperson's gender, vocal pitch and accent and the use of music in radio advertising. *BRQ rly* 18, 143–160. doi: 10.1016/j.brq.2014.06.001

Martín-Santana, J. D., Reinares-Lara, E., and Reinares-Lara, P. (2017). Influence of radio spokesperson gender and vocal pitch on advertising effectiveness: the role of listener gender. *Spanish J. Market. ESIC* 21, 63–71. doi: 10.1016/j.sjme.2017.02.001

Michelon, A., Bellman, S., Faulkner, M., Cohen, J., and Bruwer, J. (2020). A new benchmark for mechanical avoidance of radio advertising. *J. Adv. Res.* 60, 407–416. doi: 10.2501/JAR-2020-007

Nagano, M., Ijima, Y., and Hiroya, S. (2021). Impact of emotional state on estimation of willingness to buy from advertising speech. *Proc. Interspeech* 2021, 2486–90. doi: 10.21437/Interspeech.2021-827

Nordström, H. (2019). *Emotional Communication in the Human Voice. [dissertation thesis].* Stockholm: Stockholm University Sweden.

Pajupuu, H., Altrov, R., and Pajupuu, J. (2019). The effects of culture on voice likability. *Trames J. Hum. Soc. Sci.* 23, 239.—257. doi: 10.3176/tr.2019.2.08

Pajupuu, J. (2022). *Samples of Speech Style Transfer for Estonian*. Available online at: https://github.com/pajupuujh/CycleGAN (accessed April 3, 2023).

Pan, S., and He, L. (2021). Cross-speaker style transfer with prosody bottleneck in neural speech synthesis. *Proc. Interspeech.*2021, 4678–4682, doi: 10.21437/Interspeech.2021-979

Pralus, A., Fornoni, L., Bouet, R., Gomot, M., Bhatara, A., Tillmann, B., et al. (2019). Emotional prosody in congenital amusia: impaired and spared processes. *Neuropsychologia* 134, 107234. doi: 10.1016/j.neuropsychologia.2019.107234

R Core Team (2022). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Evanston, IL: Northwestern University. R package version 2.1.6.

Ribeiro, M. S., Roth, J., Comini, G., Huybrechts, G., Gabry,ś, A., Lorenzo-Trueba, J., et al. (2022). "Cross-speaker style transfer for text-to-speech using data augmentation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing.* England: ICASSP, 6797–6801.

Rodero, E. (2017). Effectiveness, attention, and recall of human and artificial voices in an advertising story. Prosody influence and functions of voices. *Comput. Hum. Behav.* 77, 336.—346. doi: 10.1016/j.chb.2017.08.044

Rodero, E. (2020). Do your ads talk too fast to your audio audience? *J. Adv. Res.* 60, 337–349. doi: 10.2501/JAR-2019-038

Rodero, E., and Larrea, O. (2021). "Audio design in branding and advertising," in *Innovation in Advertising and Branding Communication*, ed. L. Mas-Manchón (New York, NY: Routledge Research in Communication Studies) 69–85.

Rodero, E., Larrea, O., and Vázquez, M. (2013). Male and female voices in commercials: Analysis of effectiveness, adequacy for the product, attention and recall. *Sex Roles* 68, 349–362. doi: 10.1007/s11199-012-0247-y

Rodero, E., and Potter, R. F. (2021). Do not sound like an announcer. The emphasis strategy in commercials. *Psychol. Market.* 38, 1417–1425. doi: 10.1002/mar.21525

Schnell, B., and Garner, P. N. (2022). Investigating a neural all pass warp in modern TTS applications. *Speech Commun.* 138, 26–37. doi: 10.1016/j.specom.2021.12.002

Statista (2022). *Digital Audio Advertising – Worldwide*. Available online at: https://www.statista.com/outlook/amo/advertising/audio-advertising/worldwide#ad-spending (accessed April 3, 2023).

Voße, J., Niebuhr, O., and Wagner, P. (2022). How to motivate with speech. Findings from acoustic phonetics and pragmatics. *Front. Commun.* 7, 910745. doi: 10.3389/fcomm.2022.910745

Weiss, B., Trouvain, J., and Burkhardt, F. (2021). "Acoustic correlates of likable speakers in the NSC database," in *Voice Attractiveness. Studies on Sexy, Likable, and Charismatic Speakers*, eds. B. Weiss, J. Trouvain, M. Barkat-Defradas, and J. J. Ohala (Singapore: Springer Verlag) 245–262.

Westermann, C. F. (2008). "Sound branding and corporate voice–strategic brand management using sound," in *Usability of Speech Dialog Systems. Listening to the Target Audience*, ed. T. Hempel (Berlin, Heidelberg: Springer), 147–155.

Wu, Z., Watts, O., and King, S. (2016). Merlin: an open source neural network speech synthesis system. *Proc. ISCA Workshop SSW* 9, 202–207. doi: 10.21437/SSW.2016-33

Zhu, X., and Xue, L. (2020). Building a controllable expressive speech synthesis system with multiple emotion strengths. *Cognit. Syst. Res.* 59, 151–159. doi: 10.1016/j.cogsys.2019.09.009

TYPE Original Research
PUBLISHED 17 April 2023
DOI 10.3389/fcomm.2023.1124513

Check for updates

OPEN ACCESS

EDITED BY
Plinio Almeida Barbosa,
State University of Campinas, Brazil

REVIEWED BY
Heliana Mello,
Federal University of Minas Gerais, Brazil
Oliver Niebuhr,
University of Southern Denmark, Denmark

*CORRESPONDENCE
Massimo Moneglia
✉ massimo.moneglia@unifi.it

†These authors have contributed equally to this work and share first authorship

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

RECEIVED 15 December 2022
ACCEPTED 16 March 2023
PUBLISHED 17 April 2023

CITATION
Cresti E and Moneglia M (2023) The role of prosody for the expression of illocutionary types. The prosodic system of questions in spoken Italian and French according to Language into Act Theory.
Front. Commun. 8:1124513.
doi: 10.3389/fcomm.2023.1124513

The role of prosody for the expression of illocutionary types. The prosodic system of questions in spoken Italian and French according to Language into Act Theory

Emanuela Cresti† and Massimo Moneglia*†

Laboratory LABLITA, Department Lettere e Filosofia, University of Florence, Florence, Italy

This article presents a corpus-based study of the correlations between prosodic contours and question speech acts in Italian and French from the perspective of the Language into Act Theory (L-AcT). A rich taxonomy of question illocutionary types is derived from two comparable corpora of informal speech taken from the C-ORAL-ROM collection and illustrated through prototypic examples. The number of questions in speech is evaluated as <10% of utterances. Despite their syntactic and accentual differences, the two languages share comparable pragmatic values conveyed by defined prosodic variations. Total questions, which can be answered with *yes* or *no*, are expected with canonical *rising* contours. Still, a good percentage shows the so-called declarative prosody (26% in FR and 36% in IT). They have been supposed to be *biased* by presupposing *the proposition's truth* and considered *Requests for confirmations* instead of genuinely *Seeking for information acts*. But presupposition, depending on intentional states, is hard to be detected, and no clear linguistic correlation was found. The corpus-based study of the prosody/pragmatics relation allows a better understanding of the system. Total questions should be framed within the larger category of *Directives aimed at the addressee's linguistic behavior*, which does not foresee *seeking information* as the only goal. When the speaker makes a *hypothesis* on what he asks—that is not an actual presupposition—he performs a *Request for confirmation*. This is true in most Total questions, and prosody complies with the canonical contour. In contrast, declarative contours correlate with corpus contexts in which the speaker does not make any *hypothesis* but puts *pressure* on the addressee (*Challenging Questions*). *Partial questions* genuinely *seek information* and comprise *Open questions,* which are verbless utterances where the addressee is requested to freely provide information on a given topic. *Tag questions*, *Double questions*, and *Alternative questions* correspond to *Illocutionary patterns* that, according to L-AcT, are composed of two pragmatic units framed within one prosodic strategy.

KEYWORDS

questions, prosody, speech acts, illocutionary force, spoken corpora, Italian, French

1. Introduction

The investigation of how prosodic information varies across the most diverse situations of real life requires access to spontaneous speech corpora and linguistic methodologies, which can bootstrap the relevant prosodic information conveying meanings from a continuous stretch of speech.

This article presents the main results achieved during a corpus-based analysis of a representative selection of the C-ORAL-ROM Corpus of Spoken Italian and French (Cresti and Moneglia, 2005) dedicated to the induction from corpora of the pragmatic and prosodic properties of questions in these two languages. Corpora mainly represents the Florentine variety and the south France variety.

The Language into Act Theory (L-AcT), which is the framework adopted for this study, foresees that prosody is the interface between the affective/pragmatic programming of the utterance (*illocutionary act*, Austin, 1962) and its linguistic fulfilling (*locutionary act*), and it is necessary to speech acts performance and interpretation (Cresti, 2000; Cresti, 2020; Moneglia, 2011).

This article focuses explicitly on the prosodic variations that characterize Yes/No questions in Italian and French, constituting a problematic case study for the pragmatics/prosody correlation. There are many Y/N questions departing from the canonical prosodic model foreseen in the two languages. They convey a so-called nuance of meaning which has been questioned in previous literature.

Indeed, speech acts taxonomies based on spoken corpora have been recently developed, enriching the tradition that origins in Austin (Austin, 1962; Searle, 1979; Searle and Vanderveken, 1985). These taxonomies provide concepts that ground new-generation pragmatic classifications in language usage.

The DART-DAMSL repertory (Weisser, 2018) sets questions under the label of *seeking information* acts and divides the category into the following types: Yes/No, Declarative Y/N, Back-channel, Reformulating, Open, Rhetorical, Alternative, and finally, Tag Questions, which are Requests for Confirmation (RC).[1] Illocutionary types are identified through prototypical sentences that substantiate each definition.

According to the DITT++[2] tag-set (Bunt et al., 2017), questions share the goal that the speaker wants to know something, which he assumes the addressee to know, and puts pressure on the latter to provide this information. The class records the following: Set Question (WH), Proposition Question (Y/N), Check Question (RC), and Choice Question (Alternative questions). An explicit distinction between Proposition questions (Y/N) and Check questions (RC) is proposed: in Y/N, the speaker wants to obtain information about the truth of a proposition, while in RC, the speaker seeks to get information about the truth of a proposition he *weakly believes* is true. This distinction is problematic for speech act interpretation since it is not explicitly reflected in any linguistic condition. Still, it is based on the speaker's state of mind, which is unavailable. Therefore, any Y/N question (*has the train left?*) is a possible candidate for both interpretations. Neither DART nor DIT++ considers prosodic cues.

In conclusion, corpus-based pragmatic taxonomies foresee a pragmatic distinction within Y/N questions among those which are *genuinely* seeking information acts, considered *unbiased*, and those characterized by the speaker's presupposition of the truth of the proposition, considered *biased*, i.e., *Requests for confirmation*. However, this distinction is neither correlated with prosodic performance nor can be predicted because of objective linguistic features.

However, a variation in the prosodic form of Y/N questions has been found in corpora; in short, both *rising* and *rising-falling* contours are widely reported in Italian and French studies. For Italian, a substantial body of research deals with the prosodic performance of Y/N questions in most Italian regional varieties (D'Imperio, 2001; Grice et al., 2005; Gili-Fivela, 2008). Within the Auto-segmental Model (Pierrehumbert and Hirschberg, 1990; Ladd, 2008), a *falling-rising* prosodic performance (L*+H, LH%) is assumed to express an overall meaning relative to the *interrogative modality* as opposed to the *declarative* one, which in Italian are not distinguished from a syntactic point of view.

Research on task-oriented speech corpora was run on the CLIPS corpus, which collects Map Task dialogues (Anderson et al., 1991) across the main Italian regional varieties. In Crocco (2006), the prosodic features and a fine-grained pragmatic analysis of Y/N questions have been annotated according to the HCRC Map task coding scheme, where Y/N questions are considered *Dialogue moves* which vary according to their informational value. However, the expression of the different nuances of meaning conveyed by prosodic contours appears fragmented, depending on various aspects (linguistic filling, interaction, and context).

In Savino (2012), only *unbiased* Y/N questions have been prosodically analyzed in all Italian varieties of CLIPS. For what concerns the Florentine variety, two contours have been observed, both characterized by a rise on the tonic syllable, respectively, followed by a low boundary tone (L+H* L-L%) and a rise boundary tone (L+H* L-H%). Crucially, no strict pragmatic correlation of this variation has been observed.

Rossano (2010) grounded his corpus-based research on spontaneous conversations by people from the Northern variety and observed a prosodic variation in Y/N questions. He found that Y/N questions (60% of all questions in his corpus) are used to implement various actions. Half of them are *requests for confirmation*, while only 26% seek information. However, also in this work, no specific correlation between the so-called social activities and prosodic variations is stated.

For what concerns the French, the volume by Grundstrom and Léon (1973) remains one of the most accurate descriptions of the prosodic varieties carried out through f0 measurements and experiments. Grundstrom provides the prosodic analysis of a significant group of utterances derived from a spontaneous conversation, developing every type of illocution. Three contours

[1] The terminology of question illocutionary acts varies in the different tag-sets. To the purpose of clarity, when reporting previous literature, we use the most widespread Y/N label for questions that can be answered negatively or positively and the WH label for questions that satisfy the interrogative variable in the answer. Within our approach, we will use the label Total question for all kinds of Y/N and the label Partial question for all kinds of WH.

[2] http://dit.uvt.nl/

have been associated with interrogative values (*raising, flat rising*, and *rising-falling)*. Validations through perception tests confirmed that the *rising* and *flat-rising* contours are the most relevant for the recognition of questions, primarily if they are associated with a short climb duration and high intensity of the last syllable, while *the rising-falling* contour is mainly paired with assertive utterances.

Fonagy and Brerard (1973), in their work on Y/N questions, found an overall *rising* contour with a moderate *climb* signaling the performance of *neutral* instances. The contour is described as a *slower climb* at the beginning, which then accelerates and remains high, forming a *convex curve* on the final syllable. However, they also noticed other types, the so-called *implicative* questions, which from a semantic perspective should imply a speaker's doubt. They still bear a *rising contour* that, after a sudden rise on the penultimate syllable, shows a significant *fall* on the last syllable. In comparison with assertive contours, they are called *declarative questions*. However, the peak of declarative questions is higher, and the interval between peak and fall is sudden in the latter and broader in the assertive contours. Declarative questions fall abruptly and deeply, while in assertive, the fall rarely reaches the base landing level.

The prosodic characters of the *interrogation* have been recently taken up inside general treatises on French prosody (Hirst and Cristo, 1998; Lacheret-Dujour and Beaugendre, 1998; Rossi, 1999; Martin, 2009). All authors agree on an overall description of a *rising* contour on the last syllable, *Contour melodique* Ci (H*H%), in Martin (2015) terminology. The variability of performance depending on regional varieties and attitudinal aspects is often highlighted.

Delais-Roussarie et al. (2015), in their masterwork, develop a semantic classification of the different types of prosodic performance of questions. They are exemplified through best examples produced in the laboratory or derived from corpora collecting different regional varieties. The typology regarding Y/N questions can be summarized in two major types: *information seeking* that is realized by a *rising* contour (H*H%), mostly very high, and *imperative* Y/N that is performed by a *rising-falling* contour (LH*L%).

Portes (2020) investigates the meaning of intonation questions with *rising-falling* contour. She wonders whether to link this specific contour to a biased trait, noticing that, for instance, in Catalan, the positively biased questions bear a *rising-falling* contour (Vanrell et al., 2013). Moreover, the author disputes the relevance of an independent contour with an f0 peak on the penultimate syllable, which would be due to uncertainty and, therefore, *negatively* biased. An experiment has been conceived where the same sentence was proposed with four prosodic contours. Participants have been requested to correlate the different performances to a previously determined set of semantic/pragmatic values. The *falling* contour was coupled with assertion and the rising contour with question, and the form of uncertainty has mainly been assimilated to the *rising* contour and interpreted as a generic question. The most controversial results are those regarding the *rising-falling* contour, which is supposed to be less grammaticalized, more expressive, and sensitive to context (Portes and Lancia, 2017; Portes, 2020).

Beyond Italian and French, relevant work on the prosodic analysis of questions in American English-spoken corpora was run by Hedberg et al. (2017). It explicitly correlates pragmatic values and prosodic performance. In American English, questions have canonical prosody: Y/Ns are *rising*, while WHs are *falling*. However, non-canonical prosodic contours have also been found, although in a small minority. They are used to signal when the speaker does not genuinely seek information, since they fail to fit with one or more standard felicity conditions of questions, i.e., *the speaker's ignorance of the answer*, *his desire for an answer*, and *his belief in the addressee's knowledge*. In short, they are *biased*. The small percentage corresponding to non-genuine Y/N questions shows a *Level* prosodic contour that can be either (L*H-L%) or (L*L-L%).

Based on corpus evidence, we will argue that Y/N questions—*Total questions* in our terminology—are not only seeking information acts centered on the speaker but must be framed within the overall Direction class and the larger pragmatic concept of *request for linguistic behavior to the addressee*. Specifically, *two* pragmatic types result in *Request for confirmation* (RC) and *Challenging questions* (CHQ). RCs are the most frequent type in the two languages and are performed by dedicated contours, named *valley* for Italian and *final rising* for French, while the *rising-falling* contour signals CHQs in the two languages.

Beyond *Total* and *Partial questions,* rich taxonomy of illocutionary types is derived from our corpus-based study when considering the pragmatics/prosody correlations. Types that have received less attention in previous literature have been recovered systematically in our corpora and will be framed in a comprehensive classification that turns out valid in both languages. *Open questions* are verbless utterances where the addressee is requested to provide information on a given topic freely. *Tag questions*, *Double questions*, and *Alternative questions*, considered *Illocutionary patterns* in L-AcT, are a sequence of two pragmatic units framed within the same prosodic strategy. All question types present differential prosodic contours that will be described here using prototypical utterances taken from corpora.

In section 2 we will briefly present the methodology, first sketching the essential features of the L-AcT framework and its annotation scheme and then the strategy of the corpus-based work. The criteria for selecting questions are stated, and the dataset derived from corpus analysis is displayed in the tables. The quantitative distribution of question types in our Italian and French corpus samplings is derived.

Section 3 is dedicated to results describing the main illocutionary types and corresponding prosodic contours in the two languages. Within Total questions, *Requests for confirmation* and *Challenging questions* are distinguished in correlation to their canonical or declarative prosodic contours. In our analysis, Partial questions, the only genuine *seeking for information* activities, are then considered and the specific category of *Open questions* is highlighted. Finally, we will give a unified picture of the prosodic and pragmatic properties of *Tag questions*, *Alternative questions* (positive and negative), and *Double questions*, which are all framed into the *Illocutionary pattern* strategy foreseen by the L-Act approach.

2. Methodology

2.1. The L-AcT framework

L-AcT is based on the relation between the accomplishment of speech acts and their prosodic performance, which is considered a necessary means for assigning a pragmatic value to speech. L-AcT develops a methodology that is adequate for studying speech in real contexts and has been conceived in Italy, independently, but in parallel, to the Macro-syntactic approach developed in France (Blanche-Benveniste, 1997). L-AcT has been applied to large-scale Romance and English corpora (Cresti and Moneglia, 2005; Mello and Raso, 2012; Cavalcante and Ramos, 2016).

The utterance is assumed as the primary reference unit for speech analysis since it is the minimal entity that can be pragmatically interpreted (Biber et al., 1999; Cresti, 2000; Izre'el et al., 2020) being the linguistic counterpart of a speech act (Austin, 1962). Since corpus data are not only composed of isolated nuclear utterances, as is often the case for the examples reported in competence-based approaches, the study of speech acts also requires the identification of the stretch of speech that accomplishes the illocutionary activity.

Prosody is crucial to speech segmentation (Izre'el et al., 2020). The L-AcT methodology is based on recognizing prosodic breaks relevant to perception (Swerts, 1997). The segmentation consists of a two-step process: the utterance is first recognized within the flow of speech, then it can be segmented into information units. Utterances are marked by perceptively relevant terminal prosodic breaks (//; ?), while information units by non-terminal ones (/)[3].

For instance, (1) is a dialogic turn of a girl who is supposed to make photocopies for some students. She is wondering whether the professor also needs a copy. The turn is a continuous flow without pause and corresponds to four utterances, each marked by a terminal break. Listening in isolation to each of them confirms their autonomy and independent illocutionary value. Given this segmentation, according to the tag-set in Cresti (2020), we can detect a *Request for confirmation*, an *Alternative question*, a *Self-answer*, and an *Ascertainment*.

(1) (a) lei / gliene serve una anch' a lei ? (b) (una) in più / o no ? (c) no // (d) lei ha questa //
[you, (do) you need one also for you? One more, or not? no. You have this one.]

The second step is the identification of the Comment unit. An utterance can be segmented into information units (Comment, Topic, Appendix, and Parenthesis), and each unit has a one-to-one correspondence with a dedicated prosodic unit (*root, prefix, suffix, parenthetical*) within an information pattern (Chafe, 1994; Cresti, 2000; Moneglia and Raso, 2014). Thus, each information function is shaped by a perceptively relevant prosodic contour (Hart et al., 1990). In L-AcT, the correspondence between the information unit and prosodic unit is compulsory, while, for instance, in the macro-syntactic approach, the alignment of macro-segment boundaries with prosodic events is ensured in most cases but is not guaranteed (Martin, 2015).

The *root* prosodic unit carries out the Comment unit, which accomplishes the illocution. It records prosodic variants corresponding to specific illocutionary types (Firenzuoli, 2003; Rocha, 2016; Cresti and Moneglia, 2018; Cresti, 2020).

The *root* unit can be composed of a *preparation,* a *nucleus*, and a *tail*. The *nucleus* is necessary and sufficient to convey the pragmatic information of the whole utterance (Cresti, 2020). It corresponds to a *contour* that can be composed of a simple movement (*rising*, *falling*, and *holding*) or several movements. Movements are aligned to the syllables of the word(s) participating in the nucleus. The point of start and end (*low*, *middle*, *high*, and *very high*) characterizes each movement, which can be *short* or *long* (with respect to the canonical vowel's length). Still, the contour can be *spread* on the sequence of syllables composing the nucleus.

The identification of the *nucleus* is made on the f0 track evaluating the perceptual relevance of the selected part of the *root unit*. Once the nucleus, as the minimal prosodic contour sufficient to perform the speech activity, is perceptively identified, each movement participating in the prosodic contour is manually annotated on the f0 track. The perceptual relevance of every single movement is then confirmed by observing whether the annotation fits the *glissando threshold* provided by WinPitch. [4]

Glissando is the rate of f0 change above which a melodic change is supposed to be perceived. The glissando threshold determines the perceptual boundary between a static pitch and a melodic variation. If the variation is less than the threshold, the perception will correspond to a static tone; if it is higher, it will be perceived as a melodic variation. The threshold was established for synthetic vowels by Rossi (1971, 1978, 1999), Hart (1976), and Martin (2022) using a semitone scale.

According to these premises, the present research is grounded by identifying the nuclear part of the Comment unit within each question. Then, the pragmatic correlations are investigated. For instance, (1a)—f0 track in Figure 1[5]—shows a Topic-Comment information pattern. Even if the Topic unit prefix is erased, the second unit, a Comment (*root*), can still be pragmatically interpreted since it bears the illocutionary force of the whole utterance (*Request for confirmation*).

In (1), while the verb phrase "*gliene serve una*" works as a prosodic *preparation,* the *nucleus* of the *root* unit shows a *falling-rising* contour, named *valley* in Cresti and Moneglia (forthcoming)[6]. It is timed on the semantic content (*anch'a lei*), which is structured in a unique prosodic word. The falling movement (over the glissando) is timed on the first tonic vowel and

[3] Their identification has been proven consistent across languages reaching a high interrater agreement (Cresti and Moneglia, 2005; Raso and Mello, 2012; Barbosa and Raso, 2018; Panunzi et al., 2020; and references therein).

[4] http://www.winpitch.com

[5] Figures report the f0 tracks calculated with Winpitch. Given that noisy signals are frequent in spontaneous speech corpora, we will present the f0 face to the first or second harmonic to verify the correctness of its calculation. The nuclear f0 movements are highlighted in the f0 tracks. One wav. file with the acoustic signals in each figure is delivered. The C-ORAL-ROM file is reported for each example.

[6] The term *valley* must be distinguished from the usage proposed, for instance, by Kohler (2004), indicating a simple *rising* contour with no reference to the necessary previous *falling* movement on the tonic vowel.

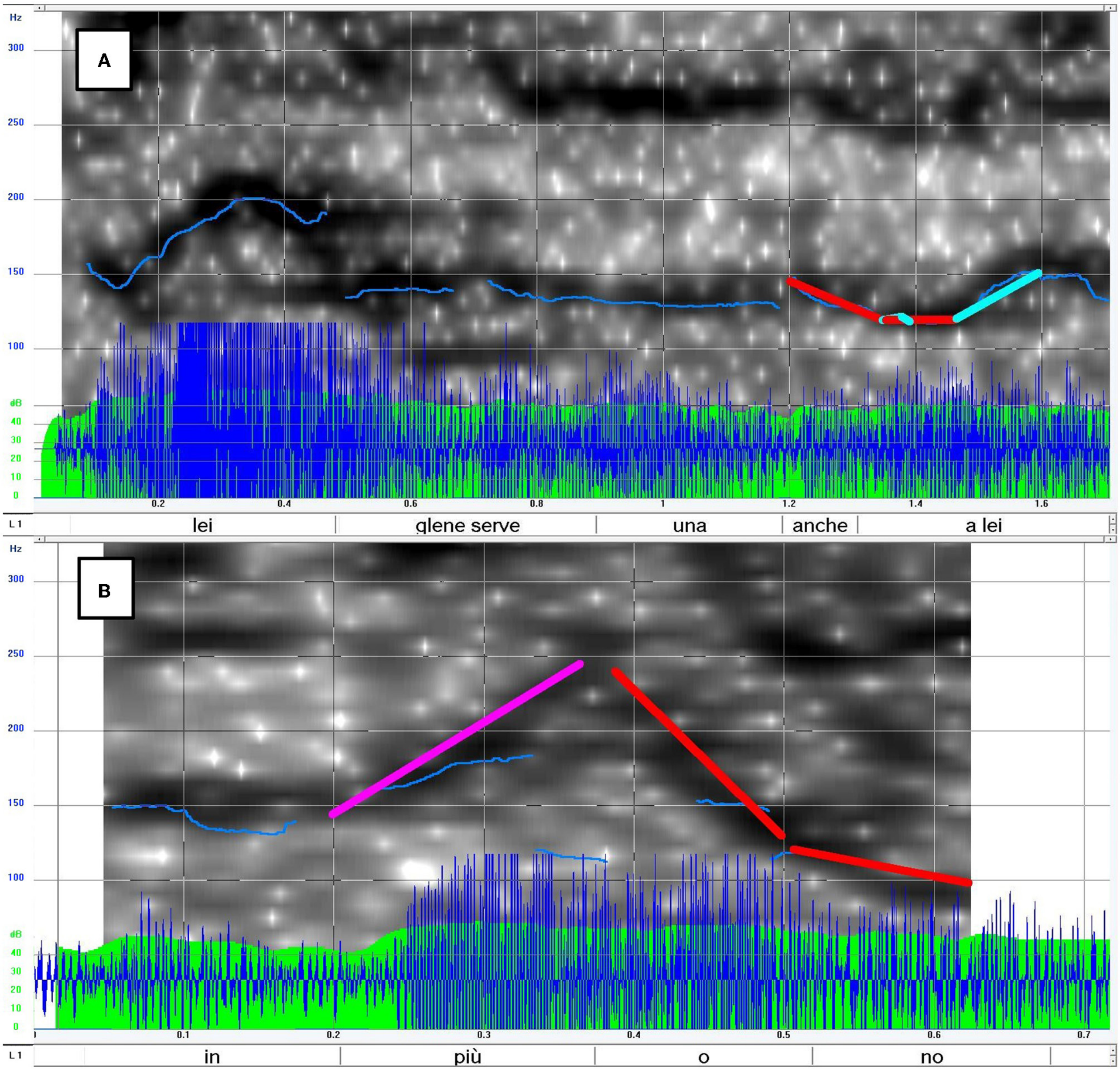

FIGURE 1
The comment and the illocutionary pattern **(A, B)** (Supplementary Audio 1.wav).

is followed by a short holding (under the glissando) and a rising movement (over the glissando) on the post-tonic diphthong.

An utterance may also correspond to a chain of Comments, called Multiple Comments (CMM), that give rise to *Illocutionary patterns*. They are conceived according to a natural rhetoric model of two or more pragmatic units (*Reinforcement, List, Comparison, etc.*), performed within a prosodic pattern (Cresti, 2000; Panunzi and Saccone, 2018). Recognizing *Illocutionary patterns* is crucial in the domain of questions since it allows gathering types of questions, such as *Alternative questions, Tag-questions,* and *Double questions,* within the same prosodic strategy.

For instance, (1b)—f0 track in Figure 1—is an *Alternative negative question* performed within one Illocutionary pattern. So, in this case, there are two CMMs and two nuclear contours.

The first is a *rising* contour (over the glissando) timed on the tonic vowel of the first CMM, and the second CMM is composed of a *falling movement* on the tonic vowel (over the glissando), followed by a *holding* movement, which we identified, as a whole, as a *level* contour (see Hedberg et al., 2017 and the following sections).

2.2. The dataset

The dataset is designed to investigate the occurrence of questions in real-life interactions, ensuring some probability of occurrence to speech act types present in the language. To this end, a highly variated sub-corpus of dialogues and

TABLE 1 The Italian data set.

	Illocutionary patterns				Total		Partial		Unclassified		Number of questions	Utterances
	db	an	ap	tag	val	moun	wh	op	ret	un		
ifamdl12 - two *young women trying to cook a Tiramisù*					7	3	13	3	4	3	33	353
Ifamdl02 - *souvenir of a family*				1	5	5	8		1	1	21	408
ifamdl20 - *gossips of two women about a friend's love story*				1	3	2	9		5		20	190
ifamcv14 – *poker game among friends*	4	1	2	4	10	12	20			7	60	458
ifamdl15 - *chat with the beautician during the depilation*			1	1	22	3	8	2		6	43	342
ifamcv01 - *showing the family Photo album to the daughter-in-low*	3	3	1	2	19	9	35	2	2	7	83	1030
ifamdl09 - *chat after a rock concert experience*	2		1		4	2	12	3		1	25	307
ifamdl19 - *driving school*			1		3	1	7	2		5	19	387
ipubcv03 - *setting the agenda for young children's assistance*				3	7			2		3	15	283
ifamdl17-*developing photographs in a dark room*	1	2	1		14	15	9	4	1	4	52	404
Total	10	6	7	12	94	52	121	18	13	37	371	4,162

multi-dialogues derived from the Italian and French C-ORAL-ROM collection has been selected to avoid repeating the same type of interactive context.[7] The contexts chosen in each language corpus are briefly reported in Tables 1, 2. The finding can provide at least some quantitative measures for the different types of questions in the corpus. This is the opposite strategy of using the Map-task dialogues, which aims at characterizing speakers' performance face to the same tasks in a controlled environment and may not represent the actual variation.

Corpus methodology based on real-life data, however, causes a loss of acoustic quality of the dataset, characterized by frequent overlapping and background noise. We accepted this bias on the f0 calculation since WinPitch ensures, in any case, the minimal conditions to verify the f0 movements displaying them face to the first or second harmonic.

The C-ORAL-ROM source provides an independent segmentation into utterances (Moneglia, 2005) that are delivered aligned to the acoustic source. The selection of questions within the dialogues undergoes the following working criteria:

- Recognition of the utterance as a question by at least two out of three mother tongue annotators.
- Contextual adequacy.
- Reaction by the addressee (answer).
- Syntactic form (if applicable).

We fixed a minimal quantitative target of up to 4000 utterances for annotating samples of each corpus. The Italian corpus reaches the target with 10 different C-ORAL-ROM files recording 4,162 utterances and 371 questions. It represents the language variety mostly spoken in Tuscany. The corpus collects data from 22 speakers, which are not balanced for gender (7 male and 15 female speakers).

The French corpus reached the target with 14 files recording 4,179 utterances, out of which we retrieved 335 questions. It mainly represents the language variety spoken in South France. The corpus collects data from 27 different speakers, not balanced for gender (13 male and 14 female speakers).[8]

It should be noted that considering an average of 50% of assertive illocutions recorded in spontaneous speech sampling (Firenzuoli, 2003), questions represent a consistent percentage given that in the Italian corpus, they record ∼9%, and in the French one, about 8%.

[7] The dataset is available from https://github.com/labiu/coralrom_questions_it-fr

[8] Although both genders are represented, the dataset is not settled for dealing with the gender characteristics found in questions (Niebuhr, 2015). Corpus data testify, however, that the correlation between prosodic contours and the speech act types are testified in both female and male speakers recorded in the corpus.

TABLE 2 The French data set.

File	Illocutionary patterns				Total		Partial		Unclassified		Number of questions	Utterances
	db	an	ap	tag	fin-rise	r/f	wh	op	ret	un		
fnatte02 - *explaining linguistic concepts in a class*	1	3	1		5	2	2			16	30	344
fpubcv02 - *students dealing with the University administration*		1		2	6	1	4			1	15	191
ffamdl01 - *chat among friends after the weekend*	3		1	1	8	1	9		3	3	29	530
fnatbu01 - *the sale of a car*	3		3	1	9	4	3	2	2	5	32	403
famdl08 - *criticism of the odd behavior of a family*		1			4	3	3	1	4	1	17	184
famdl23 - *visiting the grandmother*	2			3	6	4	5		3	8	31	303
famdl02 - *how the market of goods is organized*	3		1	2	14	6	9	1		1	37	566
fpubdl01- *discussion with workers on strike*	2		1	3	4	1	16	1	3	7	38	444
fpubdl02- *door-to-door sale*			1	1	4	2	2	2	1		13	161
ffamcv03 - *talking about skinheads in France*				1	4		4	1			10	230
ffamcv05 - two women talk *of the home décor*	1			2	9	1	10	1		2	26	203
ffamdl04 - *the life of a student who wants to become a singer*	2		2	3	12	2	5	2		1	29	278
ffamdl10- *a journey to Cuba*			1		1	1		1	1		5	199
ffamdl17 - *life and memories of a miner*				1	7	4	3				15	143
Total	17	5	11	20	93	32	75	12	17	45	327	4,179

We also counted what we can call "spurious" instances, such as *rhetorical questions*, *self-questions*, *echo questions*, questions used as *dialogic moves*, and *surprise questions* (14% in Italian and 22% in French). They are usually considered questions by mother-tongue speakers, although the criterion regarding the addressee's answer is not satisfied. According to a general system of illocutionary assignment (Sbisà and Turner, 2013), indeed, it is the addressee's behavior that guarantees the directive value of an act of question, which is an essential illocutionary feature.

Even if spurious questions are not an object of the present research, special attention must be given to *rhetorical* questions, which are presently the object of the large-scale Konstanz project "Question at the Interface".[9] In our tradition, rhetorical questions are generally characterized by a negative presupposition and are frequent in monologs and formal texts. They do not expect any answer and lack a dialogical value since the speaker addresses them to himself or a virtual audience. In conclusion, they cannot be classified as directive acts, as genuine questions should be from a pragmatic point of view.

In the contribution to German rhetorical questions by Braun et al. (2019), questions that cannot be considered seeking information—since their content is known or inferable—are classified as rhetorical, regardless of their actual directive force.

All question types we deal with are genuine since they are directive acts, accomplishing a clear request of linguistic behavior to the addressee, regardless of the semantic features of their content.

The working classification of questions into types is a two-step process. Questions are first assigned to a syntactic/semantic category, and then the correlation of each type's prosodic variation, if any, has been observed in detail. The list of types presented in Table 3 combines traditional grammatical classifications based on morpho-syntactic, semantic aspects (Serianni, 1998; Abeillé and Godard, 2021) and types emerging from corpus data.

Main question types have been gathered in three groups: (a) the Total one, which we will see comprehend two illocutionary types, i.e., Request for confirmation and Challenging questions; (b) the Partial one, which also extends to Open questions; and (c) the

[9] https://typo.uni-konstanz.de/questionsInterfaces/

TABLE 3 Question types adopted for corpus annotation.

Question type	Working definition	Italian example	French example
Total	The scope of the question is the whole proposition and must be answered by Y/N.	Vieni a cena stasera? [are you coming for dinner tonight]	Tu viens dîner ce soir ? [are you coming for dinner tonight]
Partial	The scope is an interrogative variable to be saturated in the answer.	Dove vai? [where are you going ?]	Où tu vas ? [where are you going ?]
Open	The scope is a verb-less phrase that must be freely explained in the answer.	(Bella macchina) e il prezzo ? [(nice car) and the price?]	(belle voiture) et le prix ? [(nice car) and the price?]
Alternative	Composed of two questions whose scope is the choice between them to be given in the answer.	Vieni la mattina? o il pomeriggio? [are you coming in the morning or in the afternoon?]	Tu viens le matin ? ou l'après-midi ? [are you coming in the morning or in the afternoon?]
Double	Composed of two question units hierarchically structured; the second specifies the question, determining the type of answer.	Quando parti? domani? [when do you leave? tomorrow?]	Tu pars quand ? demain ? [when do you leave? tomorrow?]
Tag	Composed of an assertive proposition and the request for its confirmation, whose scope is the proposition to be confirmed in the answer	Oggi è il tuo compleanno, no? [today is your birthday, isn't it?]	Aujourd'hui c'est ton anniversaire, n'est-ce pas ? [today is your birthday, isn't it?]

Illocutionary patterns, which correspond to all types composing two units (Tables 1, 2 summarize the quantitative values).

Regarding Italian, from the 371 questions, once the "spurious" set is eliminated, actual questions reduce to 320 instances. The French corpus records 4,179 utterances and 337 questions. Limiting the analysis to the main questions, 260 cases have been classified.

As Figure 2 shows, in Italian, Total and Partial questions record a similar percentage (39 and 38%), while in French, the percentage of Total questions (37%) is higher than that of Partial ones (26%). In Italian, Illocutionary patterns (Alternative, Double, and Tag questions) are a minority (9%) but reach 15% in French.

Relative frequency distributions are in some way comparable, and a significant difference between the two distributions is not reported (chi-square = 4.581, *p*-value = 0.205). Despite this, a strong variability in the number of occurrences for some types (e.g., in Italian, we have between 0 and 35 wh-) and the low coverage of other types (e.g., in Italian, there are only six occurrences of Alternative negative in the whole corpus) do not allow to draw a clear picture of the question type distributions across languages. Indeed, the two samplings follow a criterion of context variation to give some probability of occurrence to different question types but are not variated following a corpus design.

The following paragraphs will present the prosodic and pragmatic properties of questions through prototypic instances. Prototypes are those utterances in the dataset which best represent the link between an illocutionary type and its prosodic contour in the nuclear part of its Comment unit(s). Given the large variety of pragmatic contexts in spontaneous speech, we selected one Italian and one French utterance for each question type where the specific pragmatic activity under consideration is more evident. From the prosodic side, we choose utterances performed with sufficient phonetic accuracy, avoiding as much as possible the sandhi phenomena, voices overlapping, and preferring female voices. Listening to the acoustic source in the audio files joined to the paper, prototypes allow us to appreciate that the f0 contours identified on the f0 track are perceptively relevant and adequate to convey the foreseen illocutionary information.

3. Results

3.1. Total questions

In Italian, there is no special syntax for Total questions, which share their structure with assertions (SVO order); therefore, prosody is their distinctive marker. French grammar foresees both the unmarked word order and the following syntactic structures:

- With inversion of the personal pronoun, working as a suffix of the verb (*est-il parti?*).
- Without inversion but introduced by *est-ce que (est-ce qu'il part? est-ce que Paul part?).*

A relevant difference between the two languages concerns the status of the subject. In Italian, the null subject is grammatical (preferred in speech), while the subject is obligatory in French. The possibility to postpone the subject is a syntactic correlation of this property. In Italian, questions with the postposed stressed subject are unmarked (*parte Paolo*?; *parte lei*?), while this structure is not grammatical in French (**Part Paul?; *part elle*?).

Looking at syntax, while traditional grammatical descriptions give little space to verbless utterances, they represent a significant percentage of spontaneous speech productions. C-ORAL-ROM data estimate verbless utterances, 38% in Italian and 24% in French. In other words, utterances such as *prima di Natale*? and *la Pina*? in Italian and such as *et le prix*? and *à douze*? in French are frequent and fully adequate to perform a question. Of course, prosody is the sole linguistic index allowing their interpretation as Total questions or, as we will see in Sections 3.2.2 and 3.2.3, as Open questions.

Despite the range of syntactic possibilities available to French, among the 125 Total questions in the dataset (37%), only one utterance presents a suffixed pronoun and just seven are introduced by *est-ce-que*. The remaining part is divided between 20 verbless instances and complete utterances with unmarked word order, showing a clear preferential trend in speech. Therefore, despite

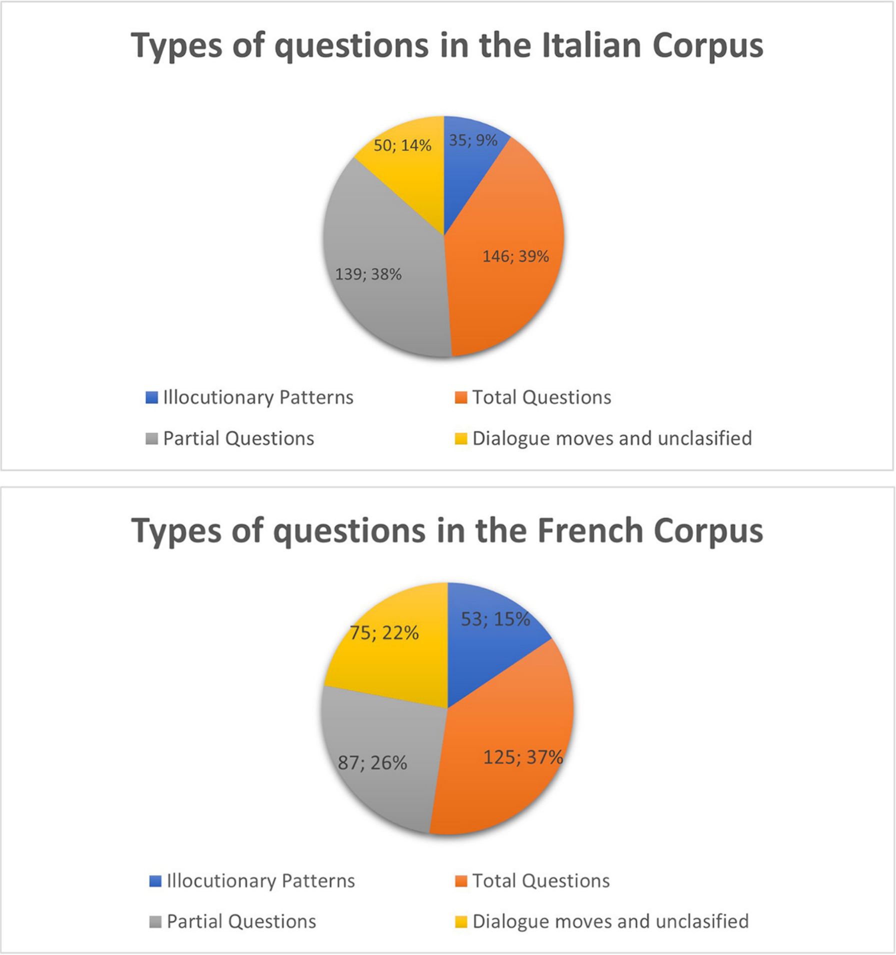

FIGURE 2
Types of questions in the Italian and French corpus.

the grammatical alternatives available, prosody is also distinctive in French in most cases.

As the pie in Figure 3 shows, French prosody is consistent with the assumption of a *rising* contour on the last tonic vowel as the standard prosodic performance of *interrogation*, which characterizes 93 instances of Total questions (74%). Corpus study also confirms that the rising contour can be performed with a higher or lower pitch (Delais-Roussarie et al., 2015; Abeillé and Godard, 2021), apparently without a nuance of meaning.

It is worth noticing, however, that in French, 32 Total questions are performed through a *rising-falling* contour, where the *falling* movement is the most relevant part (Fonagy and Brerard, 1973). As reported, this contour (*implicative*) has recently been the object of detailed investigations (Portes, 2020). Its high frequency in our spoken data (more than 25% of Total questions) is surprising since it is usually considered a marginal case.

We did not find any correlation between the prosodic and the syntactic variations in French. In short, utterances with and without inversion match both prosodic contours (*final rise* and *rising-falling*). The few utterances introduced by *est-ce-que*, too, appear in both prosodic forms.

The distribution of the syntactic structures shows specific characters in Italian. The incidence of verb less Total questions is higher than in French (36 vs. 26%). Given the null subject condition for Italian, most verbal Total questions (65%) are subject-less. The remaining part is equally divided between utterances with postponed and not postponed subjects.

The Italian prosodic performance of Total questions records two possibilities, as French. Still, instead of a simple *rising* contour, as foreseen by a large part of Italian literature, the most common type is a *falling-holding-rising* contour on the last tonic and post-tonic vowels. The contour is complex because it is composed of a *falling* part, mainly on the *tonic* vowel, and a *holding-rising* part on the *post-tonic* syllable(s). We call it *valley* contour[10]. The second prosodic contour is also complex; it corresponds to a *rising* plus a *falling* movement that, by preference, regards only the *tonic* vowel. However, the *falling* part is the most relevant, as in French.

[10] In the absence of corpus examples, it is not clear if this contour can be assimilated to the H* L-H%, as proposed by Grice et al. (2005) and evaluated as the most frequent in Florentine.

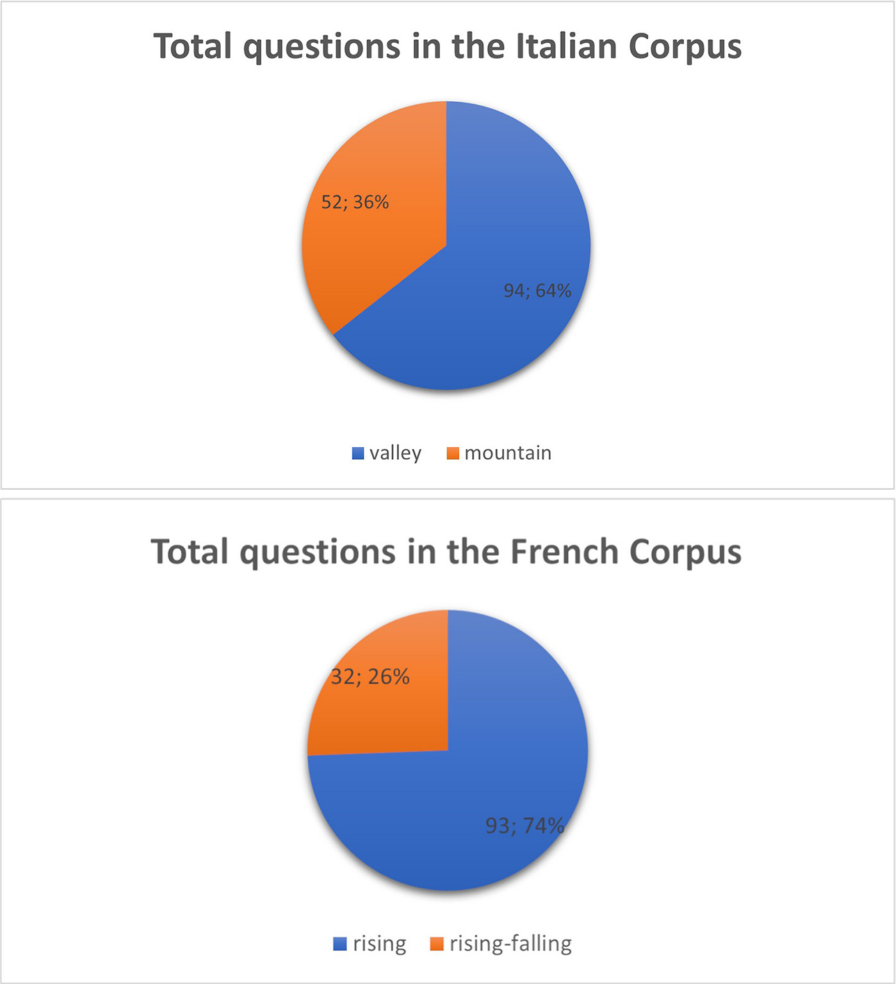

FIGURE 3
Total questions in the French and Italian corpus.

We call it *mountain* contour. All movements but the holding one are perceptually relevant according to the glissando threshold. Perceptual tests have been conducted in a laboratory to verify the distinction by mother-tongue speakers between the two contours and the change of their adequacy in different eliciting contexts (Cresti and Moneglia, forthcoming).

Figure 3 shows that the *valley* contour characterizes 94 instances of Total questions (64%) and is the most common. The *mountain* contour, however, scores a very relevant probability of occurrence (36%). Corpus-based study confirms for Italian also that the *falling-rising* contour can be performed with a higher or lower f0 range.

There is also no correlation in Italian between the possible syntactic structures (verbless, subject-less, and postponed subject) and the prosodic performance through a *valley* or a *mountain* contour. Therefore, we found each syntactic type in Italian and French in both prosodic forms. In the following paragraphs, we will explore the possibility that the prosodic variation within Total questions correlates to specific pragmatic conditions emerging from corpora.

3.1.1. Total questions as RC

In the L-AcT repertory, the question domain belongs to one of the sub-classes within the Direction Class. Question illocutionary types are defined as those activities directed to *the addressee's linguistic behavior* rather than *seeking information* acts. In other words, the definition considers that the linguistic behavior prompted by the question may not be pragmatically equivalent to giving a piece of information (Cresti, 2020).

Some intentional preparatory conditions are involved in the following questions: (a) the speaker's interest in the addressee, (b) the estimation of the background knowledge of the latter, and (c) the speaker's engagement in the linguistic interaction. Question types may vary depending on different degrees of the speaker's affective activation, in force, and type of linguistic behavior objective of the direction.

When performing a Total question, the speaker should in principle positively represent an eventuality. In so doing, he should at least assume its possible occurrence as a *hypothesis*. This seems a minimal intentional preparatory condition to keep the question as a meaningful speech act. Under this assumption, the linguistic

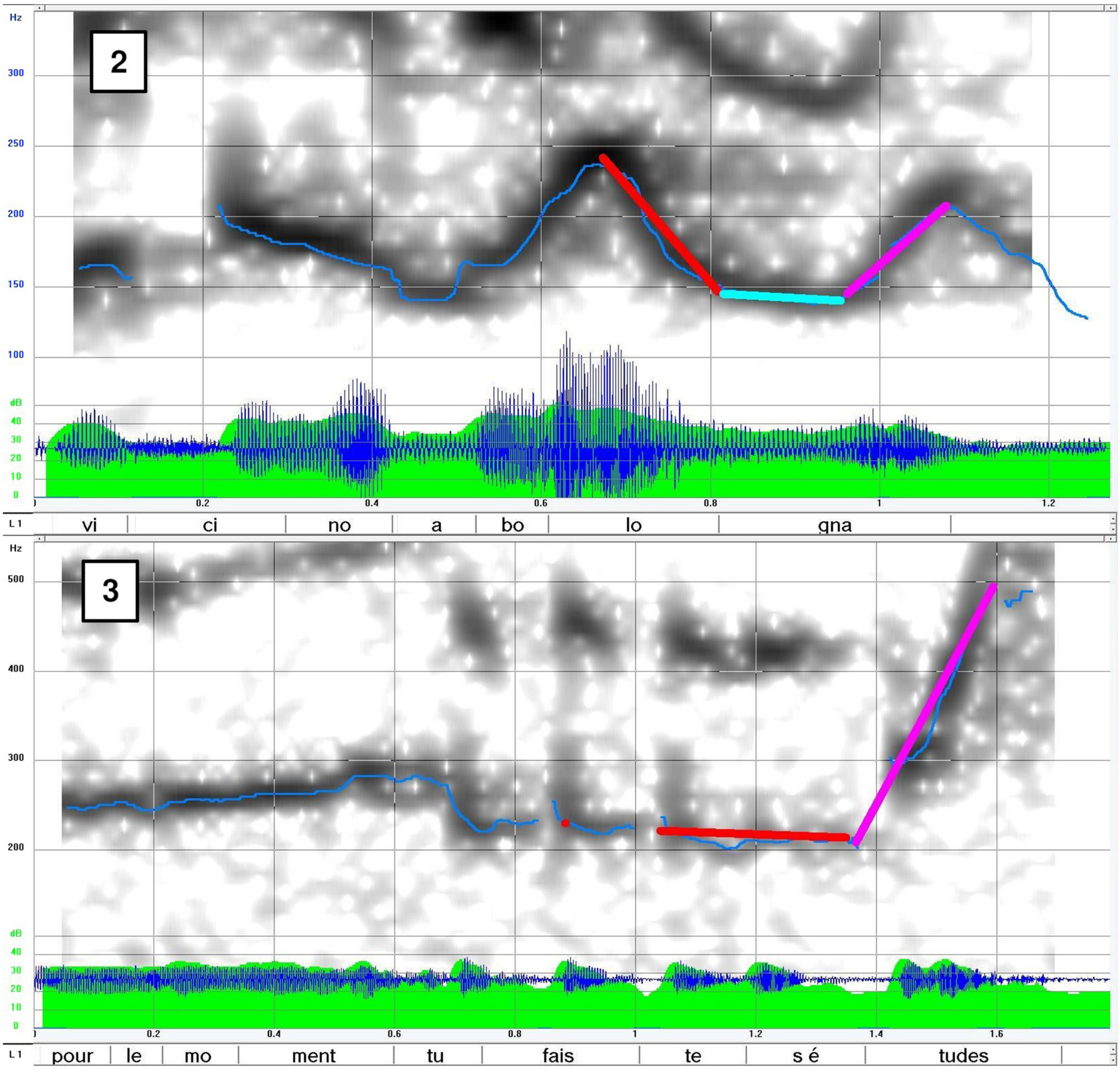

FIGURE 4
Prototypic contours of *Requests for confirmation* in Italian and French (2; 3) (Supplementary Audio 2.wav).

behavior enacted by the speaker should not be defined as a *seeking for information* act since it is instead a *request for the addressee's agreement on his hypothesis*.

The condition is independent of the speaker's presupposition of the truth of his hypothesis and must hold in all circumstances (*biased* or *unbiased*). Being a state of mind, the presupposition of the truth of the hypothesis remains underdetermined, but the consistency of the *hypothesis* condition has been verified in all the corpus contexts for questions bearing a *valley* contour in Italian and *final rising* contours in French.

For instance, in Italian example (2), ELA is wandering where the village mentioned by LIA (*Castiglion de' Pepoli*) is, and she is uncertain about the place suggested by LIA (*Sasso Marconi*). Thus, based on her rough knowledge of the region, ELA hypothesizes that the latter is near Bologna, asking for confirmation of what she supposes. The prosodic movement of (2) in Figure 4 corresponds to the *valley* contour.

(2) *ELA: [. . .] ndo' l' è / Castiglion de' Pepoli ?
*LIA: dunque / avanti di arrivare a Sasso Marconi // sull' autostrada //
*ELA: [. . .] **vicino Bologna** ? [ifamcv01]
[ELA: where is Castiglion de' Pepoli? LIA: well, before reaching Sasso Marconi. On the highway. ELA: near Bologna?]

The French example in (3) fits with the same pragmatic description. MAR, a young man, explains that he is trying to become an artist. SOP hypothesizes that MAR is doing something useful in the meanwhile and asks for confirmation. According to the description reported in previous literature, the prosodic movement of (3) in Figure 4 corresponds to a *rising* contour in its high f0 level, aligned to the last vowel, necessarily tonic in French.

(3) *MAR: j' essaierai tout / pour y <arriver / de toute façon> //
*SOP: <ouais donc> **pour le moment tu fais tes études** ? #
*MAR: ouais // [ffamdl04]
[MAR: I'll try everything, to succeed, anyway. SOP: yeah, so are you currently studying? MAR: yeah.]

There are variants of the f0 range in prosodic contours in Italian and French. However, they bring only attitudinal meanings with no changes in the illocutionary type. For instance, in (4) and (5) MAR, the same French speaker of (3) performs the two questions with a *rising* contour at a lower f0 range (still over the glissando threshold).

Both questions may be, in principle, biased. In (4), a negative presupposition is probable—it is pretty challenging to be published for a young artist—while in (5), a positive presupposition is more probable—young artists get supported by the group of pairs. The f0 of both examples, reported in Figure 5, shows a *final rise* at a lower range than in Figure 4. The choice to perform an RC at a lower f0 range may convey mitigation in the force of the request connected to a *courtesy* attitude (Moraes and Rillard, 2014).

(4) *MAR: [...] ben essayer justement d' écrire des [/] écrire des [/] des textes andeuh essayer de les envoyer / à des producteurs / andeuh voir un peu ce qu' ils en pensent quoi // [. . .]
*SOP: que oui // **que ça intéresse quelqu' un** ? [ffamdl04-161]
[MAR: well, I try to write texts, I try to send them to producers to see what they think about it. SOP: yes, is anyone interested?]
(5) *MAR: [. . .] on fait des trucs / et puis après ben / on est coincé / on a plus de [/] de pouvoir d' agir quoi // ça c' est clair // c' est un phénomène de société / ça //
*SAN: dans votre mouvement / andeuh justement andeuh **est-ce que vous avez été qu and même soutenu** ? <par les par les jeunes> //
*MAR: <ah ben oui on est soutenu hein> // [fpubdl01]
[MAR: we do some stuff, and then afterwards well, we're stuck, we have more power to act, what. That's clear. It's a phenomenon of society. SAN: in your musical movement, justly, did you get anyway support? By young people. MAR: ah well, yes, we are supported.]

In the Italian example (6), ELA, the same speaker of (2), asks for confirmation of her hypothesis that American soldiers were still in the country at the end of the second world war. The presupposition is probable as in (2), but the *valley* profile, marked in Figure 5, is performed at a much lower f0 range. Its perceptual relevance is confirmed by the glissando values of each movement (*falling* and *rising* over the glissando and *holding* under the glissando). Keeping the illocutionary value *request for confirmation*, we can still perceive that the choice manifests the speaker's *courtesy* attitude toward the addressee.

(6) *ELA: [. . .] alla fine / della guerra //
*LIA: sì / insomma / c'era sempre un po' // sì / sì // [. . .]
*ELA: ma / **che c' era sempre gli americani / <ancora>** ?
*LIA: [<] <sì> // quello sì // [ifamcv01]
[ELA: at the end of the war. LIA: yes. Well, there were still some (soldiers). Yes, yes. ELA: but there were always the Americans, yet? LIA: yes. That yes.]

The connection between attitudinal meanings and prosodic cues has been proposed to explain prosodic variations of questions found in corpus-based research beyond the Italian and French traditions. Kohler (2004), in his research on German syntactically marked questions taken from the Kiel corpus, finds a *default* correlation between the *rising* prosodic contour and the syntactic structure marked by the verb in the first position (*Total questions*) and the *falling* contour and the syntactic structure marked by wh- variables (*Partial questions*). In his proposal, the two prosodic contours bear different *attitudinal meanings*: the *rising* one expresses an orientation toward the addressee with interest and openness toward him, while the *falling* is oriented toward the speaker and asks for a response.

Dealing with the prosodic variation of syntactically marked questions (Total and Partial) in the corpus, he makes the hypothesis that during communication, the default link can be restructured, and each syntactic type can be found with the not-expected contour, because of the change in the relationship with the attitudinal meaning. Therefore, according to Kohler, the prosodic contour is a function of attitudinal meanings that apply to syntactic constructions.

The L-AcT perspective assumes that prosody is the direct manifestation of the illocutionary enactment, i.e., the prosodic contour is a function of illocution, which in turn shapes syntax. It is worth remembering that nearly 30% of utterances in spoken Romance languages are verbless (Cresti and Moneglia, 2005), and their pragmatic interpretation only results from the prosodic performance. The attitudinal meanings concern spheres of subjective (uncertainty, aggressiveness, and seductiveness) or social connotations (courtesy and social role), that can change the range of f0, speech rate, and quality of voice of a prosodic contour without altering its holistic form and the illocutionary value (Mello and Raso, 2012).

In the case just seen, the *valley* contour correlates with the illocutionary type of *Request for confirmation*. The f0 range of the contour can be at a lower level depending on *courtesy* attitude, but if the form changes, as we will see below, this is a function of the illocutionary type performed.

In conclusion, our corpus data does not verify the variation of a contour in correspondence either with a positive or negative presupposition (*bias*) (Savino, 2012; Vanrell et al., 2013; Portes, 2020) or with attitudinal nuances. On the contrary, the prosody is sensitive to the *hypothesis* made by the speaker that determines the *valley* form in Italian and the *rising* in French, performing RCs.

3.1.2. Challenging questions

The occurrence of questions in spoken French bearing a *rising-falling* movement (*implicative*) on the last tonic syllable has been noted by Fonagy and Brerard (1973) and Rossi (1999). Because of their prosodic closeness, they have been referred to as *declarative questions* or even *peremptory assertions*.

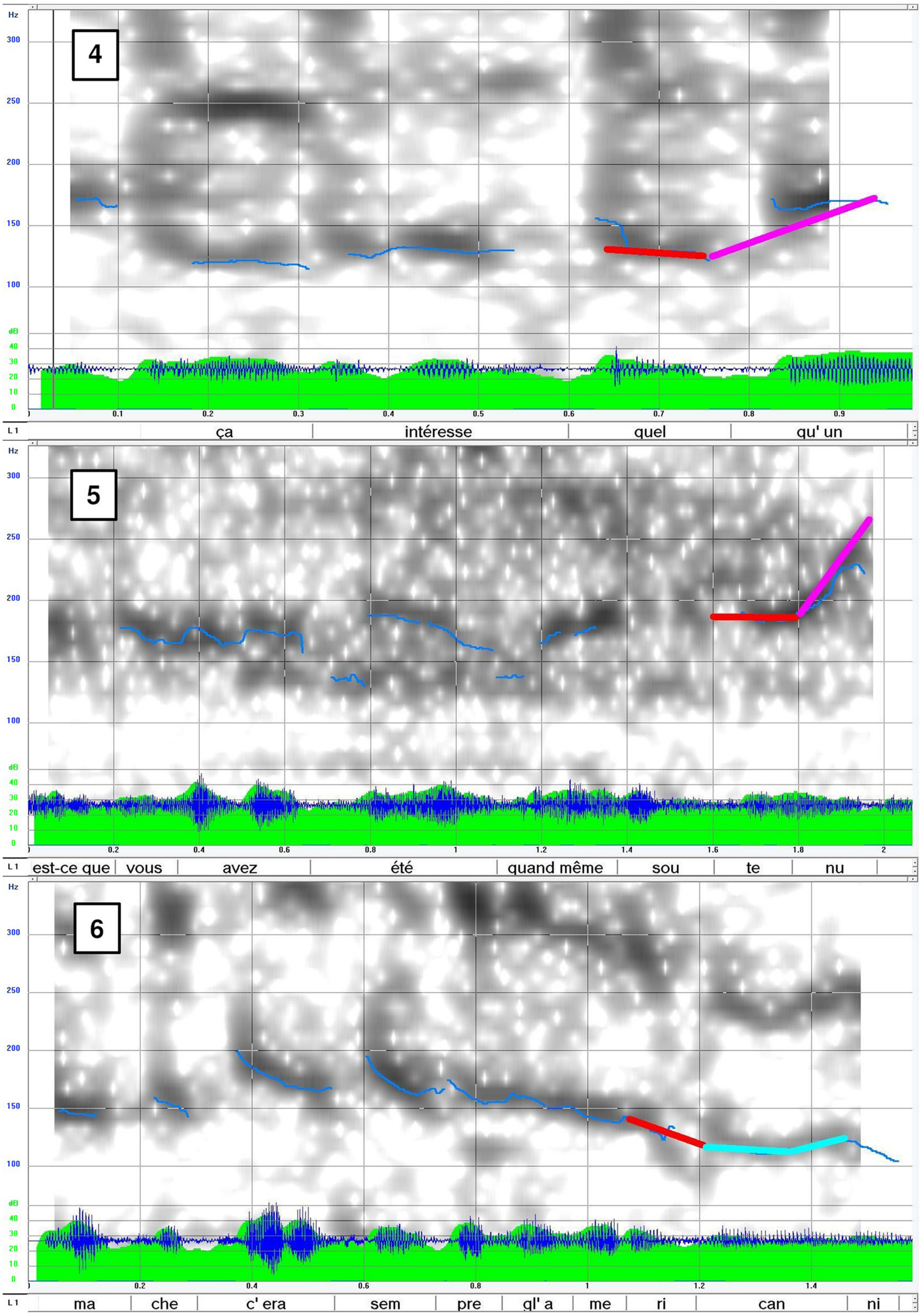

FIGURE 5
Prosodic variants of *Requests for confirmations* in French and Italian (4–6) (Supplementary Audio 3.wav).

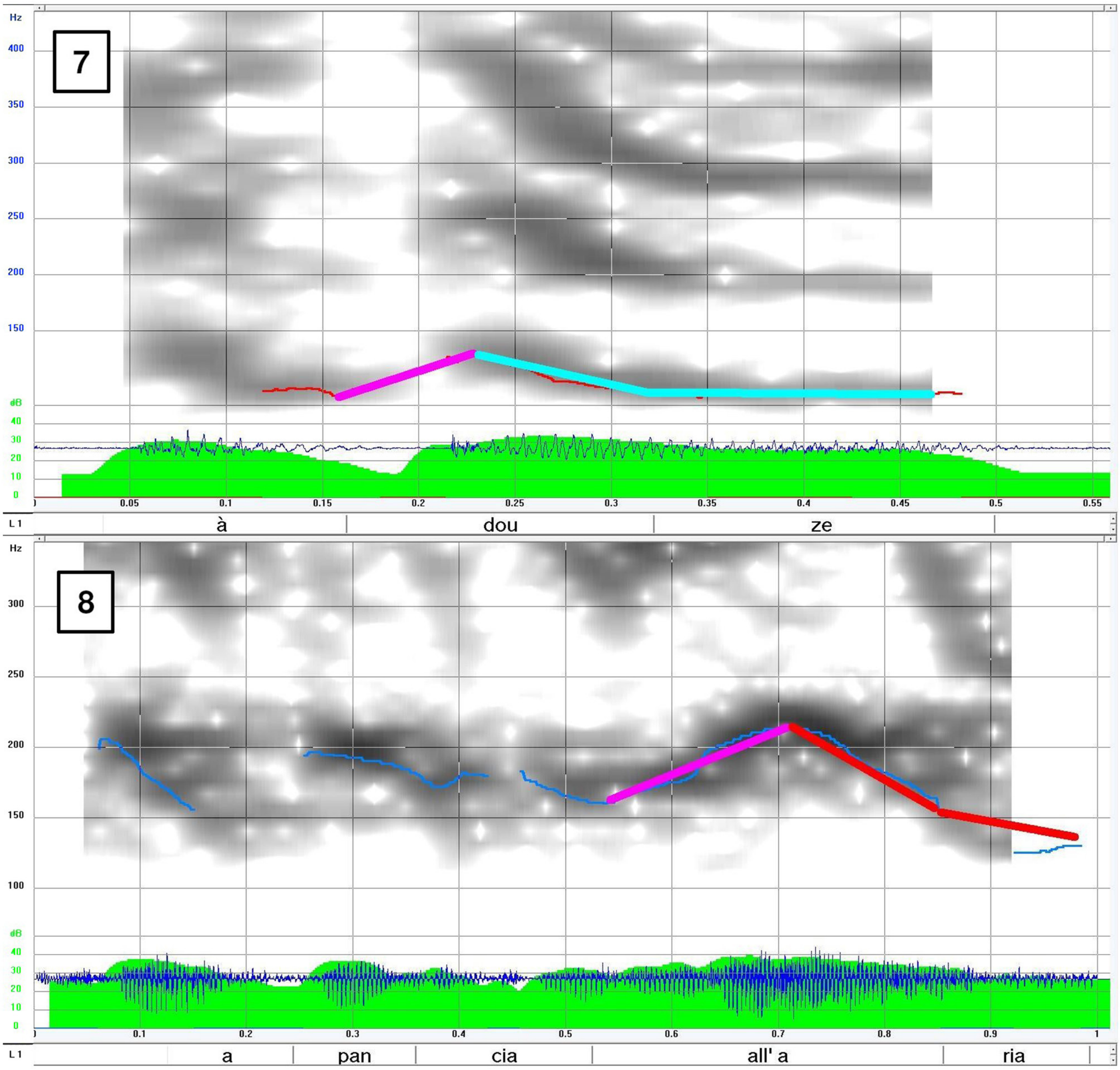

FIGURE 6
Prototypic contours of *Challenging questions* in French and Italian ("Validation Pattern" 7; 8) (Supplementary Audio 4.wav).

The so-called nuance of the meaning of these contours has been the topic of corpus-based and experimental research in French (Portes and Lancia, 2017; Portes, 2020), and similar research has been conducted in Italian and English (Cresti and Moneglia, forthcoming; Rossano, 2010; Hedberg et al., 2017).

Corpus data are a significant source of information to characterize the pragmatics/prosody relation within Total questions. Both in Italian and French, we found recurrent pragmatic patterns which must be distinguished from those conveyed by RCs that are performed through a mountain contour.

A first pattern occurs in dialogic contexts in which the speaker repeats or reformulates a previous addressee's assertion in the immediate precedent dialogic turn or picks up in the common ground the object/argument of the question. For instance, in (7), f0 contour in Figure 6, while talking about the Cuban revolution, DAM questions a circumstance (*à douze*)—indeed, 12 people starting a guerrilla is remarkable—which DAV has just recalled.

> (7) *DAV: donc ils ont débarqué avec andeuh un minimum d' armes / une douzaine / dans un endroit complètement pourri / des marécages // [...] et puis ils ont organisé la révolution / là dans les montagnes / *à douze* //
> *DAM: **à douze** ?
> *DAV: *à douze* // [efamdl10]
> [DAV: so, they landed with a minimum of weapons, a dozen, in an entirely rotten place, swamps. [...] and then they organized the revolution, there in the mountains, at twelve. DAM: at twelve? DAV: at twelve.]

In (8), f0 track in Figure 6, ELA reformulates the typical position of a happy cat—showing his belly—roughly recalled by LID.

> (8) *LID: [. . .] gli facevo / Pallino / Pallino // avanti ha fatto / fr fr ha girato / va' // poi s' è messo / a schiena 'n su / e come ands / si gongolava // tu vedessi <buffo> //
> *ELA: [<] <cioè> /**a pancia all' aria** ?
> *LID: a pancia all' aria // [ifamdl02]
> [LID: I called him: Pallino, Pallino. Before he did: fr fr. He turned, finally. Then, he put himself to his back 'n up and gloated. You should have seen how funny it was. ELA: you mean: belly in the air? LID: belly in the air.]

In the above contexts, the information questioned is immediately at the disposal of the addressee. Crucially, the question is not supported by any speaker's *hypothesis*. In no circumstance, the above context can be paraphrased as *the speaker makes the hypothesis of the occurrence of the given eventuality, asking the addressee for its confirmation*. In other words, corpus data, both in Italian and French, positively show that this kind of Total question lacks the necessary intentional preparatory condition we identified in RC.

We can describe the pragmatic activity considering that the speaker assumes what is said (or is emerging from the context) and challenges the addressee to *validate* it through his linguistic behavior as an intersubjectively shared assumption. The act may be peremptory in connection to the greater or lesser speaker's interest in validating the ongoing linguistic exchange.

Even this kind of Total question cannot be defined as *seeking for information acts* since the information is at the disposal, and the speaker wants just a guarantee. We refer to them as *Challenging questions* (CHQ).

All questions sharing this pragmatic pattern, both in the Italian and French corpora, show a *rising-falling* contour which we holistically call *mountain*—in parallel to the term *valley*. It corresponds to a complex contour composed of a rising movement (over the glissando) and a rapid falling movement to the baseline (over the glissando).[11] Movements are by preference, aligned to the tonic vowel, or the falling movement can be placed on the post-tonic, as in the examples (9) and (10) below. The contour may be followed optionally by a *holding* movement on words after the *nucleus* within the root unit.

In parallel, no Total questions, performed in French through a *rising* contour or in Italian through a *valley* contour, have been found in pragmatic contexts lacking the speaker's intentional feature *hypothesis on the eventuality*.

Thus, positive corpus data corroborate the suggestion that the variation of the prosodic contour of the Total question is at least sensitive to the existence or not of an underlying hypothesis by the speaker, leading to differential pragmatic activities (RC or CHQ).

However, in Italian and French, a set of Total questions with a *rising-falling* contour does not entirely match the previous pragmatic description, i.e., challenging the addressee to validate information already given as a shared assumption. Nonetheless, it is also confirmed in these pragmatic contexts that the speaker does not make any *hypothesis* on what is asked.

These are probably those language usages that led authors such as Riegel and Pellat (1999) to propose the term *injunctive questions*. The (9) and (10), f0 tracks in Figure 7, are good examples:

> (9) *NIC: [<] <ne ho fatte> tre //
> *CEC: **posso averne una** ?
> *NIC: sì // [ifamdl17]
> [NIC: I made three (pictures). CEC: can I have one? NIC: yes.]
> (10) *UBR: [. . .] pour les groupes et les cars / c' est seulement trente-huit francs //
> *PEL: oui c' est assez cher // et nous sommes obligés d' inclure le prix dans / notre produit de la journée // andeuh **est-ce que on peut voir un choix** ? [fpubdl02]
> [UBR: For groups and coaches, it's only thirty-eight francs. PEL: yes it's quite expensive. And we must include the price in our ticket of the day. Can we see anything else?]

In (9), NIC wants permission from his brother to keep a picture as a sign of affection. In (10), PEL pushes the seller to offer a different tour. Both utterances challenge the addressee by bearing a strong directive force instead of making a hypothesis. The answer is strongly requested to the addressee, and the interest of the speaker on the matter is even higher than in the validation context, thus fitting well with the term *Challenging questions*.[12]

The specific goal of Challenging questions varies and does not define the language activity. In (7) to (10), the speaker wants to validate a piece of information as a shared assumption; in (9), the interest is relative to the speaker's behavior (to keep the picture), and she asks for permission (linguistic behavior by the addressee). Conversely, in (10), the interest regards an action to be performed by the addressee[13].

The pragmatic similarities between Challenging question groups may be summarized as follows:

- The speaker does not make any hypothesis or assumption on the proposition.
- The speaker challenges the addressee with force.
- The interest of the speaker in the question is high.

In conclusion, Total questions find two main prosodic contours in Italian and French. The prosodic variation is sensitive to pragmatic features such as the presence or lack of the intentional preparatory condition regarding the speaker's hypothesis on one side and the force of the challenge toward the addressee on the other. If the *hypothesis* is present, the *force is weaker,* and the question enacts a Request for confirmation. If there is *no speaker's hypothesis* and the *force is strong,*

[11] Movements can be characterized by speech rate as *slow* or *rapid* (with respect to the rate of the entire root unit).

[12] Dehé and Braun (2020) consider declarative questions in English as rhetorical because they ask a question to create a dramatic effect or to make a point rather than to get an answer. Thus, they are considered *challenging* rather than *seeking information*. In our frame, CHQs genuinely challenge an answer, but cannot be classified as rhetorical for their directive force.

[13] According to the tradition, (12) can be seen as an *indirect speech act,* i.e., requests of *action* by the addressee instead of *linguistic behavior*.

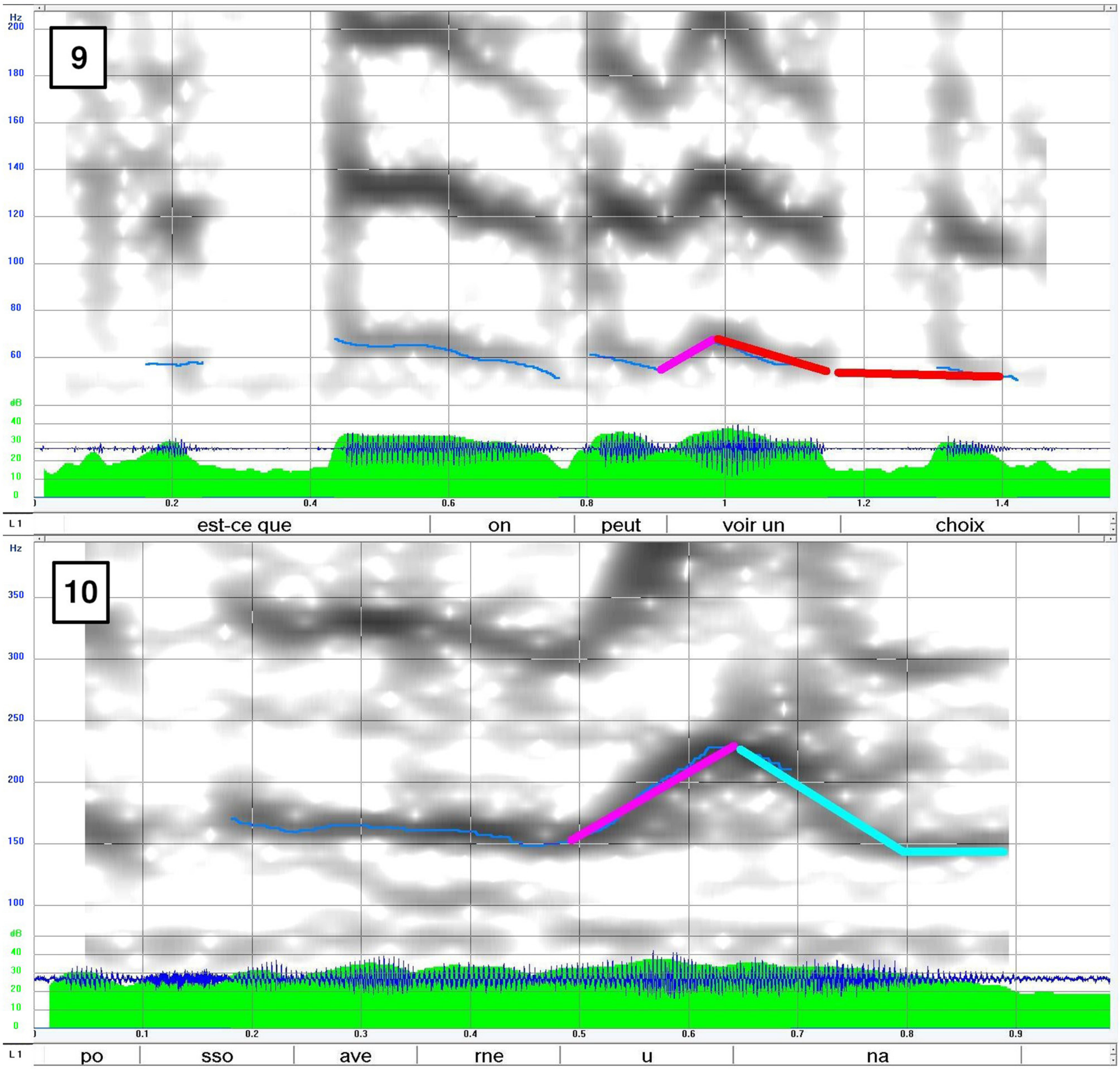

FIGURE 7
Prototypic contours of *Challenging questions* in French and Italian ("Injunctive Pattern" 9; 10) (Supplementary Audio 5.wav).

the question enacts the Challenging illocutionary type. Prosody varies accordingly.

3.2. Partial and open questions

3.2.1. Italian partial and open questions

Partial questions can be defined as those interrogative utterances whose scope is a variable expected to be saturated in the answer. In Italian, they represent nearly 38% of questions. Their syntactic structures correspond mostly to verbal phrases that the Wh- interrogative morpheme precedes (*chi, che cosa, quale, dove, quando, quanto, come, perché,* and *come mai*)[14]. The morpheme can occur alone or at the beginning of the Comment unit.

Partial questions are properly *seeking information* acts and do not imply any hypothesis by the speaker on the proposition satisfying the query.

The prosody of Partial questions is a *falling* contour (over the glissando) starting on the tonic vowel of the variable,

[14] The percentage of each single Wh variable is not reported since it is not relevant in such a small dataset.

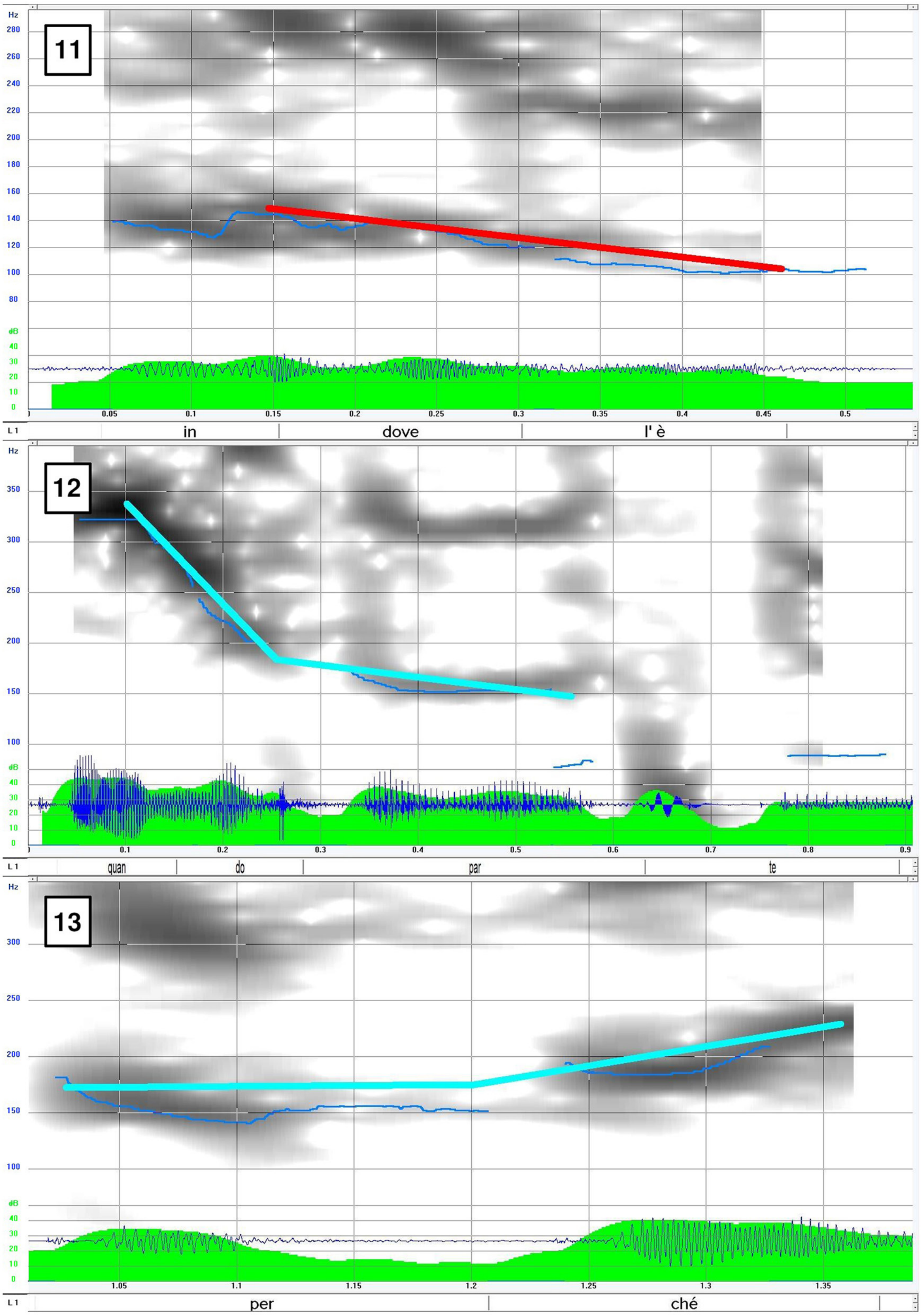

FIGURE 8
Prototypic contours of *Partial questions* in Italian (11–13) (Supplementary Audio 6.wav).

which can be characterized by a more or less rapid declination, depending on the height at the start, the length of the information unit, the speed of speech, and sometimes on stance. See, for instance, two WH questions performed by female speakers. (11) is introduced by *dove* (where) and is characterized by a neutral attitude. The *falling* movement, reported in the f0 track of Figure 8, starts at a middle height and declines slightly.

(11) *ELA: **n' dove l'è** ? [ifamcv01]
[ELA: where is it?]

(12), still showing a *falling* movement, is more lengthened and starts from higher f0 values. The question seems characterized by an attitude of *insistence*.

(12) *EST: **quando parte**? [ifamdl15]
[EST: When does she leave?]

In Italian, the only prosodic difference within Partial questions regards the performance of *perché* (*why*). In (13), this variable is realized through an upward movement that can be *slightly rising* (over the glissando), as in Figure 8. [15]

(13) *ELA **perché**? [ifamcv01]
[ELA: Why?]

Perché frequently occurs in isolation; however, other words can follow it, taking part in the question as a *tail* with no functional value. From a pragmatic point of view, *perché* questions are distinguished from other WHs because of their prosody and because they are expected to be saturated by an explanation rather than a noun or prepositional phrase.

The idiosyncratic prosodic performance of *perché* allows passing to the description of *Open questions* (OP), which is rarely considered in previous literature (Weisser, 2018). Its syntactic structure corresponds to a verb-less phrase (noun phrase, adverbial phrase, adjectival phrase, and verb nominalization) filling the Comment unit (13% of PQs).

In (14), f0 track in Figure 9, PAO performs a question (*per quest'estate?*), showing a *continuous rising contour* aligned to the last tonic vowel (over the glissando), optionally followed by holding movement on the post-tonic.

Although the question does not specify precisely what kind of answer is expected, it just seeks an explanation. The addressee cannot answer Y/N.

(14)*PAO: perché la domenica / tu se' qui / a <negozio> //
*FRA: da urlo //
*PAO: ma / anche / **pe' quest' estate** ? come / tu se' messa ? [ifamdl12]
[PAO: because on Sunday, you are here, at the store. FRA: screaming. PAO: but, and for this summer? how do you manage it?]

In (14), "*for this summer*" might be considered the topic of the Partial question, "*how do you manage it?*". But this is not the case since the first phrase bears an illocutionary force, and it is a Comment that can be interpreted as an independent question. The difference is reflected in the prosodic performance since OPs, as in the prototype in Figure 9, show a *continuous rising contour*.

The *continuous rising* contour distinguishes from *valley* and *mountain* contours of the Total types and resembles the one found with *perché*. Maybe not by chance that *perché* too foresees an *explanation* as an answer.

OPs are filled by a verb-less phrase, as is frequent in Total questions but cannot be answered by Y/N. The only reason for it can be up to their specific prosody. For instance, (15), f0' tracks in Figure 9, shows that the two prosodic contours on the same proper name (*la Pina*) determine their different pragmatic interpretations. CLA wants to hear about a girl whose name she does not remember well. EST assumes that she is named *Pina* performing a Request for confirmation. CLA asks an open question on this subject.

(15) *CLA: <e quella> ragazza / <la> ...
*EST: <**la**> **Pina** ?
*CLA: **la Pina** ?
*EST: eh // lei / prima veniva tutte le settimane // [ifamdl15]
[CLA: and that girl, the... EST: Pina? CLA: Pina? EST: eh. She used to come every week.]

The different contours of the same lexical word can be easily compared. While the first utterance shows the typical *valley* of RC, the second, the OP, is performed through a *continuous rising* contour.

The pragmatic function of OP can be summarized in the following conditions: a) as all WH, the question *does not convey a speaker's hypothesis*; b) it is genuine seeking for information activity; and c) it *waits to be freely completed* by the answer.

3.2.2. French partial and open questions

In French data, Partial questions represent nearly 26% of the total amount and are composed of the whole range of Wh-morphemes (*comment, où, quand, quel, quoi, combien, pourquoi*) but can also be headed by *q'est-ce que/qui*.

Wh- morphemes are distributed not only at the beginning of the Comment unit, as in Italian, but also *in situ* at its end and (only occasionally) within it. The most frequent distribution is at the beginning (nearly 65%) and only 25% is at the end. However, it must be stressed that the initial position is filled in 45% of cases by *q'est-ce que/qui, lequel, quel/quelle*, and *pourquoi*, that in our data do not find any other position. Therefore, the remaining Wh-variables are distributed fifty/fifty in the initial and final positions, which is prominent.

When the interrogative morpheme occurs at the beginning of the Comment, its prosodic performance corresponds to a *slightly falling* contour (over the glissando) starting from the tonic vowel of the Wh-. A slightly *rising tail* can optionally complete the movement, as in the f0 track of (16) in Figure 10.

(16) *MIR : **qu'est-ce qu'elles deviennent** maintenant ? [ffamdl17]
[MIR: what they become now?]

In (17), the prosodic contour of the most common Wh-morpheme (*comment*), distributed in the first position, is consistent with this description (f0 in Figure 10).

(17) *SAN: **comment ils vivent** ? [ffamdl02]
[SAN: how do they live?]

When the interrogative variable is *in situ,* the contour is nearly *flat* or weakly *rising* on the morpheme, marked by a minimal reset, strong intensity, and/or lengthening (still over the glissando). The f0 track of example (18) shows how prosody works on "*comment*".

[15] Depending on attitudinal connotations, for instance, *surprise*, it can also reach a higher peak.

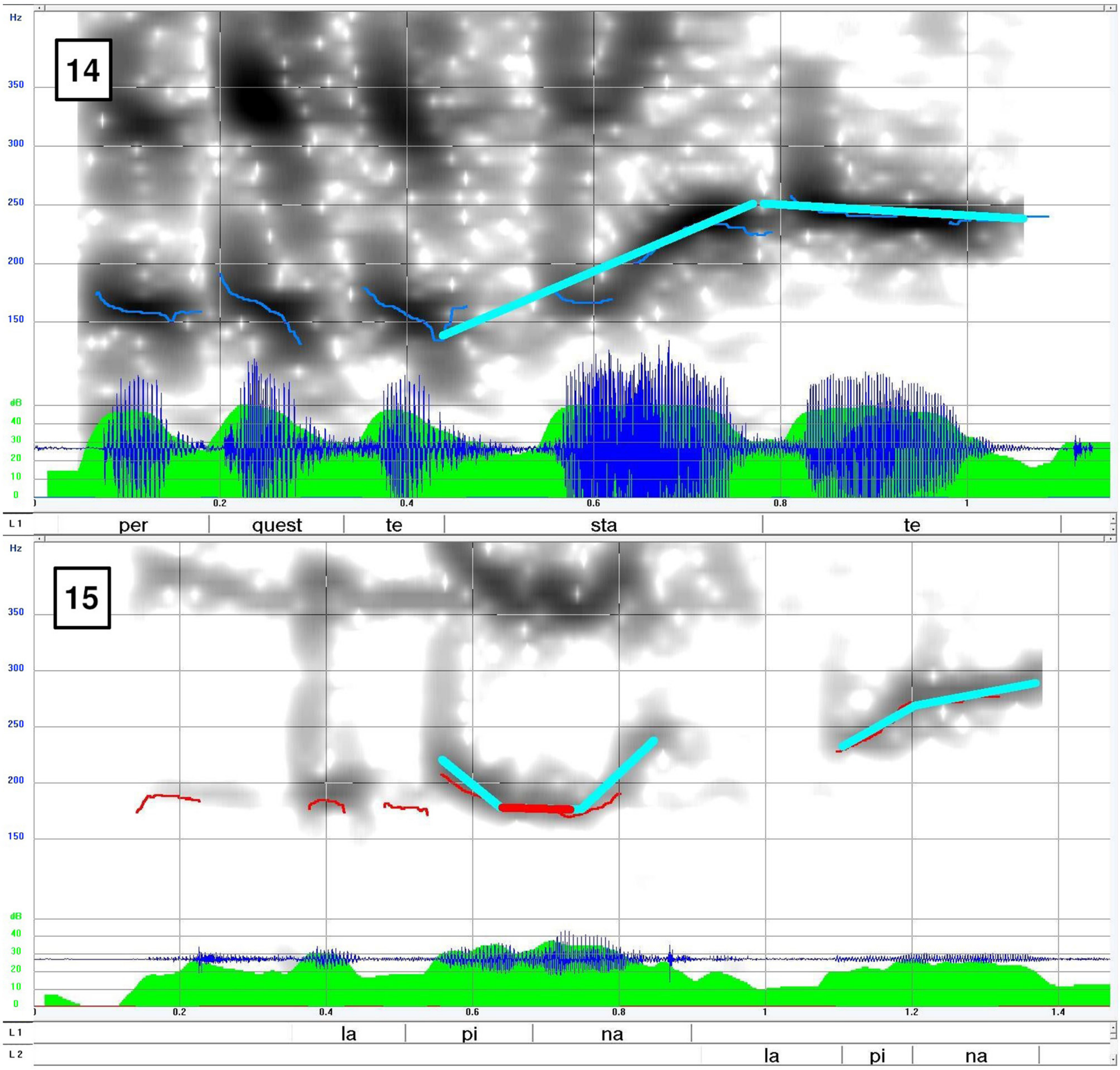

FIGURE 9
Prototypic contours of *Open questions* in Italian (14; 15) (Supplementary Audio 7.wav).

(18) *JOS: vous avez fait **comment** ? [ffamcv05]
[JOS: did you do it how?]

Open questions also occur in French (13.7% of PQs). Their syntactic structure is verbless phrases. The question is expected to be freely completed in the answer by an explanation, as in Italian. Their prosodic performance is *a continuous rising contour* (over the glissando). It can be only roughly compared to the Italian prosodic performance, as verified in the f0 track of (19) reported in Figure 10.

(19) *UBR: euh je vous garantis que andeuh [/] qu'avec ça / vos clients auront passé / une journée mais alors vraiment mémorable hein //
*PEL: **et le prix** ?
*UBR: andeuh le prix / andeuh /deux cents francs tout compris // [fpubdl02]
[UBR: uh, I guarantee that that with this, your customers will spend a day but then really memorable, eh. PEL: and the price? UBR: uh, the price, uh. Two hundred francs all inclusive.]

3.3. Questions as illocutionary patterns in Italian and French

3.3.1. Tag questions

Double questions, Alternative questions (both positive and negative), and Tag questions are performed within Illocutionary

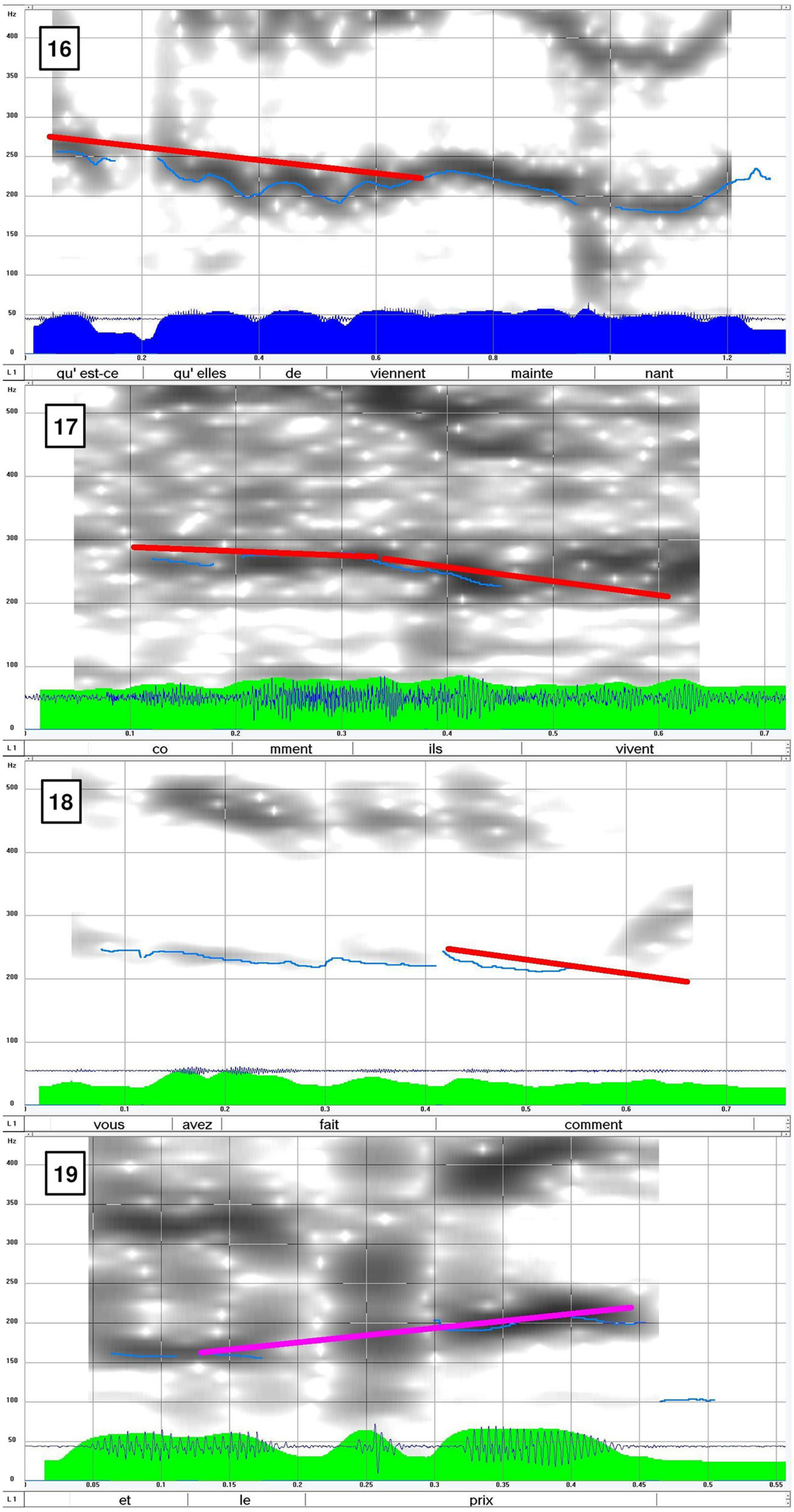

FIGURE 10
Prototypic contours of *Partial questions and Open questions* in French (16–19) (Supplementary Audio 8.wav).

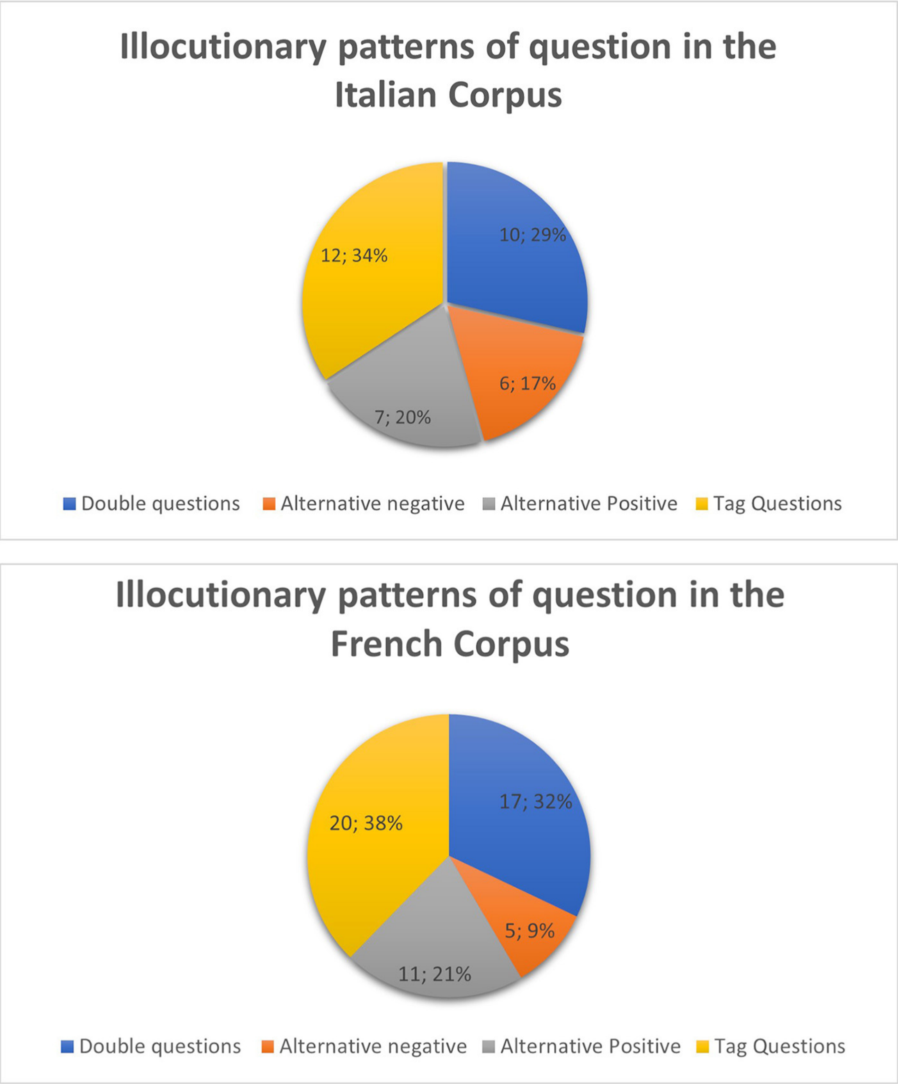

FIGURE 11
Illocutionary patterns of question in the Italian and French corpus.

pattern structures, which require the distribution of the utterance in two pragmatic units (CMM) in a prosodic pattern. All types are well documented in both corpora (from 9 to 15% of questions) and vary similarly in the two languages, as pies A and B of Figure 11 show. [16]

In Italian and French, Tag questions are composed of a first assertive CMM, shaped according to the prosodic variation allowed by the assertive Illocutionary types (Cresti, 2020), and a second CMM filled by a morpheme working as a tag unit. Without the tag unit, the utterance would not be considered a question. The morpheme directs to the addressee the previous assertion, which functions as a positive hypothesis, stimulating him to confirm it. From a pragmatic perspective, Tag questions are almost equivalent to weak RCs.

A restricted number of morphemes can constitute the lexical filling of the second CMM. In Italian: *forse, vero, no, eh,* and *giusto*; in French, *hein*, *non*, *vraiment*, and *c'est ça*. (20) and (21) show Tag questions in the two languages, where the more common lexical expression filling the tag label are, respectively, *eh* and *hein*. Their respective f0 track is displayed in Figure 12.

(20) *ELA: **Fiordalice / eh** ?
[ELA: Fiordalice, isn't it?] [ifamdl02]
(21) *EMA: **il y en a beaucoup / hein**? [efamdl23]
[EMA: there are a lot, isn't it?]

Prosody is crucial to enact the illocutionary role, which requires the morpheme to be a stressed vowel *rising* (over the glissando) or *lengthened*. This condition can be verified on corpus data. The morpheme *eh* in Italian and *hein* in French frequently also play

[16] Supplementary instances have been detected in IPIC database which stores the full informal section of C-ORAL-ROM Italian (Panunzi and Gregori, 2012).

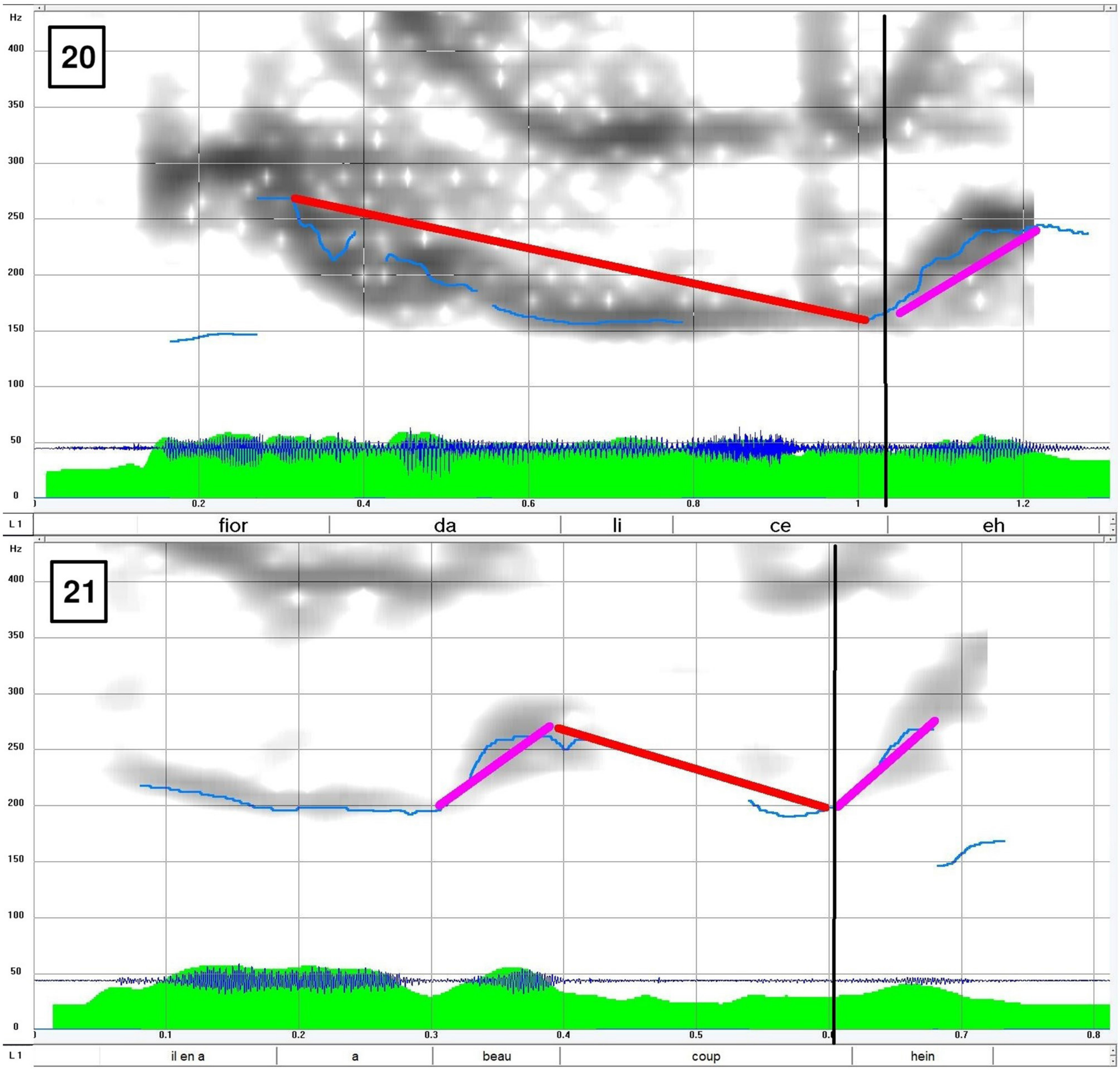

FIGURE 12
Prototypic contours of *Tag questions* in Italian and French (20; 21) (Supplementary Audio 9.wav).

a *Phatic* function, strengthening the previous assertion's force. In those cases, the morpheme is defocused and is performed at a lower f0 level (under the glissando).

3.3.2. Alternative questions

Half of the Illocutionary patterns in Italian and French are Alternative questions: either positive or negative. (22) and (23) are, respectively, Alternative negative (ANQ) and Alternative positives of the Italian Dataset, while (24) and (25) are comparable French examples. Figure 13 presents their f0 contours.

(22) *GIA: **le metto / o no**? [ifamcv14]
[GIA: do I put them on, or not?]
(23) *ANT: **ma lui è più grande / o più piccolo**? [ifamdl05]
[ANT: is he older, or younger?]
(24) *SOP: **ça vous dit quelque chose / ou pas** ? [fnatte02]
[SOP: does that mean anything to you or not?]
(25) *SOP : **c'était d'abord anglais / ou c'était d'abord français** ? [ffnatte02]
[SOP: was it English first / or was it French first?]

In no case, the Alternative patterns are equivalent to the performance of one sole question since the interpretation of the utterance does not concern a question on the truth of disjunctive coordination, as it might be in "*is it true that (chess pieces are white*

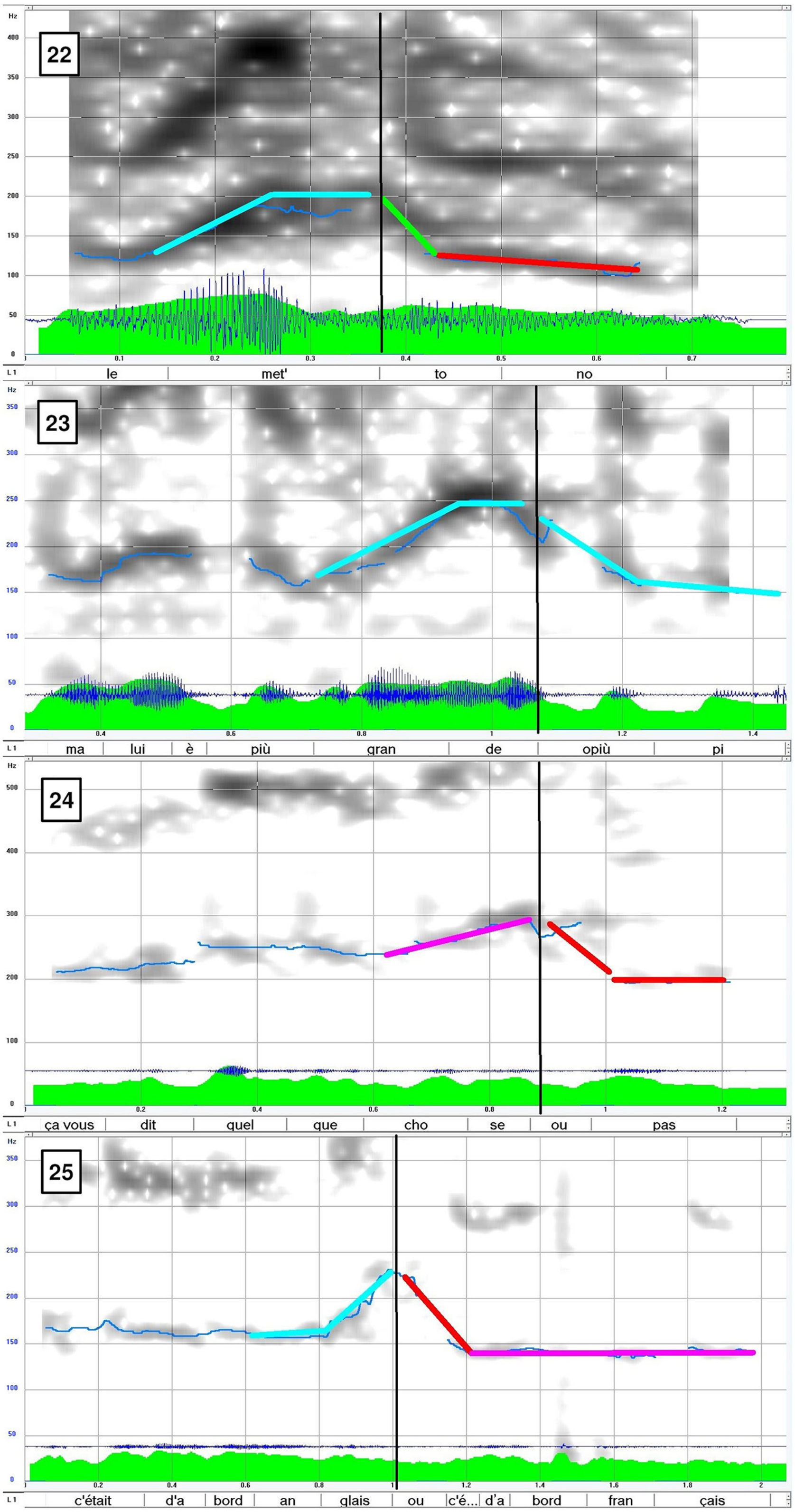

FIGURE 13
Prototypic contours of *Alternative questions* in Italian and French (22–25) (Supplementary Audio 10.wav).

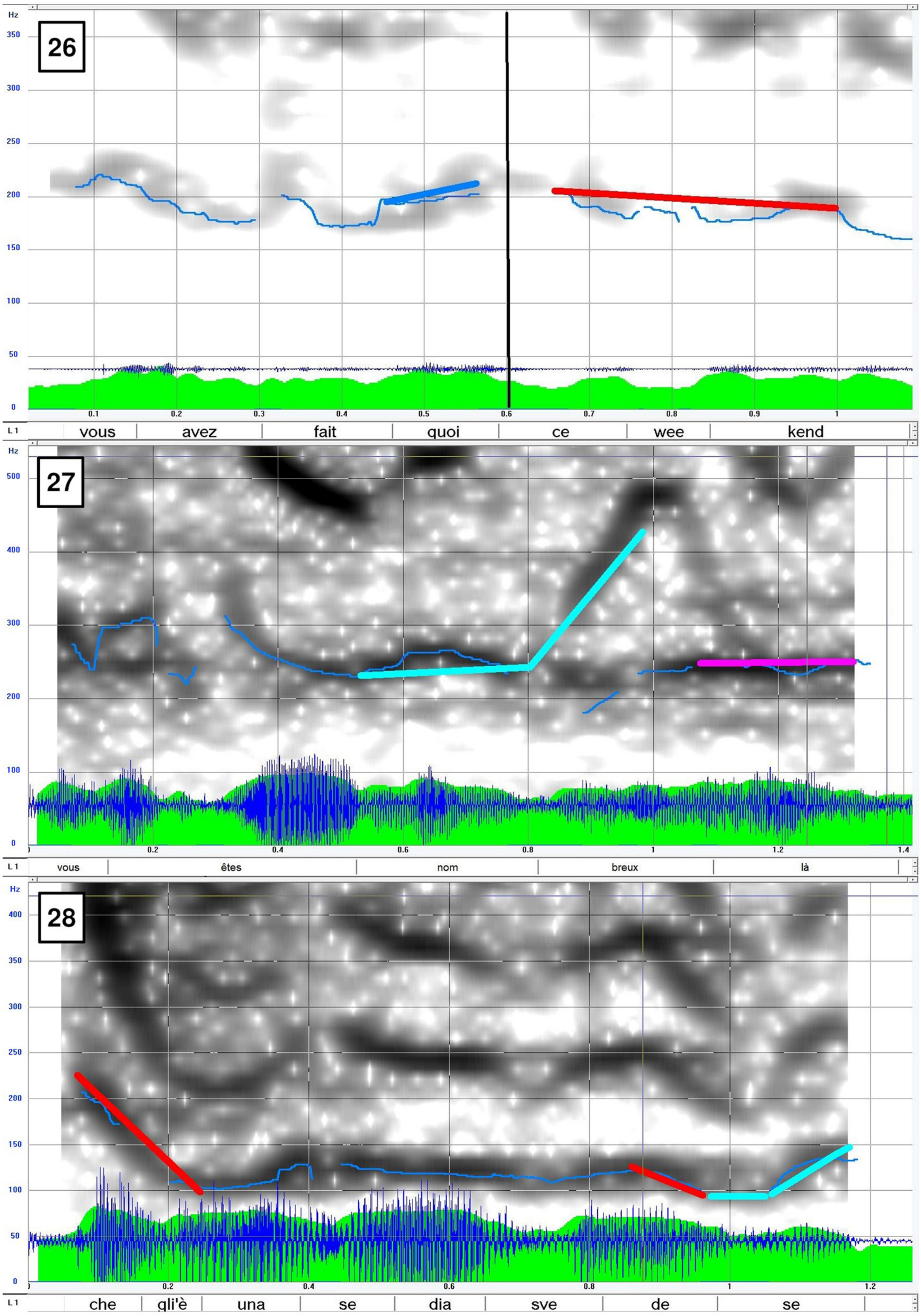

FIGURE 14
Prototypic contours of *Double questions* in French and Italian (26–28) (Supplementary Audio 11.wav).

or black)?".[17] Each CMM bears an illocutionary force of question, but the actual request is up to their association within the pattern, even if, as sometimes happens, the disjunctive conjunction (*or*) is not realized.

APQ and ANQ have different semantic values. In ANQ, one sole proposition is questioned and, for this reason, can be answered Y/N. On the contrary, APQ requires a choice between two propositions and cannot be answered Y/N.

ANQs work as a Total question; if the answer is negative, the question is solved. In contrast, the positive response is insufficient and requires supplementary information. Although APQs cannot be answered Y/N, a negative answer is possible, but it denies both possibilities and is by preference integrated by additional information.

From a prosodic point of view, quite surprisingly, APQ and ANQ show the same prosodic pattern, which is also the same in Italian and French. The first CMM bears a *rising* movement (over the glissando) with a peak on the final tonic vowel. In Italian, the rising contour is followed by a *hold* (under the glissando) if the word is paroxytone. The second CMM bears a *level* contour composed of *rapidly falling* on the first tonic vowel, followed by a *holding* movement at low f0 values (over the glissando) spread on the subsequent syllables. This description roughly replicates the prosodic description of alternative constructions in French (Delais-Roussarie and Turco, 2019).

Although the semantics of alternative questions might be extensively treated, from corpus data that emerge on the pragmatic ground, both APQ and ANQ, to be meaningful, cannot imply a *hypothesis* by the speaker on the questioned alternative and cannot be interpreted as RC but rather function as CHQ.

3.3.3. Double questions

Although virtually ignored in Italian or French grammar, Double questions (DBQ) are well testified in language usage. As with the other Illocutionary patterns, they split the question into two prosodic contours (CMMs). They were described by Fonagy and Brerard (1973) as a *double-rising* movement belonging to the same prosodic contour. They report that Dubois already considered it like a *prosodic redundancy* on a unitary content that could constitute a single interrogative sentence. However, this description does not fit with the speech performance since each CMM signals a distinct act.

Italian and French DBQ present different types. Still, the overall structure always consists of a first CMM performing a question and a second CMM restricting the first. For instance, French features Double patterns with either a WH or a Total question in the first CMM, as in (26) and (27).

(26) *MAI: **vous avez fait quoi / ce week-end** ?
[MAI: what did you do this weekend?] [ffamcv05]
(27) *SAN: **vous êtes nombreux / là** ?
[SAN: there were many of you, there?] [fpubdl01]

From a prosodic point of view, the contour of the WH is strictly compatible with the overall WH model, showing a prosodic marking of the Wh- variable, which is placed mainly *in situ*, as the f0 track of (26) in Figure 14. The second CMM shows a *level* contour (over the glissando) spread over the sequence of syllables. The answer should satisfy the Wh- variable.

The prosodic pattern of (27), displayed in Figure 14, corresponds to the sequence of a *rising* contour on the last vowel, strictly compatible with French RCs in the first CMM and a *level* contour (still over the glissando) in the second CMM, as in (26). The answer is necessarily Y/N.

It is worth noting that it is hard to interpret the second CMM as an autonomous question. However, the unit is not defocused as it might be the case, in the L-AcT perspective, for an Appendix of Comment (Cresti, 2021). As Martin (2008, 2018) noticed discussing similar questions within the Macro-syntactic approach (Blanche-Benveniste, 1997), the second unit of a DBQ works as a *postfix*. It must be distinguished from both an independent *noyau* and a *suffix* unit. In other words, the *high-level* contour bears a functional value in French, working as a *specification act* within the Illocutionary pattern.

DBQ in our Italian sampling always corresponds to a first CMM performing a WH question, followed by a second CMM restricting its scope.[18] Let us see (28).

(28) *POL: **che gli'è / una sedia svedese**?
[POL: what is (it), a Swedish chair?] [ifamcv27]

Figure 14 shows that the prosodic contour in the first CMM of (28) is *falling*, matching the Italian WHs. In contrast, the second CMM fits with the contour assigned to RC (*valley*), with the falling part (over the glissando) on the last tonic vowel and a hold (under the glissando) followed by a rising (over the glissando) on the post-tonic. In (28), the second CMM saturates the Wh- variable, thus modifying the status of the first WH question, which becomes an RC and can be answered Y/N.

Despite the variation found in DBQ, which may correspond to WH (*seeking information*) and RC, we never found Illocutionary patterns working as *Challenging questions*. In other words, the restrictive role of the second CMM develops a mitigation effect.

4. Conclusion

In the L-AcT frame, the pragmatic domain of questions is a sub-class within the overall Direction class and is defined as the speaker's *activity aimed at the addressee's linguistic behavior*, passing from being considered the speaker's *seeking for information activities* as corpus-based taxonomies assume, to a more general concept.

Two selections of ∼4,000 utterances have been derived from the Italian and French C-ORAL-ROM corpora. Focusing on the Comment information unit, we derived the main prosodic properties of questions in the two languages and their pragmatic correlations.

[17] In Weisser's Taxonomy, the Alternative question is foreseen but is composed of only one syntactic unit introduced by the conjunction *or*.

[18] The type of DBQ composed of a first CMM with a Total question, as in French, is missing in the Italian corpus.

TABLE 4 The system of *Requests aimed at the addressee's linguistic behavior* in Italian and French.

	Corpus prototype	Illocutionary activity	Prosodic contour
Types of Utterance			
Wh-question (PQ)	(12) quando parte ?	seeking information request conditioned by a wh-variable	*falling* contour
	(17) comment ils vivent ? (18) vous avez fait comment ?		*slightly falling* contour or *minimal reset* (on the Wh *in situ*)
Open question (OP)	(14) pe' quest'estate ?	seeking information request without conditions on the answer	*continuous rising* contour (spread on the syllabic sequence)
	(19) et le prix ?		
Request for confirmation (RC)	(2) vicino Bologna ?	Agreement request based on the speaker's hypothesis	*valley* contour (*falling* on the tonic vowel and *holding-rising* on the post-tonic syllable(s))
	(3) pour le moment tu fais tes études ?		*(high) rising* contour on the last tonic syllable
Challenging Question (CHQ)	(8) a pancia all'aria ?	A pressing request based on linguistic/contextual evidence to satisfy the query	*mountain* contour (*rising-falling* on the tonic syllable followed by a *hold*)
	(7) à douze ?		
Types of Illocutionary pattern			
Tag questions (TAG)	(20) Fiordalice / eh ?	Agreement request on an asserted utterance	*assertive rising-falling* contour + *rising* contour
	(21) il en a beaucoup / hein?		
Alternative negative questions (ANQ)	(22) le metto / o no ?	A request of choice between the content of an utterance and its negation	*rising* contour + *level* contour
	(24) ça vous dit quelque chose / ou pas ?		
Alternative positives questions (APQ)	(23) ma lui è più grande / o più piccolo ?	A request of choice between contrastive contents in the pattern	*rising* contour + *level* contour
	(25) c'était d'abord anglais / ou c'était d'abord français ?		
Double questions (DBQ)	(28) che gli'è / una sedia svedese?	A seeking information request restricted to an area of relevance through a specification act	*falling* contour + *valley* contour
	(26) vous avez fait quoi / ce week-end ? (27) vous êtes nombreux / là ?		*minimal reset* on the Wh (or *rising* contour) + *level* contour

The first finding is that <10% of utterances are questions in language usage compared to assertive over 50%. The corpus-based study allows sketching the quantitative proportion among the main question types. Total questions (37% FR −39% IT) and WH questions (26% FR −38% IT) are the most frequent type. The relevance of questions performed through Illocutionary patterns (Tag-questions, Alternative questions, and Double questions) is highlighted (9% IT −5% FR).

Beyond the grammatical distinction between Total and Partial questions, the syntax finds little relevance in the typological classification of questions. Semantic aspects, such as positive and negative presuppositions (*bias*), are not predictive of the pragmatic typology. Total questions, which are expected to be answered by the addressee positively or negatively, and which are supposed to fit standard prosodic models, show a consistent number of instances bearing a so-called declarative contour, respectively 26% in French and 36% in Italian. This is a relevant corpus finding which asks for an explanation.

We provided evidence correlating the two contours with pragmatic values. On one side, the *valley* contour (in Italian) and the *final rising* (in French) aim to obtain the addressee's agreement on a *speaker's hypothesis*, accomplishing a Request for confirmation (RC). Conversely, the *mountain* contour (in both languages) *challenges the addressees* (CHQ). Both types depart from the shared pragmatic definition of questions as *seeking information acts*. The prosodic contours reflect intentional and affective differences, mainly depending on the balance between the weight of the speaker's hypothesis, which only pertains to RC, the interest toward the addressee, and the force of the request, which is highest in CHQ.

Seeking for information activities are limited to Partial questions. PQs are strong requests to the addressee, performed by a *falling* contour in both languages and, in French, by marking the Wh- variable when *in situ*. PQs comprise Open questions. OPs show a specific prosodic contour (*continuous rising*) shared by the two languages and correspond to a vague underlying WH question the addressee is free to satisfy as he prefers.

The Illocutionary patterns are natural rhetorical figures linking two illocutionary units (CMMs) within a prosodic structure. Tag questions manifest the speaker's hypothesis in a first assertive CMM, while the Tag morpheme signals an RC to the addressee. The prosodic pattern corresponds to an *assertive* contour in the first CMM, followed by a focused *rising* or *lengthened* contour.

In Alternatives questions, positive and negative, the content is composed of two pragmatic units put in alternation. In both languages, the prosodic pattern is composed of a *rising* contour followed by a *low-level* contour. From a pragmatic point of view, contrary to Tag questions, the Alternative Illocutionary pattern develops a CHQ.

Double questions enact a speaker's way of realizing complex questions that are hierarchically structured: the first CMM always demands to be further specified in the second. This general model is reflected in various types depending on the semantic relations between the two CMMs. Although our description is limited by the small number of Double questions in the dataset, some difference between the two languages has emerged.

In Italian, Double questions are mainly composed of a first WH question performed through a canonical *falling* contour. The second CMM saturates the Wh- and the Illocutionary pattern works as an RC. In French, we found the first CMM recording either a WH question or a Total question; however, the second CMM always restricts it to a circumstantial argument. Accordingly, the Illocutionary patterns work as a WH or as an RC.

Table 4 summarizes the pragmatic system of activities aimed at the addressee's linguistic behavior shared by the two languages and the differential prosodic contours allowing their performance.

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. For the availability of the complete data set, see footnote 7.

Author contributions

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

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/fcomm.2023.1124513/full#supplementary-material

References

Abeillé, A., and Godard, D. (2021). *La Grande Grammaire du Français*. Paris: Actes du Sud.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., et al. (1991). The HCRC map task corpus. *Lang. Speech* 34, 351–366. doi: 10.1177/002383099103400404

Austin, J. L. (1962). *How to Do Things with Words*. Oxford: Oxford University Press.

Barbosa, P., and Raso, T. (2018). Spontaneous speech segmentation: functional and prosodic aspects with applications for automatic segmentation / A segmentação da fala espontânea: aspectos prosódicos, funcionais e aplicações para a tecnologia. *Rev. Estudo Ling.* 26, 397–1433. doi: 10.17851/2237-2083.26.4.1361-1396

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *The Longman Grammar of Spoken and Written English*. London: Longman.

Blanche-Benveniste, C. L. (1997). *Approches de la Langue Parlée en Français*. Paris: Ophrys.

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

Bunt, H., Petukhova, V., Traum, T., and Alexandersson, H. (2017). "Dialogue acts annotation with the ISO 24617-2 standard," in *Multimodal Interaction with W3C Standards: Towards Natural User Interfaces to Everything*, ed D. Dahl (Berlin: Springer), 109–135.

Cavalcante, F. A., and Ramos, A. (2016). The american english spontaneous speech mini-corpus. *CHIMERA. Romance Corpora Linguist. Stud.* 3, 99–124. doi: 10.15366/chimera2016.3.2

Chafe, W. (1994). *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago, IL: The University of Chicago Press.

Cresti, E. (2000). *Corpus di Italiano Parlato*. Firenze: Accademia della Crusca.

Cresti, E. (2020). "The pragmatic analysis of speech and its illocutionary classification according to the language into act theory" in *Search of Basic Units of Spoken Language: A Corpus-Driven Approach*, eds S. Izre'el, H. Mello, A. Panunzi, and T. Raso (Amsterdam: Benjamins), 177–216.

Cresti, E. (2021). The appendix of comment according to language into act theory: corpus-based research. *CHIMERA* 8, 46–69.

Cresti, E., and Moneglia, M. (2018). "The illocutionary basis of Information Structure. Language into Act Theory (L-AcT)," in *Information Structure in Lesser-Described Languages: Studies in Prosody and Syntax*, eds E. Adamou, K. Haude, M. Vanhove (Amsterdam: Benjamins), 359–401.

Cresti, E., and Moneglia, M. (forthcoming). "The role of prosody for identifying illocutionary types. Polar questions vs. Request for confirmation in the LAcT Taxonomy," in *Studi Italiani di Linguistica Teorica e Applicata*, eds G. M. Alfonsetti, F. Orletti, and E. Banfi.

Cresti, E., and Moneglia, R. O. M. (2005). *Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: Benjamins.

Crocco, C. (2006). *Prosodic and Informational Aspects of Polar Questions in Neapolitan Italian. Proceedings of Speech Prosody 2006*. Dresden: TUDPress.

Dehé, N., and Braun, B. (2020). The prosody of rhetorical questions in English. *English Lang. Ling.* 24, 607–635. doi: 10.1017/S1360674319000157

Delais-Roussarie, E., Post, B., Avanzi, M., Buthke, C., Di Cristo, A., Feldhausen, I., et al. (2015). "Intonational phonology of French: developing a ToBI system for French," in *Intonation in Romance*, eds Frota S, Prieto P (Oxford: Oxford University Press), 63–100.

Delais-Roussarie, E., and Turco, G. (2019). "Intonation of alternative constructions in French" in *Romance Language and Linguistic Theory 15*, eds I. Feldhausen, M. Elsig, I. Kuchenbrandt, M. Neuhaus (Amsterdam: Benjamins), 135–136.

D'Imperio, M. (2001). Focus and tonal structure in Neapolitan Italian. *Speech Commun.* 33, 339–356. doi: 10.1016/S0167-6393(00)00064-9

Firenzuoli, V. (2003). *Le forme intonative di valore illocutivo dell'italiano parlato: analisi sperimentale di un corpus di parlato spontaneo (LABLITA)*. [PhD thesis]. [Firenze]: University of Firenze.

Fonagy, I., and Brerard, E. (1973). Questions totales simples et implications en français parisien. *StudiaPhonetica* 8, 53–97.

Gili-Fivela, B. (2008). *Intonation in Production and Perception. The Case of Pisa Italian*. Alessandria: Edizioni Dell'Orso.

Grice, M., D'Imperio, M., Savino, M., and Avesani, C. (2005). "Strategies for intonation labeling across varieties of Italian" in *Prosodic Typology*, ed S. Jun (Oxford: Oxford University Press), 363–89.

Grundstrom, A., and Léon, P. (1973). *Interrogation et Intonation en Français Standard et Canadien*. Montréal: Didier.

Hart, J. (1976). Psychoacoustic backgrounds of pitch contour stylization. *IPO Ann. Prog. Rep.* 11, 11–19.

Hart, J., Collier, R., and Cohen, A. (1990). *A Perceptual Study on Intonation. An Experimental Approach to Speech Melody.* Cambridge: Cambridge University Press.

Hedberg, N., Sosa, J. M., and Görgül,ü, E. (2017). The meaning of intonation in yes-no questions in American English: a corpus study. *Corpus Ling. Ling. Theor.* 13, 1–48. doi: 10.1515/cllt-2014-0020

Hirst, D., and Cristo, D. A. (1998). *Intonation Systems: A Survey of Twenty Languages*. Cambridge: Cambridge University Press.

Izre'el, S., Mello, H., Panunzi, A., and Raso, T. (2020). *In Search of Basic Units of Spoken Language: A Corpus-Driven Approach*. Amsterdam: Benjamins.

Kohler, K. (2004). "Pragmatic and attitudinal meanings of pitch patterns in german syntactically marked questions," in *From Traditional Phonology to Modern Speech Processing,* eds G. Fant, H. Fujisaki, J. Cao (Bejing : Foreign Language Teaching and Research Press), 205–214.

Lacheret-Dujour, A., and Beaugendre, F. (1998). *La Prosodie du Français*. Paris: CNRS editions.

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

Martin, P. H. (2008). *Postfixes et Suffixes Interrogatifs : Un cas D'ambiguïté Prosodique? Conférence de la Section Tchéco-Slovaque de l'ISPhS*, 111–119.

Martin, P. H. (2009). *Intonation du Français*. Paris: Armand Colin.

Martin, P. H. (2015). *The Structure of Spoken Language. Intonation in Romance*. Vol. 1. Cambridge: Cambridge University Press.

Martin, P. H. (2018). *Intonation, Structure Prosodique et Ondes Cérébrales*. London: ISTE.

Martin, P. H. (2022). "Intonation of telephone conversations in a Customer Care service" in *Corpora and Linguistic Studies*, eds E. Cresti and M. Moneglia (Milano: Officina21), 325–336. doi: 10.17469/O2106SLI000021

Mello, H. C., and Raso, T. (2012). "Illocution Modality Attitude: different names for different Categories," in Pragmatics and Prosody. Illocution Modality Attitude, Information, Patterning and Speech Annotation, eds H. Mello, A. Panunzi, T. Raso (Florence: Florence University Press), 1–18.

Moneglia, M. (2005). "The C-Oral-Rom resource," in *Integrated Reference Corpora for Spoken Romance Languages*, eds E. Cresti, and M. C. Moneglia (Amsterdam: Benjamins), 1–70.

Moneglia, M. (2011). Spoken corpora and pragmatics. *Rev. Brasileira Ling. Apl.* 11, 479–519. doi: 10.1590/S1984-63982011000200009

Moneglia, M., and Raso, T. (2014). "Notes on the language into act theory," in *Spoken Corpora and Linguistics Studies*, eds T. Raso, and H. Mello (Amsterdam: Benjamins), 468–494.

Moraes, J., and Rillard, A. (2014). "Illocution attitudes and prosody: a multimodal analysis, in *Spoken Corpora and Linguistics Studies*, eds T. Raso, and H. Mello (Amsterdam: Benjamins), 233–70.

Niebuhr, O. (2015). "Gender differences in the prosody of German questions," in *Proceedings of the 18th International Congress of Phonetic Sciences*, eds M. Wolters, J. Livingstone, B. Beattie, R. Smith, M. MacMahon (Glasgow: Scotland), 1–5.

Panunzi, A., and Gregori, L. (2012). "An XML database for the representation of information structure in spoken language," in *Pragmatics and Prosody. Illocution, Modality Attitude, Information, Patterning, and Speech Annotation*, eds A. Panunzi, T. Raso, H. Mello (Firenze: Florence University Press), 133–150.

Panunzi, A., Gregori, L., and Rocha, B., (2020). "Comparing annotations for the prosodic segmentation of spontaneous speech," in *Search of Basic Units of Spoken Language: A Corpus-Driven Approach*, eds S. Izre'el, H. Mello, A. Panunzi, T. Raso (Amsterdam: Benjamins), 403–432.

Panunzi, A., and Saccone, V. (2018). Complex illocutive units in L-AcT: an analysis of non-terminal prosodic breaks of bound and multiple comments. *Rev. Ling.* 26, 1647–1674. doi: 10.17851/2237-2083.26.4.1647-1674

Pierrehumbert, J., and Hirschberg, J. (1990). "The meaning of intonation contours in the interpretation of discourse," in *Intentions in Communications*, eds R. Cohen, J. Morgan, M. E. Pollack (Cambridge, MA: MIT Press), 271–311.

Portes, C. (2020). *Intonation Meaning in Question(s) the Case of Some French Rising-Falling Tunes Paper Presented at the Online Workshop Non-Canonical Questions at the Syntax-Prosody Interface*. Paris: Université de Paris.

Portes, C., and Lancia, L. (2017). Earlier or higher? Comparing french rising-falling contour with rising contour in a corpus of conversation. *J. Phonetics.* 63:35–52. doi: 10.1016/j.wocn.2017.04.002

Raso, T., and Mello, H. C., (2012). *Corpus de Referência de Português Brasileiro Falado Informal*. Belo Horizont: Editora UFMA. (2012).

Riegel, M., Pellat, J. C. (1999). *La Grammaire Méthodique du Français*. Paris: P. U. F.

Rocha, B. (2016). Uma Metodologia Empírica Para a Identificação e Descrição de Ilocuções e a Sua Aplicação Para o Estudo da Ordem em, PB, e Italiano. [Ph, D., and thesis]. [Belo Horizonte]: UFMG.

Rossano, F. (2010). Questioning and responding in Italian. *J. Pragmat.* 42, 2756–2771. doi: 10.1016/j.pragma.2010.04.010

Rossi, M. (1971). Le seuil de glissando ou seuil de perception des variations tonales pour la parole. *Phonetica.* 23, 1–33. doi: 10.1159/000259328

Rossi, M. (1978). La perception des glissando descendants dans les con-tours prosodiques. *Phonetica* 35, 11–40. doi: 10.1159/000259920

Rossi, M. (1999). *L'intonation, le Système du Français. Description et Modélisation*. Paris: Ophrys (1999).

Savino, M. (2012). The intonation of Polar Question in Italian: where is the rise? *J. Int. Phon. Assoc.* 42, 23–48. doi: 10.1017/S002510031100048X

Sbisà, M., and Turner, K. (2013). *Pragmatics of Speech Actions*. Berlin: Mouton de Gruyter (2013).

Searle, J. (1979). "A taxonomy of illocutionary acts," in *Expression and Meaning: Studies in the Theory of Speech Acts*, ed J. Searle (Cambridge: Cambridge University Press), 1–29.

Searle, J., and Vanderveken, D. (1985). *Foundations of Illocutionary, Logic.* Cambridge: Cambridge University Press.

Serianni, L. (1998). *Grammatica Italiana*. Torino: UTET. (1998).

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *J. Acoustic. Soc. Am.* 101, 514–521. doi: 10.1121/1.418114

Vanrell, M. M., Mascaró, I., Torres-Tamarit, F., and Prieto, P. (2013). Intonation as an encoder of speaker's certainty: information and confirmation yes–no questions in Catalan. *Lang. Speech* 56, 163–190. doi: 10.1177/0023830912443942

Weisser, M. (2018). *How to Do Corpus Pragmatics on Pragmatically Annotated Data: Speech Acts and Beyond.* Amsterdam: Benjamins.

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 26 April 2023
DOI 10.3389/fcomm.2023.1116955

Check for updates

OPEN ACCESS

EDITED BY
Sandra Madureira,
PUCSP, Brazil

REVIEWED BY
Oliver Niebuhr,
University of Southern Denmark, Denmark
Ingo Siegert,
Otto von Guericke University
Magdeburg, Germany

*CORRESPONDENCE
Nicole Holliday
✉ nicole.holliday@pomona.edu

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

RECEIVED 06 December 2022
ACCEPTED 20 March 2023
PUBLISHED 26 April 2023

CITATION
Holliday N (2023) Siri, you've changed! Acoustic properties and racialized judgments of voice assistants. *Front. Commun.* 8:1116955.
doi: 10.3389/fcomm.2023.1116955

Siri, you've changed! Acoustic properties and racialized judgments of voice assistants

Nicole Holliday*

Department of Linguistics and Cognitive Science, Pomona College, Claremont, CA, United States

As speech technology is increasingly integrated into modern American society, voice assistants are a more significant part of our everyday lives. According to Apple, Siri fulfills 25 billion requests each month. As part of a software update in April 2021, users in the U.S. were presented with a choice of 4 Siris. While in beta testing, users on Twitter began to comment that they felt that some of the voices had racial identities, noting in particular that Voice 2 and Voice 3 "sounded black." This study tests whether listeners indeed hear the different Siri voices as sounding like speakers from different groups, as well as examines voice quality features that may trigger these judgments. In order to test evaluations of the four voices, 485 American English listeners heard each Siri voice reading the Rainbow Passage, *via* online survey conducted on Qualtrics. Following each clip, listeners responded to questions about the speaker's demographic characteristics and personal traits. An LMER model of normalized ratings assessed the interaction of voice and race judgment revealed that indeed, Voice 2 and Voice 3 were significantly more likely to be rated as belonging to a Black speaker than Voices 1 and 4 ($p < 0.001$). Per-trait logistic regression models and chi-square tests examining ratings revealed Voice 3, the male voice rated as Black, was judged less competent ($X^2 = 108.99$, $x < 0.001$), less professional ($X^2 = 90.97$, $p < 0.001$), and funniest ($X^2 = 123.39$, $x < 0.001$). Following analysis of listener judgments of voices, I conducted *post-hoc* analysis comparing voice quality (VQ) features to examine which may trigger the listener judgments of race. Using PraatSauce, I employed scripts to extract VQ measures previously hypothesized to pattern differently in African American English vs. Mainstream American English. VQ measures that significantly affected listener ratings of the voices are mean F0 and H1–A3c, which correlate with perceptions of pitch and breathiness. These results reveal listeners attribute human-like demographic and personal characteristics to synthesized voices. A more comprehensive understanding of social judgments of digitized voices may help us to understand how listeners evaluate human voices, with implications for speech perception and discrimination as well as recognition and synthesis.

KEYWORDS
voice assistants, voice quality, ethnic identification, language and race, speech perception

1. Introduction

As the line between real life and online interactions becomes increasingly blurred, researchers seek to understand how linguistic production and perception may operate in digital spaces. In particular, understanding how humans interact with computational systems such as voice assistants, will likely become increasingly important for improving the function and fairness of the technologies as well as describing how language change may proceed in a digital world. Additionally, recent work has also begun to examine sociolinguistic questions related to bias and algorithmic inequality. According to Apple, Siri voice assistant fulfills more than 25 billion requests each month worldwide (Eckel, 2021). Siri debuted with a single U.S. English voice in 2010, but by April of 2021, U.S. users were presented with a choice of four different American English Siri voices[1]. While these four voices were undergoing beta testing, users on Twitter began to comment that two of the voices had racial identities, noting in particular that one of the voices "sounded black" (Waddell, 2021). The linguistic research on the perception of racialized voices, and especially voiced that are classified as "sounding Black" has been fairly robust since the 1950s, but such perceptions of synthesized voices as having racial identities have not yet been studied in sociolinguistics. The current study therefore aims to address the following questions:

- Do listeners hear the different Siri voices as sounding like speakers with different demographic characteristics, including region of origin, age, and race?
- What personality traits do listeners associate with the different Siri voices?
- What voice quality features may be associated with different listener judgments of voice assistants like Siri?

The paper is structured as follows. First, it begins with a discussion of how previous studies have examined listener judgments of race when presented with human voices. It then moves on to discuss work on the perception of synthesized voices in general. Subsequently, the methods and results of the perception study on how listeners evaluate the four Siri voices are discussed. The paper then presents a *post-hoc* analysis of voice quality features that correlate with these listener judgments, and finally concludes with a discussion of the results and their impacts for our understanding of racialized perception and interaction with voice assistants.

1.1. Perception of racialized voices

The question of how listeners engage in racial and ethnic identification of human voices has been explored in sociolinguistic research in the U.S. for over 60 years. In a comprehensive review article, Thomas and Reaser (2004) provide an overview of studies on ethnic identification up to that point, but research in this area has continued to grow over the past two decades. Despite intense interest, sociolinguists still have a number of remaining questions about the mechanisms by which listeners make racial and ethnic judgments, especially about disembodied voices. In general, the research has shown that American English listeners are extremely adept and accurate at racial identification tasks, even when presented with stimuli that have been filtered, or that consist of very little phonetic information, such as a single vowel. Thomas and Reaser (2004) discuss the results of 30 studies conducted between the early 1950s to the early 2000s that examined whether American English speakers can reliably differentiate Black and white speakers. In general, the studies they discuss find rates of accurate racial identification over 75%, and in some cases, over 90%. studies These studies employ a wide variety of speaker populations and methodologies and focus on different aspects of the linguistic signal that may be involved in speaker judgments. This is particularly of interest due to the fact that many of the studies have found accurate judgments on the basis of stimuli as limited as one vowel, indicating that the evaluations are about the properties of the voice itself as opposed to enregistered morphosyntactic features of ethnolinguistic varieties. As a result, the current discussion will focus primarily on those studies that aimed to understand the prosodic features that listeners rely on in forming such judgments. However, it is important to note that the relationship between segmental phonological features and prosodic features in perception is not well-understood, and has not been examined at all in studies of ethnic identification [though see Clopper and Pisoni (2004) and for discussion of the relationship in MAE]. In one of the earliest studies, Koutstaal and Jackson (1971) found that listeners were over 80% accurate in their racial identifications, though they were more accurate with white speakers than Black speakers. Importantly, this result demonstrates that listener confusion in such studies is not necessarily bidirectional, which is a consistent finding across many of the studies discussed by Thomas and Reaser (2004). Koutstaal and Jackson (1971) is particularly of interest because the study specifically observes intonation and timing differences between groups, though the author does not actually claim that these differences trigger the differences in listener judgments.

Purnell et al. (1999) found that naive listeners were over 70% accurate in differentiating African American English (AAE), Chicano English and Standard American English (SAE) guises on the basis of one word (in this case, "hello"). This study is unusual compared to others in the vein of ethnic identification, because it utilizes one speaker employing three different dialect guises. While there are certainly downsides to such a methodology, it has the advantage of being particularly useful for researchers interested in the phonetic properties that listeners may rely in making such judgements. The fact that the speaker was held constant across guises means that differences in voice quality features related to recording environment and the speaker's physical characteristics are significantly reduced, which was a persistent challenge for earlier studies. As a result, Purnell et al. (1999) are isolate a number of voice quality features that they believe may be involved in triggering listener judgments of ethnicity. Ultimately, the authors concluded that harmonics to noise ratio, select formant measurements, F0 peak and vowel duration play a significant role in influencing listener judgments. Of particular interest for the current study is the fact that the authors observe that lower HNR differentiates the AAE guise from the SAE guise, indicating a possible role of phonation type in influencing guise identification.

In addition to providing a summary of earlier work, Thomas and Reaser (2004) also conduct an experiment testing test ethnic identification among North Carolina speakers and listeners in original, monotonized, and low pass filtered stimuli. They find high levels of accuracy for the original and monotonous treatment among listeners, and close to chance results for the low-pass filtered condition. As a result, they conclude that manipulation

[1] As of this writing in November 2022, there are now 5 Siri voices. The fifth voice debuted in Spring 2022, and was specifically marketed as sounding "non-binary".

of intonational contours does not significantly reduce listener accuracy, but that eliminating high portions of the acoustic signal containing segmental information does. These results provide evidence that perhaps listeners are especially attuned to segmental information and less attuned to intonational features, however this study does not specifically examine the role of voice quality or phonation. Overall, these studies show that both voice quality and segmental information likely play a role in listener judgments of race, but the specific features that may be involved are not yet well-understood. This is likely due at least in part to the methodological and technical difficulty of isolating variables from one another, as well as controlling natural variation in the speech signal due to speaker or recording quality properties. As a result, the role of suprasegmental features in ethnic identification is still not well-understood. The *post-hoc* analysis in the current study aims to expand our understanding of the contributions of these features, employing a novel method that controls for both recording quality and speaker variation that utilizes synthesized voices.

1.2. Perception of synthesized voices

Recent studies in the realm of linguistics and human-computer interaction have aimed to describe how humans respond to synthesized voices with different types of pseudo-demographic characteristics. This research complements more traditional linguistic work on topics such as speaker identification and ethnic identification by testing not only new types of voices, but also introducing a greater level of control over the properties of the voices used in such experiments. In general, these studies have found that listeners due attribute demographic and personality-type characteristics to both synthesized and natural voices. Additionally, listeners in studies of synthesized voices also reproduced the types of social biases that researchers have observed in studies of natural voices, especially with respect to gender. For example, several studies have found that humans are more likely to be abusive to digital assistants with female names and voices than those with male names and voices (Penny, 2016; Fossa and Sucameli, 2022). Similarly, Jackson et al. (2020) found that listeners judge "female-sounding" assistants more harshly than "male-sounding" robots when they do not comply with user directions, indicating gendered expectations about robot compliance (Jackson et al., 2020). While less work has been conducted on how listeners respond to voices that are evaluated differently based on perception of race, the results from these studies focused on gender provide evidence that listeners utilize social information to respond to voices, even when they are aware that the voices are non-human (Tamagawa et al., 2011; Baird et al., 2017).

Though the evidence is robust that humans readily attach social information to voices in both real-life and experimental situations, the specific linguistic criteria that are involved in such judgments is still not well-understood. Examining how listeners make such judgments on synthesized or partially synthesized voices provides a promising new area for social perception of voices. To begin with, speech synthesis technology has now advanced to the point that listeners can be deceived about whether they are hearing a natural or synthesized voice, allowing us to control for the effects of naturalness (Kühne et al., 2020). More importantly however, synthesized voices allow researchers to tightly control micro-level variation in the realms of intonation and voice quality, which is nearly impossible for naturalistic speech produced in the real-world, due to noise and the extreme level of both vocal control and metalinguistic awareness that would be required to elicit precise productions from humans. Synthesized speech therefore allows us to test and create stimuli that are more tightly controlled than the types previously employed in judgment tasks with natural human voices. In this way, we can isolate specific variables in order to arrive at a better understanding of which of them are most important for triggering social judgments on the part of listeners.

The current study focuses on Apple's proprietary voice assistant, Siri. Siri has undergone numerous updates and changes since it debuted in 2010, generally trending in the direction of more user options for Siri's voice. When Siri was first introduced, the only available American English voice was female, with an American English male voice later added in 2013, in part as a result of user complaints about gender stereotyping (Bosker, 2013). From 2013 to 2021, Apple's two options for Siri in the U.S. were explicitly labeled "American English female" and "American English male." With the April 2021 upgrade, these voices were renamed, with the former "American English male" voice now labeled as "Voice 1," and the former "American English female" voice now labeled as Voice 4. Voices 2 and 3 also debuted at this time, and while Apple never explicitly provided them with gendered labels, the introduction of the new "gender-neutral" Siri voice option, "Quinn," in 2022, reinforced user claims that the previous 4 voices were explicitly gendered. As a result, the current study does not focus on gender, because unlike for other demographic characteristics, Apple explicitly stated the gender of the American English Siri voices prior to April 2021. During beta testing of the new 4-voice Siri paradigm introduced in 2021, users and the media began to express strong social impressions of the voices, especially the new options, Voice 2 and Voice 3. In a 2021 article in Consumer Reports, Waddell reports that some Twitter users explicitly labeled the new Voice 2 and Voice 3 as "sounding Black." The perception study reported in the next section aims specifically to test claims about the demographic and personal characteristics that users attribute to each of the 4 Siri voices in order to better understand listener perceptions of digital voices. The study then builds on the results of that perception study to explore which voice quality features may be involved in triggering such judgments, which will help researchers and the public gain a better understanding of the properties of the voice involved in ethnic identification.

2. Methods and analysis: Listener perception of Siri voices

2.1. Methods

In order to address the question of what types of social and personality judgments listeners make about the 4 Siri voices, I designed a survey-based experiment, presented *via* Qualtrics. The study was conducted over 1 week in April 2021, while Apple's new Siri voices were still in the beta-testing stage in order to reduce the likelihood that listeners would recognize the voices. 485 listeners were recruited *via* the platform Prolific, which is designed to allow researchers to obtain high-quality research participants with specific demographic characteristics (www.prolific.co). All 485 participants were speakers of American English residing in the U.S. at the time of the survey. Prolific provides detailed demographic information about participants, which also allows researchers to examine potential effects of participant race, gender, region, etc. The listener group was composed of 50% participants who identified as female, 48% as male, and 2% as Non-binary or Other. For Race, 70% of listeners identified as white, 8% as Asian, 7% as Black, 6% as Multiracial, 5% as Latino/a/x, and the remaining 4% as Other. 27% of listeners were from the Northeast, 23% from the Midwest, 21% from the Southeast, 20% from the Southwest, and 9% from the Northwest.

Participants were told that they would be participating in a survey about how people react to different voices, and following the completion of a consent form, they heard each of the 4 Siri voices reading the Rainbow Passage (Fairbanks, 1960), in randomized order. Listeners were initially asked if they heard the clip well, and then were permitted to play the clip as many times as they wanted. Following the presentation of each voice, listeners were asked to respond to questions about the voice's race, region, and age (as categorical) and its personal characteristics (as 7-point Likert scales), following the methods employed by Holliday and Tano (2021). Participants were compensated at a rate of $7.50 per hour upon completion of the study, using Prolific's built-in payment methods.

After the study data was collected, analysis was conducted using a series of logistic and linear mixed effects regression models of normalized ratings for each property in order to assess the interaction of voice and demographic property/personality trait. The final models contained normalized demographic ratings by voice with main effects and interactions of the listener traits and random intercept per listener. I also then conducted a likelihood ratio test for omnibus testing of the demographic properties. With respect to the demographic characteristics of the voices, models for age, region and race all showed that listeners evaluated the voices differently from one another. However, listener demographic characteristics (including age, race, and region) had no significant effects for any of the models, so the results presented here will demonstrate overall judgments. Results for each demographic characteristic will be presented in turn, followed by the results for the ratings of personality traits.

2.2. Results of perception experiment testing demographic judgments by voice

2.2.1. Region

Overall, listeners are predisposed to rate the voices as more likely to be from the Northeast or Midwest than the Southeast, Southwest, or Northwest U.S. This may be in part a result of the fact that since the Siri voices all read the same passage, there is no morphosyntactic variation available for the listeners in their evaluations. As a result, they must rely primarily on prosodic and segmental phonological information in their evaluations. Varieties of English spoken in the Southwest and Northwest and somewhat less enregistered than those of other regions, and the Midwest is frequently ideologically painted as more "neutral" or general," so in the absence of salient morphosyntactic variation, participants may be more likely to default to the less marked varieties (Carmichael, 2016). Figure 1 shows the results for listener judgments for each voice's region, with error bars representing the standard error.

There are some significant and informative differences between region judgements for the four Siri voices. 37% of listeners rated Voice 2 as from the Northeast, while none of the other voices had ratings that were significantly different between Midwest and Northeast. Listeners displayed a higher rate of confusion for Voice 3 than any of the others, with judgements fairly split between Northeast (22%) and Midwest (23%), with slightly fewer participants selecting Southwest (17%). Of particular interest for Voice 3 however, is the fact that it was significantly more likely to be labeled as "Southeast" (32%) than any of the other voices (<8%). This result will be discussed in greater detail below in connection with the age and race ratings for Voice 3.

2.2.2. Age

Listener judgments for age are skewed toward the younger options presented in the survey, a result that has also been observed in other studies of synthesized voices (Baird et al., 2017). Overall, fewer than 5% of participants labeled any of the voices as over age 45. For the three age groups 18–25, 26–35, and 36–45, we do observe some differences between the 4 Siri voices. While all of the voices are most likely to be rated as age 26–35, Voices 1 and 2 are significantly less likely to be rated as 18–25 than as 26–35 or 36–45. Voice 4 is equally likely to be rated as 18–25 and 36–45, but 56% of listeners rated it as age 26–35. As with the results for region, Voice 3 is somewhat of an outlier. While Voice 3 is most likely to be rated as 26–35 (50%), it is also disproportionately likely to be rated as 18–25 (37%) when compared to the other 3 voices, as can be seen in Figure 2.

The fact that the ages attributed to each voice differs somewhat is also informative. In particular, a picture is beginning to emerge such that Voice 3 patterns differently from the other 3 voices. Also, of note here is that fact that Voice 2 is the least likely to be rated as age 18–25, indicating that this voice heard as somewhat more mature. However, Voice 2 does pattern with voices 1 and 4 in terms of being likely to be rated as either 26–35 or 36–45, giving them roughly the same mean ratings for age.

2.2.3. Race

The original motivation for the survey was the claim by some users that Voices 2 and 3 "sounded Black," so racial judgments are of particular interest for the current study (Waddell, 2021). The categorical options presented to listeners for race judgements were Asian, Black, Hispanic/Latino, Multiracial, and White. Of interest is the fact that listeners are overall biased toward selecting Black or White, mostly ignoring the other categories. This is perhaps unsurprising given results of previous studies showing a persistent bias among Americans for imagining race as binary (Alcoff, 2003; Kushins, 2014). Figure 3 shows the results for race ratings of each voice. Note that the category of "multiracial" is excluded from this

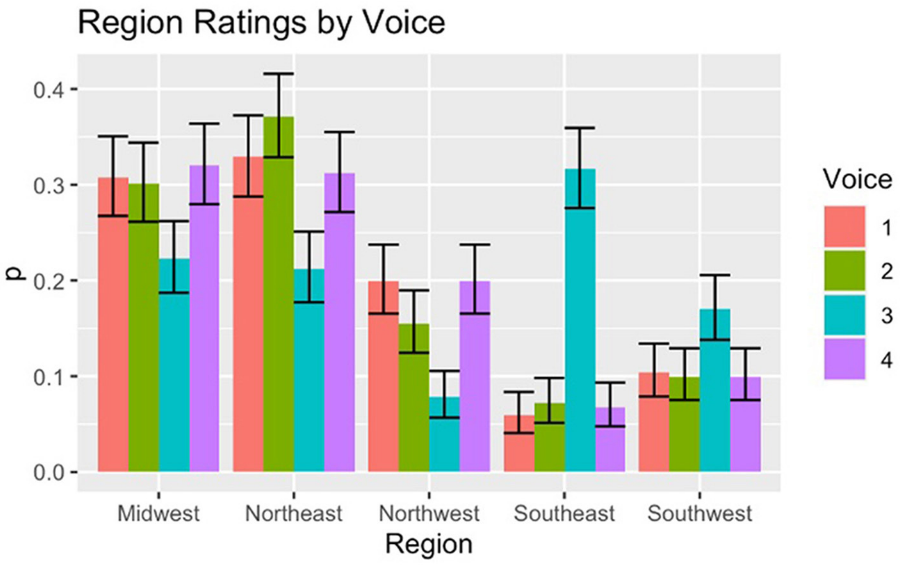

FIGURE 1
Ratings for region by voice.

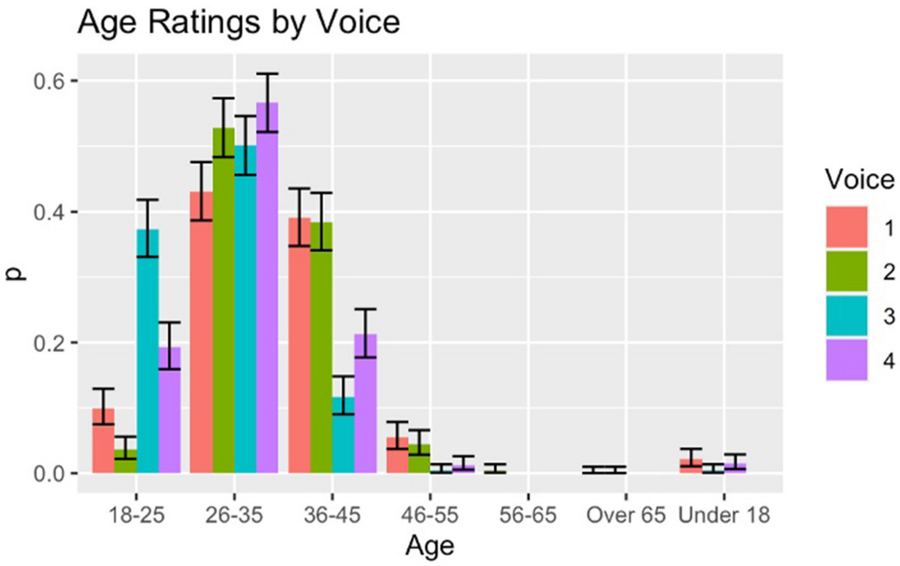

FIGURE 2
Ratings for age by voice.

plot; of the study's 485 listeners, only 8 ever selected the multiracial category for any of the voices.

When compared to the results for the ratings of other demographic traits, listeners are less divided on their ratings of race than they were for region or age. None of the listeners utilized the categories of Asian or Latino/Hispanic for any of the voices at a rate higher than 8%, likely due to the aforementioned bias toward racial binary categorization. With respect to the likelihood of the voices being rated as Black, only 3% of raters said that Voice 1 sounded Black, and only 1% of raters said that Voice 4 sounded Black. 73% of raters said that Voice 1 sounded white, and 78% of raters said that Voice 4 sounded white. Recall that Voice 1 was previously named "American English Male" and Voice 4 was previously named "American English Female," while Voices 2 and 3 were newly introduced.

With respect to the voices being rated as Black. Voices 2 and 3 pattern quite differently than the other two voices. 37% of listeners rated Voice 2 as Black, while 42% rated Voice 3 as Black. While each of these numbers is still slightly lower than the probability of Voice 2 and Voice 3 being rated as White, the fact that the ratings for race pattern so differently for these Voices than they do for Voices 1 and 4 is informative. Overall, we observe a pattern such that these newly

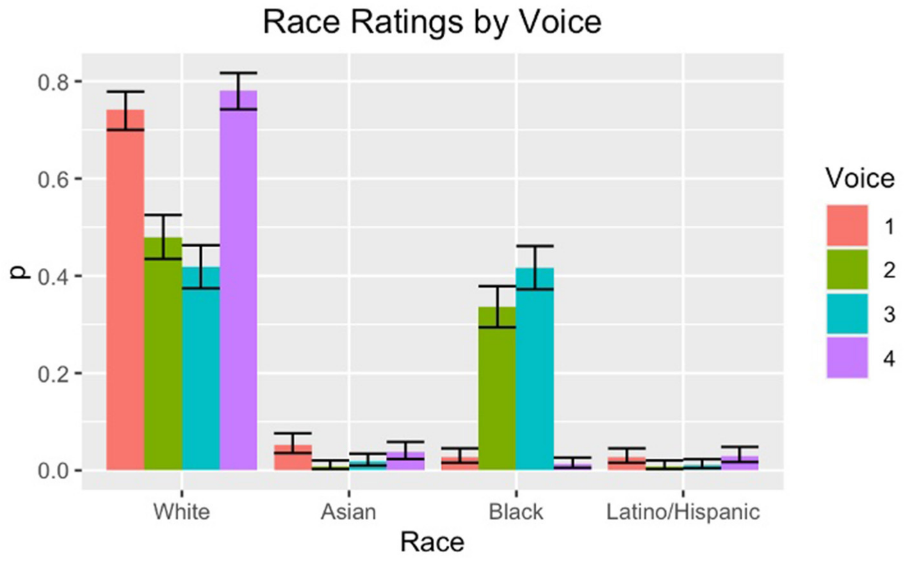

FIGURE 3
Ratings for race by voice.

introduced voices are much more likely to be rated as Black than the older Siri voices.

2.2.4. Summary of demographic characteristics for the voices

Overall, listeners are predisposed to rate the four Siri voices is from the Midwest or Northeast, aged 18–45, and white. However, the differences in the probabilities of ratings for region, age, and race between the 4 Voices does reveal that listeners as a group react to and evaluate the voices in different ways. Table 1 shows the overall results for the demographic ratings of each voice as well as Apple's gender categorization for them. Where there was no clear majority between choices for any given category, both categories are displayed alongside the percentage of listeners who chose each option.

Voice 3 is significantly more likely to be rated as from the Southeast than any of the others, likely overlapping with the fact that it is also the voice most likely to be rated as Black. Sociolinguists have documented significant feature overlap between many Southern White varieties and African American English, as result of the fact that AAE originated in the South (Wolfram, 2007). Additionally, African American English is often inaccurately stereotyped as youth slang in the public imagination, which is also likely a factor contributing to Voice 3's judgments as younger than the others (Green, 2002). Overall, listeners have a markedly different reaction to Voice 3's demographic properties than any of the other voices. When compared with Voices 1 and 4, listeners also demonstrate significantly more ambiguity in their judgments of Voice 2. Voice 2 is significantly more likely to be rated as "Black," but also as from the Northeast. Overall, these results demonstrate that listeners judge the demographic properties of Voices 2 and 3 differently than Voices 1 and 4, in particular, being much more likely to rate them as Black.

2.3. Results of perception experiment testing personality traits by voice

2.3.1. Personal characteristics by voice

In order to arrive at a more holistic picture of how listeners may evaluate the different Siri voices, the survey also asked them to rate each voice for four personal characteristics, on a 7-point Likert scale. The characteristics of interest were friendliness, professionalism, funniness and competence following the methods of Holliday and Tano (2021). Analysis of this data was conducted *via* per-trait regression models and chi square tests. In general, there are few differences between how the voices are rated for each trait, with most participants rating all of the voices fairly highly on all 4 characteristics. Figure 4 shows the results for per-trait ratings for each voice.

The primary difference in ratings, according to the per-trait logistic regression models and chi-square tests, is that Voice 3, the male voice rated as black and younger, was rated less competent ($X^2 = 108.99, x < 0.001$), less professional ($X^2 = 90.97, p < 0.001$), and funnier ($X^2 = 123.39, x < 0.001$). Interestingly, this is a pattern similar to what we observe when listeners are asked to rate human voices; they display a persistent negative bias against Black voices for traits related to competence, but usually a positive bias for traits related to sociability (Kushins, 2014; King et al., 2022).

One additional finding of interest emerges, with respect to Voice 1 and listener ratings of friendliness. The other 3 voices were rated similarly for this trait, but Voice 1 was rated significantly lower ($X^2 = 101.97, x < 0.001$). Voice 1 was the voice previously known as "American English Male," and was also rated by the participants in this study as likely to be white (73%). Section X discusses one hypothesis for this lower rating of friendliness related to voice quality, but another possible hypothesis for these ratings is also listener expectations of "male" digital assistants differ from those of "female" assistants (Jackson et al., 2020). Finally, it is also

TABLE 1 Most frequent listeners ratings by gender/region/age/race for each Siri voice.

Voice	Gender (from Apple)	Region judgment	Age judgment	Race judgment
1	Male	Midwest/northeast	26–35/36–45	White
2	Not official, but implied female	Midwest/northeast	26–35 (52%), 36–45 (42%)	Black (34%), white (48%)
3	Not official, but implied male	Southeast	26–35 (50%), 18–25 (38%)	Blacek (42%), white (43%)
4	Female	Midwest/northeast	26–35 (56%)	White

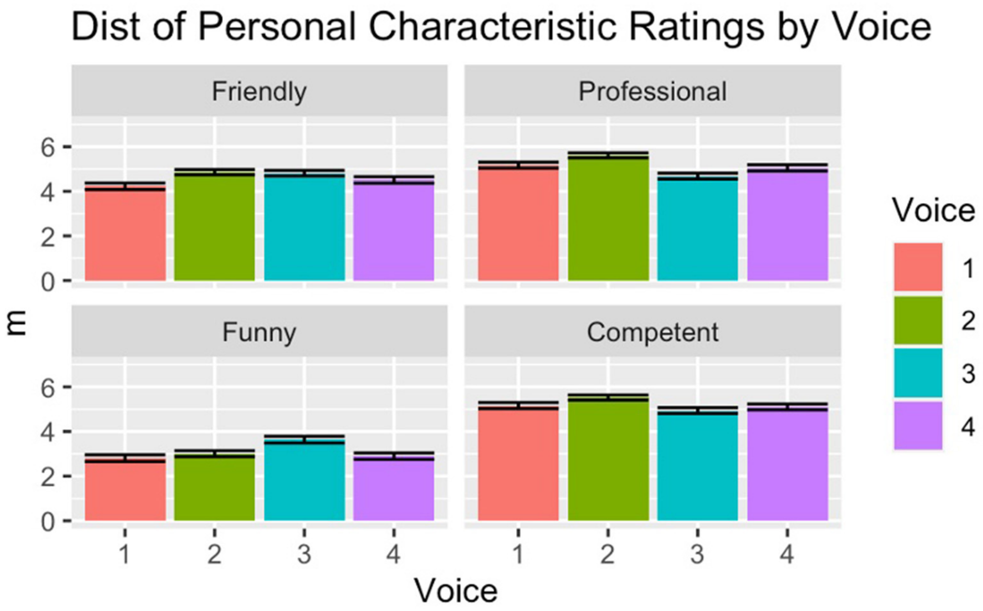

FIGURE 4
Ratings for personal characteristics by voice.

notable that when these models account for listener, age, race, gender and region, no significant differences emerge, indicating that black listeners were different than white listeners in terms of having more bias toward these personal characteristics.

As a result of these analysis of listener demographic ratings and personality trait ratings for the 4 Siri voices, we can be confident that listeners do robustly engage in racialized judgments of digital voice assistants. Additionally, we observe that listener racial judgments do interact with perceptions of the personality of the voice, further demonstrating that listeners personify these voices and attach human-like stereotypes to them. However, we still do not know which specific linguistic features are involved in triggering such listener judgments. The next section presents the results of a *post-hoc* analysis of the relationship between listeners judgments and voice quality features of the 4 Siri voices.

3. Voice quality analysis

3.1. Methods: Voice quality

Having established that listeners do in fact make systematically different demographic judgments about the four Siri voices, the next section aims to explore which voice quality (VQ) features may be involved in triggering such listener judgments. While it is important to note that listeners likely integrate both segmental and prosodic features in their judgments, the current study's analysis will be limited to VQ features, given that synthesized voices in a reading task may be more limited in their ability to display segmental variation especially as the consonantal level. VQ properties are also especially of interest in the current study due to the fact that they may be involved in judgments of voices that are totally independent of segmental phonological or morphosyntactic features, which may be more likely to operate at the level of conscious and/or have enregistered social stereotypes (Labov, 1971). Indeed, as Garellek (2022, p. 2), observes "The reason why the voice often takes center stage in phonetic research is because it is everywhere and matters for everything in the phonetic signal." For this reason, we can observe that features related to properties of the voice provide important information for how listeners evaluate speakers. However, to date, little work has examined how voice quality features may differ for synthesized voices, or how listeners may react to VQ properties of synthesized voices. Generally, voice quality features are underdescribed in part due to the fact that they frequently behave in a colinear fashion and are not fully theorized with respect to sociolinguistic variation. The current paper therefore represents a first pass at examining how VQ features may affect perceptions and judgments of such voices.

As discussed in Section 1.1, previous research on ethnic identification and production differences between Black and white speakers has posited general differences in voice quality between the groups, but to date, little is still known about the specific VQ parameters that may be involved. In general, such studies have posited that there may be differences related to F0 and the perception of pitch, as well as variable use of different phonation types (creaky, breathy, and modal) (Purnell et al., 1999; Thomas, 2015). In order to examine a maximally broad set of VQ parameters, I employed the PraatSauce suite of scripts, which is designed to extract a variety of spectral measures from acoustic data (Kirby, 2018). Praatsauce extracts 34 features that are related to VQ, a useful technique for the current study given that we want to conduct an exploratory analysis. The Praatsauce scripts take measurements by dividing each vowel in the passage and dividing it into five parts with equal duration. Measurements are then made at five points by averaging value (for each measure) of that section. Since all 4 voices in the current read the same passage and in the same room during the same 15-min interval, the sample is already internally controlled for vowel identity and coarticulatory effects, as well as external recording noise. The full list of VQ features extracted from the speech signal by the PraatSauce scripts are listed in Supplementary Appendix A, but in general, the features of interest are the harmonic amplitude components from the low-, mid-, and high-frequency regions of the signal (H1, H2, A1, A2, A3, H2k, and H5k), cepstral peak prominence, and harmonic and amplitude differences, following the methods of phonetic studies such as DiCanio (2009) and Garellek (2019).

3.2. Analysis and LASSO regression results

A major difficulty of studying voice quality parameters is that the sheer number of variables that may be involved in theoretically infinite, so this type of analysis requires statistical method that can handle both variable selection and regularization when variables behave in a colinear fashion. One way to resolve this challenge is *via* the use of a Least Absolute Shrinkage and Selection Operator (LASSO) regression. For the current study, I conducted this using the GLMnet package in R (Friedman et al., 2010). LASSO regression provides a model that improves prediction accuracy, and decreases variance by shrinking or regularizing the coefficients, effectively relying on penalties in order to prevent overfitting. The benefit is that it allows us to fit a model containing all possible predictors and use lasso to perform variable selection that simultaneously chooses a set of variables and regularizes their coefficient estimates.

LASSO works by selecting a tuning parameter (lambda) which is chosen by cross-validation. When lambda is zero, the estimates are the same as the ordinary least squares (OLS) and as lambda increases, shrinkage occurs, and variables set at zero that do not contribute to the fit of the model can be excluded. LASSO does both shrinkage and variable selection so if we have a large number of features, we can better find the model with the best fit. Figure 5 shows the LASSO tuning plot of the model for race rating by VQ parameters.

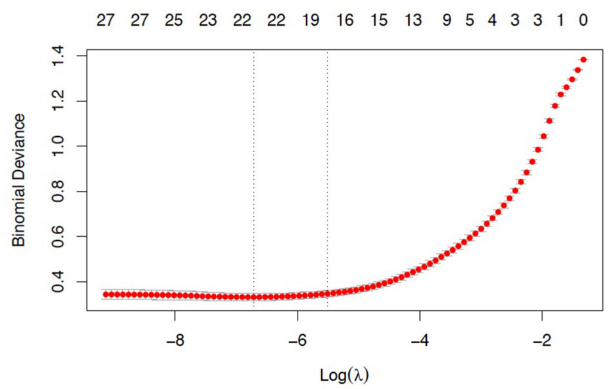

FIGURE 5
LASSO regression model for voice quality features by race rating.

The x axis is log of lambda, which corresponds to the minimum MSE and one standard error from that, and those are shown by the vertical lines. In this output, the best fit is between those with lines, and we see that with increasing lambda there are fewer variables in the model, because the penalty for inclusion starts to become weighted more heavily. From this plot, we can observe that the model with the best fit likely contains between 19 and 23 variables. However, we can also observe that 2–3 variables account for nearly half of the model fit. Figure 6 displays the cross-validation plot showing the contribution of the variables of interest.

Overall, the results of the LASSO regression show differences between the race ratings for the voices and the selected VQ features. Lower F0 and higher H1–A3c[2] correlate with the voices rated as black, especially Voice 3, but also Voice 2. F0 here is fundamental frequency, which is the main correlate of what humans perceive as pitch. H1–A3 is the difference between the amplitude of the first harmonic (H1) and the harmonic nearest F3. This correlates with the abruptness of the glottal closure. Generally, a larger positive value is perceived as more breathy, and a lower value is more creaky. Results for F0 and H1–A3 will be discussed in turn.

3.2.1. F0

As the LASSO output selected F0 and H1–A3c as the primary variables involved in differentiating the voices rated as Black vs. those rated as white, it is important to better understand the differences in mean F0 between the voices in order to interpret these results. PraatSauce measures F0 at 5 time points per vowel per phrase, so these measurements provided the F0 input for the model. Figure 7 shows the mean F0 for each of the four Siris.

From Figure 7, we can see that the outlier voice in terms of F0 is Voice 4, which has a mean F0 of 210Hz. Voice 4 is the voice that debuted in 2010 and was formerly referred to as "American English Female," and this F0 value is close to what previous studies have reported for white American women (Bradlow et al., 1996; Pépiot, 2014), and thus may not be especially surprising. However, the difference between Voice 4's F0 mean and the other "female" voice, Voice 2, is striking. Voice 2 has an F0 mean of 155 Hz, and

[2] H1–A3c here is the measurement corrected for the third formant.

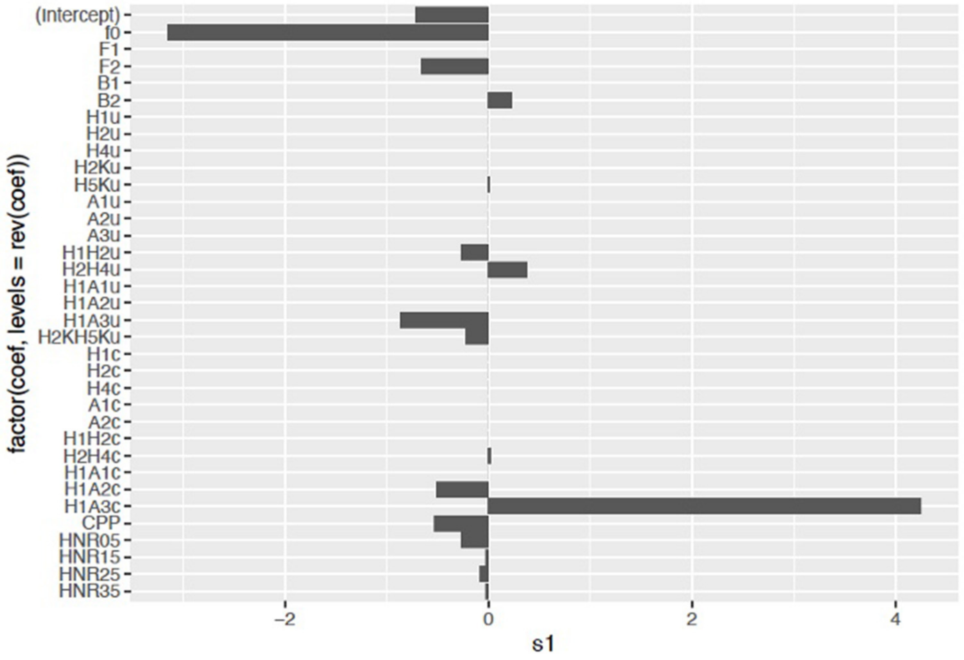

FIGURE 6
Factor plot for LASSO regression model of voice quality features by race rating.

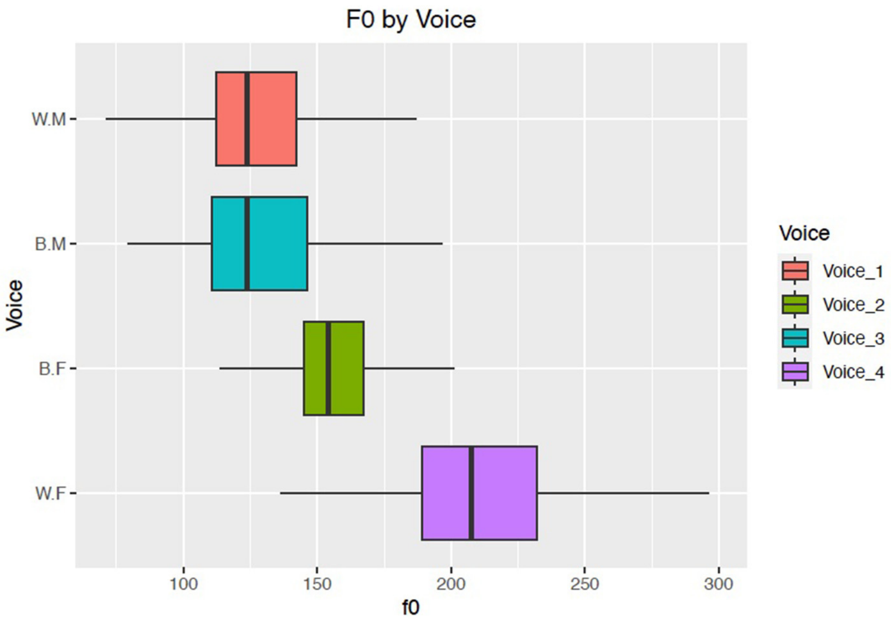

FIGURE 7
Mean F0 values by voice/race rating.

its mean values overlap with those of Voice 3, the "male voice" who was most likely to be rated as Black. The F0 mean values for Voice 1 and Voice 3, the two "male" voices, are not significantly different from one another. These results provide compelling evidence that listeners may be attuning specifically to the interaction between low F0 and gender for voices that they perceive as "female," but not for those that they perceive as "male." Recent work by Holt and Rangarathnam (2018) and Li et al. (2022), finds that in some samples of Black American female voices, speakers do generally employ a lower F0 mean value than white female voices. If these differences do exist in production and are perceptually salient for listeners, then they may also influence listener judgments of race, in part contributing to judgments of Voice 2 as more likely to be Black than Voice 4.

3.2.2. H1–A3c

The other VQ parameter selected by the LASSO regression as correlated with the voices more likely judged as Black was H1–A3c, which is related to perceptions of breathiness. According to DiCanio (2009), H1–A3 is a "mid-range measure of spectral tilt which involves a calculation of the amplitude of the different formants" (168), in this case the amplitude of the first harmonic minus the amplitude of the harmonic nearest the third formant (corrected for the formant). Crosslinguistically, H1–A3 has been shown to distinguish breathy from modal phonation (Esposito, 2010, for Chong) as well as creaky from modal phonation. Articulatorily, a high H1–A3 captures lax vocal folds, but increased H1–A3 can also be caused by more active thyroarytenoids in vocal fold vibration (Maddieson and Ladefoged, 1985). If indeed there are average differences in this measure between Black and white speakers, then the fact that the synthesized voices show different values for this parameter may an attempt to capture some differences in vocal fold position that Siri's designers have observed between groups. Generally, we should expect a higher H1–A3 for female voices due to articulatory motivations: for female speakers, the thyroid notch in cartilage is more rounded and does not lower during puberty, resulting in a less complete vocal fold closure when arytenoids are adducted. From a sociolinguistic perspective, lax vocal fold quality has been associated with agreeableness and warmth (Kreiman et al., 2008). Additionally, Babel et al. (2010) found that higher H1–A3 has been found to be rated as more attractive in American English, providing some evidence that listeners do have impressions about social information related to this variable. Figure 8 shows the mean H1–A3 values for the four voices.

Interestingly, Voices 2, 3, and 4 overlap with respect to measurements on this parameter, but Voice 1 has a significantly lower H1–A3c measurement. Taken together with the personality trait results in Section 2.3.1 showing that Voice 1 was also rated lower on friendliness, we may hypothesize that less perception of breathiness for this voice is related to the judgment as less friendly. Additionally, since H1–A3 correlates with breathiness, it is possible that listeners simply expect less breathiness from a voice that they perceive as belonging to an allegedly typical male speaker (Ishi et al., 2010). Indeed, Gobl and Chasaide (2003) specifically mention that male voices are less likely to exhibit breathiness, and that less breathy voices are perceived as less friendly.

4. Summary discussion

The results of the perception study about four different Siri voices demonstrate that listeners do robustly engage in regional, age, and racialized judgments of digital voice assistants. In particular, American English listeners from a variety of racial and regional backgrounds pattern similarly in their perceptions of these voices, but the voices are evaluated differently from one another in important ways. With respect to region, listeners generally judged all four voices as likely to be from the Midwest or Northeast, but showed much more variation in their ratings of the two newer voices, Voice 2 and Voice 3. Voice 2 was rated as most likely to be from the Northeast, and Voice 3 was most likely to be rated as from the Southeast. The ratings for age showed that Voice 3 was also the most likely to be rated younger than the other three voices, which were generally judged as belonging to a speaker aged 26–45. Finally, listeners overwhelmingly rated the original Siri voices, formerly called "American English Male" and "American English Female" as white, and showed much more variation in their racial ratings of the two newer voices that debuted in 2021. In particular, listeners were unlikely to rate any of the voices as Hispanic/Latino, Asian, or Multiracial, but much more likely to rate Voice 2, and especially Voice 3, as Black. Taken together, the results for the newer Voices 2 and 3 show that they are significantly more likely to be perceived as coming from people of color than the two older voices.

Racial judgments also interact with perceptions of the personality of such assistants, further demonstrating that listeners holistically personify these voices. While the four voices were generally rated similarly on friendliness, competence, professionalism, and funniness, the outlying judgments are revealing. Voice 1, the former "American English Male" who was rated as white, was also overall rated less friendly than the other three voices, possibly in part due to his use of lower values for the VQ parameter H1–A3, a correlate of breathiness that has been previously discussed as associated with friendliness. However, the results for the personality traits of Voice 3 are particularly of interest, especially when combined with the ratings for his demographic features. Overall, Voice 3 was rated as youngest, most likely to be from the Southeast, most likely to be Black, and less professional and competent, but funnier. When combined, these ratings give us a richer idea of who the listeners may imagine Voice 3 to be: a young, Black man from the Southeast who is funny but not especially competent or professional. This persona, the underachieving, regionally disenfranchised young Black man, is a well-worn trope in U.S. media depictions of individuals who speak African American English (AAE) (Cutler, 2007; Lopez, 2012). It is also a strong stereotype about the kind of person who is imagined to "sound Black" (Baugh, 2005). While Voice 2 was also more likely to be rated as sounding Black, the fact that it is a voice implied to be female somewhat mitigates this judgment, as several studies have found that listeners perform worse in ethnic identification tasks with female speakers (Thomas and Reaser, 2004).

Overall, these results demonstrate that human listeners attach the same types of regional, age, racial, and personality judgements to voice assistants that they do to human voices in previous studies. While this result may be seen as positive in terms of advances in naturalness of digital voice assistants as well as representation of a diverse set of voices in our everyday technologies, it is worrying that even synthesized voices that are perceived as coming from speakers of marginalized backgrounds can be evaluated with the same negative social stereotypes. Future studies should further explore how voice assistants created by different mechanisms and for different user types contribute to the linguistic ideologies of their user base. On a positive note, however, some Black users have reported positive feelings about hearing voices like theirs in digital assistants, demonstrating that tech firms' efforts toward both realistic and inclusive synthesized voices have been somewhat effective (Waddell, 2021). Going forward, researchers and tech

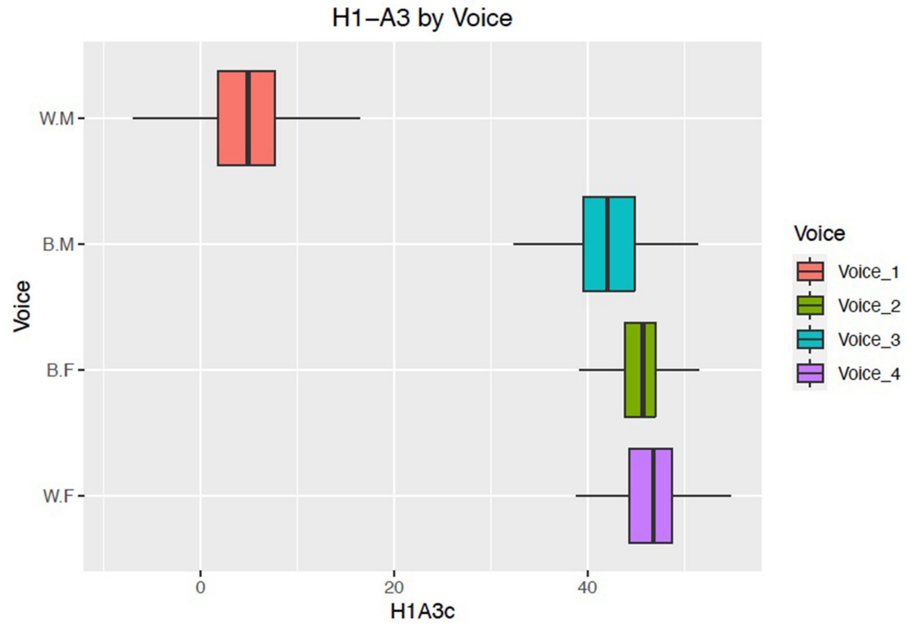

FIGURE 8
Mean H1–A3 values by voice/race rating.

firms should carefully investigate user responses to such voices and make efforts to ensure that they are fairly and accurately representing the voices they aim to synthesize.

Tech firms may also make greater efforts to understand and share acoustic information about synthetic voices; both to better understand how listeners perceive such voices but also to assist researchers in better understanding which properties of the voice may be associated with listener judgments of demographic properties and personality traits. A *post-hoc* analysis of the four Siri voices' use of 34 voice quality (VQ) parameters using PraatSauce (Kirby, 2018) shows that racial judgments in particular are linked with features related to the perception of both pitch and breathiness. The current study identified lower F0 and higher H1–A3 as correlated with judgments of a voice as Black, but many questions remain about the relationship between voice quality parameters and listener perceptions of race. Based on previous research such as Purnell et al. (1999), I expected that features related to perception of breathiness may play a role, but the results did not show significant contributions of parameters such as CPP and HNR. This may indicate that when making racial judgments, listeners attune more to VQ parameters that interact with formant measurements or other segmental phonetic features, but much more work is needed to better understand how listeners attune to these different parameters.

With respect to pitch, there are strong social stereotypes related to Black male voices and lower pitch, though these are not necessarily borne out in production studies that examine pitch as a racialized variable (Li et al., 2022). Indeed, in the current study, Voice 2, the implied female voice most likely to be rated as Black had a lower mean F0 than Voice 4, the female voice labeled as white, but the F0 mean values for the two male voices overlapped. This supports the results of Li et al. (2022) who found that the Black American English female speakers they examined did use lower F0 mean values than the white American English female speakers in their study. Whether or not future studies support the claim that Black American women employ lower F0 mean than white American women, the existence of such a stereotype might still affect listener judgments of race. Future work should examine racialized differences in both production and perception of voices of people and digital assistants of all genders.

The use of synthesized voices in studies on the perception of demographic and personal traits, as well as voice quality properties, represents a new avenue for sociophonetic research on variation in voice quality. Indeed, there is a dearth of sociolinguistic studies on voice quality in general, likely due to the fact that voice quality features demonstrate so much naturalistic instability that it may be hard to distinguish which elements represent group-level variation and which are artifacts of a particular speaker's idiolect or anatomical features. Synthesized voices, however, allow researchers a great deal of control over recording quality and may eliminate variation due to physical properties of the voice altogether, providing researchers with the ability to create a static voice with specific features in order to test the contributions of different voice quality parameters in greater isolation.

A more comprehensive understanding of judgments of digitized voices may help us better examine how listeners make judgments of human voices with implications for a variety of fields including human perception, linguistic discrimination and speech recognition and synthesis. Understanding the specific features that listeners attune to in making racial judgment could be used in future efforts to reduce linguistic bias in avenues such as education and criminal justice. They may also be useful for forensic linguistic purposes such as speaker identification, especially for speakers who have been previously disenfranchised by such technologies. Finally, understanding how voices are evaluated may also assist not only with the creation of future, more authentic and representative digital assistants, but also in the development of more realistic synthetic voices used by humans with vocal disorders.

Data availability statement

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

Ethics statement

The studies involving human participants were reviewed and approved by Pomona College IRB. 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 research was funded by internal grants to NH at the University of Pennsylvania and Pomona College.

Acknowledgments

The author would like to thank Kaveh Waddell of Consumer Reports for inspiration, Neal Fultz for statistical advice, and Christian DiCanio for advice on measuring and interpreting voice quality parameters.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2023.1116955/full#supplementary-material

References

Alcoff, L. M. (2003). Latino/as, Asian Americans, and the black–white binary. *J. Ethics* 7, 5–27. doi: 10.1023/A:1022870628484

Babel, M., King, J., McGuire, G., Miller, T., and Babel, M. (2010). Acoustic determiners of vocal attractiveness go beyond apparent talker size. *Lab. Rep. Linguist. Res. Center Univ. Calif. Santa Cruz* 2010, 1–23.

Baird, A., Jørgensen, S. H., Parada-Cabaleiro, E., Hantke, S., Cummins, N., and Schuller, B. (2017). "Perception of paralinguistic traits in synthesized voices," in *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*, 1–5. doi: 10.1145/3123514.3123528

Baugh, J. (2005). "Linguistic profiling," in *Black Linguistics* (London: Routledge), 167–180. doi: 10.4324/9780203986615-17

Bosker, B. (2013). *Will a Man's Voice Make Siri Better?* New York, NY: HuffPost. Available online at: https://www.huffpost.com/entry/siri-voice-man-woman_n_3423245 (accessed June 12, 2013).

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Commun.* 20, 255–272.

Carmichael, K. (2016). Place-linked expectations and listener awareness of regional accents. *Awareness Control Socioling. Res.* 12, 123–151. doi: 10.1017/CBO9781139680448.009

Clopper, C. G., and Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *J. Phonet.* 32, 111–140. doi: 10.1016/S0095-4470(03)00009-3

Cutler, C. (2007). Hip-hop language in sociolinguistics and beyond. *Lang. Ling. Compass* 1, 519–538. doi: 10.1111/j.1749-818X.2007.00021.x

DiCanio, C. T. (2009). The phonetics of register in Takhian Thong Chong. *J. Int. Phonet. Assoc.* 39, 162–188. doi: 10.1017/S0025100309003879

Eckel, E. (2021). *Apple's Siri: A Cheat Sheet*. San Francisco: TechRepublic. Available online at: https://www.techrepublic.com/article/apples-siri-the-smart-persons-guide/ (accessed November 3, 2022).

Esposito, C. M. (2010). The effects of linguistic experience on the perception of phonation. *J. Phonet.* 38, 306–316. doi: 10.1016/j.wocn.2010.02.002

Fairbanks, G. (1960). The rainbow passage. *Voice Articulat. Drillbook* 2, 127.

Fossa, F., and Sucameli, I. (2022). Gender bias and conversational agents: an ethical perspective on social robotics. *Sci. Eng. Ethics* 28, 1–23. doi: 10.1007/s11948-022-00376-3

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models *via* coordinate descent. *J. Stat. Softw.* 33, 1–22. doi: 10.18637/jss.v033.i01

Garellek, M. (2019). "The phonetics of voice 1," in *The Routledge Handbook of Phonetics* (London: Routledge), 75–106. doi: 10.4324/9780429056253-5

Garellek, M. (2022). Theoretical achievements of phonetics in the 21st century: phonetics of voice quality. *J. Phonet.* 94, 101155. doi: 10.1016/j.wocn.2022.101155

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

Green, L. (2002). A descriptive study of African American English: research in linguistics and education. *Int. J. Qual. Stud. Educ.* 15, 673–690. doi: 10.1080/0951839022000014376

Holliday, N., and Tano, M. (2021). "It's a Whole Vibe": testing evaluations of grammatical and ungrammatical AAE on Twitter. *Ling. Vanguard* 7, 4389. doi: 10.1515/lingvan-2020-0095

Holt, Y. F., and Rangarathnam, B. (2018). F0 declination and reset in read speech of African American and White American women. *Speech Commun.* 97, 43–50. doi: 10.1016/j.specom.2018.01.001

Ishi, C., Ishiguro, H., and Hagita, N. (2010). Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech. *EURASIP J. Audio Speech Music Process.* 2010, 1–12. doi: 10.1155/2010/528193

Jackson, R. B., Williams, T., and Smith, N. (2020). "Exploring the role of gender in perceptions of robotic noncompliance," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 559–567. doi: 10.1145/3319502.3374831

King, S., Vaughn, C., and Dunbar, A. (2022). *Dialect on trial: raciolinguistic ideologies in perceptions of AAVE and MAE codeswitching*. University of Pennsylvania Working Papers in Linguisti.

Kirby, J. (2018). *Praatsauce: Praat-Based Tools for Spectral Analysis.*

Koutstaal, C. W., and Jackson, F. L. (1971). Race identification on the basis of biased speech samples. *Ohio J. Speech Hear*. 6, 48–51.

Kreiman, J., Vanlancker-Sidtis, D., and Gerratt, B. R. (2008). "14 perception of voice quality," in *The Handbook of Speech Perception*, 338. doi: 10.1002/9780470757024.ch14

Kühne, K., Fischer, M. H., and Zhou, Y. (2020). The human takes it all: humanlike synthesized voices are perceived as less eerie and more likable—evidence from a subjective ratings study. *Front. Neurorobot*. 14, 105. doi: 10.3389/fnbot.2020.593732

Kushins, E. R. (2014). Sounding like your race in the employment process: an experiment on speaker voice, race identification, and stereotyping. *Race Soc. Probl.* 6, 237–248. doi: 10.1007/s12552-014-9123-4

Labov, W. (1971). Some principles of linguistic methodology. *Lang. Soc*. 1, 97–120.

Li, A., Purse, R., and Holliday, N. (2022). Variation in global and intonational pitch settings among black and white speakers of Southern American Englisha. *J. Acoust. Soc. Am.* 152, 2617. doi: 10.1121/10.0014906

Lopez, Q. L. T. (2012). *White bodies, black voices: The linguistic construction of racialized authenticity in US film* (Doctoral dissertation).

Maddieson, I., and Ladefoged, P. (1985). "Tense" and "lax" in four minority languages of China. *J. Phonet.* 13, 433–454.

Penny, L. (2016). Why are so many robots given female names? Because we don't want to worry about their feelings. *New Statesman* 145, 38–39.

Pépiot, E. (2014). May. Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers. *Speech Prosody* 7, 305–309. doi: 10.21437/SpeechProsody.2014-49

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

Tamagawa, R., Watson, C. I., Kuo, I. H., MacDonald, B. A., and Broadbent, E. (2011). The effects of synthesized voice accents on user perceptions of robots. *Int. J. Soc. Robot.* 3, 253–262. doi: 10.1007/s12369-011-0100-4

Thomas, E. (2015). "Prosodic features of African American English," in *The Oxford Handbook of African American Language*, ed S. Lanehart (Oxford: Oxford University Press), 420–438.

Thomas, E. R., and Reaser, J. (2004). Delimiting perceptual cues used for the ethnic labeling of African American and European American voices. *J. Sociolinguist.* 8, 54–87. doi: 10.1111/j.1467-9841.2004.00251.x

Waddell, K. (2021). *Hey Siri, Is That You? Apple's New Voices Resonate With Some Black iPhone Users*. New York, NY: Consumer Reports. Available online at: https://www.consumerreports.org/digital-assistants/apples-new-sirivoices-resonate-with-some-black-iphone-users/ (accessed March 23, 2022).

Wolfram, W. (2007). Sociolinguistic folklore in the study of African American English. *Lang. Linguist. Compass* 1, 292–313. doi: 10.1111/j.1749-818X.2007.00016.x

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 28 April 2023
DOI 10.3389/fcomm.2023.1148240

Check for updates

OPEN ACCESS

EDITED BY
Plinio Almeida Barbosa,
State University of Campinas, Brazil

REVIEWED BY
Marzena Zygis,
Leibniz Center for General Linguistics
(ZAS), Germany
Ye Zhang,
University College London, United Kingdom

*CORRESPONDENCE
Saurabh Garg
✉ srbh.garg@gmail.com
Yue Wang
✉ yuew@sfu.ca

RECEIVED 19 January 2023
ACCEPTED 06 April 2023
PUBLISHED 28 April 2023

CITATION
Garg S, Hamarneh G, Sereno J, Jongman A and Wang Y (2023) Different facial cues for different speech styles in Mandarin tone articulation. *Front. Commun.* 8:1148240.
doi: 10.3389/fcomm.2023.1148240

Different facial cues for different speech styles in Mandarin tone articulation

Saurabh Garg[1]*, Ghassan Hamarneh[2], Joan Sereno[3], Allard Jongman[3] and Yue Wang[1]*

[1]Department of Linguistics, Simon Fraser University, Burnaby, BC, Canada, [2]School of Computing Science, Simon Fraser University, Burnaby, BC, Canada, [3]Department of Linguistics, University of Kansas, Lawrence, KS, United States

Visual facial information, particularly hyperarticulated lip movements in clear speech, has been shown to benefit segmental speech perception. Little research has focused on prosody, such as lexical tone, presumably because production of prosody primarily involves laryngeal activities not necessarily distinguishable through visible articulatory movements. However, there is evidence that head, eyebrow, and lip movements correlate with production of pitch-related variations. One subsequent question is whether such visual cues are linguistically meaningful. In this study, we compare movements of the head, eyebrows and lips associated with plain (conversational) vs. clear speech styles of Mandarin tone articulation to examine the extent to which clear-speech modifications involve signal-based overall exaggerated facial movements or code-based enhancement of linguistically relevant articulatory movements. Applying computer-vision techniques to recorded speech, visible movements of the frontal face were tracked and measured for 20 native Mandarin speakers speaking in two speech styles: plain and clear. Thirty-three head, eyebrow and lip movement features based on distance, time, and kinematics were extracted from each individual tone word. A random forest classifier was used to identify the important features that differentiate the two styles across tones and for each tone. Mixed-effects models were then performed to determine the features that were significantly different between the two styles. Overall, for all the four Mandarin tones, we found longer duration and greater movements of the head, eyebrows, and lips in clear speech than in plain speech. Additionally, across tones, the maximum movement happened relatively earlier in clear than plain speech. Although limited evidence of tone-specific modifications was also observed, the cues involved overlap with signal-based changes. These findings suggest that visual facial tonal modifications for clear speech primarily adopt signal-based general emphatic cues that strengthen signal saliency.

KEYWORDS

speech style, Mandarin, facial cues, computer vision, video processing, Mandarin tones

1. Introduction

It is well known that having both audio and video information in a noisy environment, or when talking with non-native speakers or cochlear implant users can help with speech perception and intelligibility (e.g., Sumby and Pollack, 1954; Desai et al., 2008; Wang et al., 2008). In such challenging listening contexts, speakers tend to use a clear, hyperarticulated speech style (relative to plain, conversational style)[1] with exaggerated acoustic features such as increased voice intensity, fundamental frequency (F0), duration, and hyper-articulation with more extreme spectral features to help speech intelligibility (Ferguson and Kewley-Port, 2002; Cooke and Lu, 2010; 2007; Krause and Braida, 2004; Smiljanić and Bradlow, 2005; Lu and Cooke, 2008; Hazan and Baker, 2011; Kim and Davis, 2014; Smiljanić, 2021). In addition to enhanced audio features, visual articulatory cues provided by speakers' mouth movements have been found to improve speech intelligibility (Perkell et al., 2002; Traunmüller and Öhrström, 2007; Kim and Davis, 2014), and perception of such visual cues can be further enhanced in clear speech (Gagné et al., 1994, 2002; Helfer, 1997; Lander and Capek, 2013; Van Engen et al., 2014).

While most clear-speech studies focus on speech segments, little research has examined clear-speech effects on prosody (including lexical tone), especially in the visual domain, presumably because prosodic production does not rely on vocal tract configurations and may less likely provide reliable visual speech cues. However, there is evidence that head, jaw, neck, eyebrow, and lip movements may convey visual information in prosodic tonal production and perception (Burnham et al., 2001; Yehia et al., 2002; Munhall et al., 2004; Chen and Massaro, 2008; Attina et al., 2010; Cvejic et al., 2010; Swerts and Krahmer, 2010; Kim et al., 2014). Furthermore, research by our team suggests such movements provide linguistically meaningful cues to signal tonal category distinctions (Garg et al., 2019).

These findings present an interesting case with respect to how these cues are utilized in clear-speech tone modification. On the basis of acoustic characteristics, clear speech has been claimed to involve two levels of modifications (Bradlow and Bent, 2002; Zhao and Jurafsky, 2009; Redmon et al., 2020), namely, signal-based and code-based. Signal-based clear-speech modifications involve changes across the entire speech signal independent of specific sound features, resulting in enhancement of overall signal saliency rather than distinctions of specific speech sounds; e.g., longer duration across vowels (Leung et al., 2016) or higher intensity across lexical tones (Tupper et al., 2021). In contrast, code-based clear-speech changes involve sound-specific modifications resulting in enhancement of phonemic contrasts; e.g., increased F2 for front vowels and decreased F2 for back vowels (Leung et al., 2016) or steeper downward F0 slope for the falling tone (Tupper et al., 2021). Likewise, clear-speech modifications of visual articulatory features may also involve signal-based changes (e.g., greater mouth opening across vowels) vs. code-based changes (e.g., greater horizontal lip stretching for /i/ and greater lip rounding for /u/, Tang et al., 2015). Effective clear-speech modifications must involve coordination of signal- and code-based strategies to enhance as well as preserve phonemic category distinctions (Moon and Lindblom, 1994; Ohala, 1995; Smiljanić and Bradlow, 2009; Tupper et al., 2018; Smiljanić, 2021). Such coordination may be challenging in cases where cues are less definitive in serving code-based functions, as in the case of visual articulatory correlates to lexical tone. As such, lexical tone provides a unique platform for testing these clear-speech principles with respect to the extent to which signal- and code-based visual cues are adopted in visual articulatory clear-speech modifications.

In the present study, we examine how the visual tonal cues identified in Garg et al. (2019) are enhanced in clear speech in the production of Mandarin Chinese tones, using state-of-the-art computer vision, image processing, and machine learning techniques.

1.1. Background

1.1.1. Visual cues in clear speech production

Kinematic studies focusing on segmental articulatory features of speech production show that speakers articulate in a more exaggerated manner in adverse listening conditions, presumably to be more intelligible to perceivers (e.g., Tasko and Greilick, 2010; Kim et al., 2011; Kim and Davis, 2014; Garnier et al., 2018). For example, studies using an Optotrak system examined articulatory movements of clear speech produced in noise and in quiet by tracking the motion of face markers as speakers produce English sentences (Kim et al., 2011; Kim and Davis, 2014). The results of these studies revealed increased movements of the jaw and mouth in speech produced in noise (clear speech) compared to that produced in quiet (plain speech). Similarly, using electromagnetic articulography (EMA), Garnier et al. (2018) examined articulatory movements in the production of French CVC words in clear speech produced in noisy environments. They found patterns of hyperarticulation in lip movements in clear (relative to plain) speech, with greater contrasts in lip aperture between low and high vowels, and in lip spreading and protrusion between spread and rounded vowels. In another EMA study, Šimko et al. (2016) examined the production of Slovak syllables containing long and short vowels in noise, allowing the comparison of clear-speech effects on segmental and suprasegmental (durational) features. They found that overall, hyperarticulated speech produced in noise was associated with expansion of movement of the jaw, the lips and the tongue as well as increased utterance duration. Furthermore, suprasegmental-level (durational) modifications associated with jaw opening appeared to be separate from segmental-level modifications associated with lip movements. Studies have also examined tongue movements in clear vs. plain speech production using a midsagittal X-ray microbeam system to track tongue fleshpoints (Tasko and Greilick, 2010). Results revealed that, in clear relative to plain productions of the word-internal diphthong /aɪ/, the tongue began in a lower position at the onset of diphthong transition (i.e., lowered tongue for /a/) and ended in a higher position at transition offset (i.e., higher tongue position for /ɪ/), indicating that clear speech resulted in significantly larger and longer movements of the tongue toward the target of the vowel components.

[1] The use of the terms "clear (hyperarticulated) style" and "plain (conversational) style" follows the convention in previous clear-speech studies (e.g., Ferguson and Kewley-Port, 2002; Maniwa et al., 2008; Tang et al., 2015; Smiljanić, 2021; Tupper et al., 2021). These two terms refer to the more enunciated vs. normal speech styles, respectively, resulting from elicitation procedures to instruct talkers to speak an utterance "normally" first in the manner used in a plain, natural conversation, and then repeat it "clearly" with the goal of improving intelligibility.

Recent research conducted by our team has developed an approach using computerized facial detection and image processing techniques to measure articulatory movements (Tang et al., 2015; Garg et al., 2019). For example, in Tang et al. (2015), we examined front- and side-view videos of speakers' faces while they articulated English words in clear vs. plain speech containing vowels differing in visible articulatory features. The results revealed significant plain-to-clear speech modifications with greater mouth opening across vowels, as well as vowel-specific modifications corresponding to the vowel-inherent articulatory features, with greater horizontal lip stretch for front unrounded vowels (e.g., /i, ɪ/) and greater degree of lip rounding and protrusion for rounded vowels (e.g., /u, Ʊ/).

Taken together, both kinematic and video-based articulatory studies consistently show hyper-articulation in clear speech, with modifications being both signal-based and generic (e.g., increased mouth opening) and code-based and segment-specific (e.g., greater lip protrusion for rounded vowels).

1.1.2. Visual articulatory cues for tone

As discussed previously, although F0 information cannot be directly triggered by vocal tract configurations, movements of the head, jaw, neck, eyebrows, as well as lips have been found to be associated with changes in prosody, including lexical tone (Burnham et al., 2001, 2022; Yehia et al., 2002; Munhall et al., 2004; Chen and Massaro, 2008; Attina et al., 2010; Swerts and Krahmer, 2010; Kim et al., 2014). Further research has revealed that facial movements (e.g., head, eyebrow, lip) in terms of spatial and temporal changes in distance, direction, speed, and timing can be aligned with acoustic features of tonal changes in height, contour, and duration (Attina et al., 2010; Garg et al., 2019).

For prosody in general, movements of the head have been shown to be correlated with F0 changes. Specifically, greater head movements are found in sentences with strong focus (Swerts and Krahmer, 2010; Kim et al., 2014), in stressed syllables (Scarborough et al., 2009), and in interrogative intonation (Srinivasan and Massaro, 2003), suggesting that the magnitude of head motion can be aligned with the amount of F0 variation. In addition to the head, eyebrow movements are also claimed to be associated with prosodic articulation (Yehia et al., 2002; Munhall et al., 2004; Swerts and Krahmer, 2010; Kim and Davis, 2014). For example, focused, accented, and stressed words in a sentence have been found to involve larger vertical eyebrow displacement and higher peak velocity of eyebrow movements (Scarborough et al., 2009; Flecha-García, 2010; Swerts and Krahmer, 2010; Kim et al., 2014), indicating that eyebrow movements may be coordinated with F0 for prosodic contrasts. However, it has been pointed out that the specific connection to F0 changes in terms of height and direction is not straightforward or invariably evident (Ishi et al., 2007; Reid et al., 2015). Moreover, although mouth configurations typically signal segmental rather than prosodic contrasts, there has been evidence that lip movements such as lip opening and lowering may be spatially and temporally aligned with prosodic variations (Dohen and Loevenbruck, 2005; Dohen et al., 2006; Scarborough et al., 2009). For example, using a facial-motion tracking system with retro-reflectors attached to the face (Qualisys), Scarborough et al. (2009) found lip movements to be larger for stressed than unstressed syllables.

Attempts have also been made to identify visible facial cues associated with lexical tone production. In particular, computer-vision research from our team has found that spatial and temporal changes in distance, direction, speed, and timing are related to acoustic features of Mandarin tonal changes in height, contour, and duration (Garg et al., 2019). From tracking head movements, Garg et al. (2019) has revealed that Mandarin high-level tone (Tone 1), which involves minimal F0 variation compared to the other contour tones, exhibits minimal head movements and low movement velocity. These patterns are consistent with previous kinematic sensor-based results showing that head movements (e.g., nodding, tilting, rotation toward the back) are correlated with F0 changes in Cantonese tones (Burnham et al., 2006, 2022). Similar to head movements, the spatial and temporal changes in eyebrow motion also follow the trajectories of F0 height and contour in Mandarin tones. Garg et al. (2019) shows that the magnitude of eyebrow displacement along with its movement velocity is smaller for the level tone as compared to the contour tones, for which the eyebrow movements are aligned with the direction and timing of the rising (Tone 2), dipping (Tone 3), and falling (Tone 4) trajectories. The spatial and temporal events in tone production may also coordinate to mouth movements. For example, compared to the other tones, Tone 4 exhibits the longest time to reach the maximum velocity of lip closing, accompanied by the longest time for the head and the eyebrows to reach maximum lowering, suggesting later lowering movement corresponding to the falling F0 trajectory of this tone (Garg et al., 2019). Findings from sensor-based studies also corroborate a general correlation between lip movements and F0, with lip raising movements corresponding to the high F0 nature of Tone 1 and lip protrusion relating to the rising contour of Tone 2 (Attina et al., 2010).

Together, the findings based on the analyses of head, eyebrow and lip movements reveal linguistically meaningful facial cues in tone articulation. One subsequent question yet to be addressed is whether speakers make use of such cues to modify their speech, such as in clear speech in adverse listening contexts with the intention of enhancing intelligibility. Han et al. (2019) analyzed videos of Mandarin tone production teaching (clear) style by four Mandarin instructors. They found a greater total amount of facial movements and longer durations in clear relative to natural (plain) speech. There were also tone-specific differences, with greater horizontal movements for the high-level tone and greater vertical movements for the rising and falling tones in clear than plain speech. However, the measures were limited to the three general facial movement measures (total amount, horizontal, vertical) and were not associated with particular facial regions (e.g., eyebrows, lips) as revealed by other research (Attina et al., 2010; Garg et al., 2019). It is thus unclear whether the exaggerated facial movements observed in clear speech are associated with code-based linguistically meaningful tonal cues identified previously.

1.2. The present study

In this study, we compare movements of the head, eyebrows and lips associated with clear vs. plain speech styles of Mandarin tone articulation. The comparisons are based on a comprehensive set of static (distance- and time-based) cues as well as dynamic (kinematic-based) cues identified to characterize different Mandarin tone categories in our previous research (Garg et al., 2019).

Mandarin tone provides a unique case in examining clear-speech characteristics in articulation. As mentioned previously, the articulation of tone primarily involves laryngeal activities not necessarily distinguishable through visible articulatory movements. It is thus unclear if clear-speech modifications involve exaggerated signal-based facial movements in general or enhancement of code-based linguistically relevant articulatory movements. The current study aims at disentangling which articulatory features are used in clear-speech modifications across tones (signal-based) and which features are unique for individual tones, and furthermore, if such tone-specific adjustments are aligned with the (code-based) category-defining features for each tone identified in Garg et al. (2019). Such findings will have implications for unraveling which visual cues may enhance tone perception and intelligibility.

2. Methods

2.1. Speakers and stimuli

2.1.1. Speakers

Twelve female and eight male native Mandarin speakers aged between 18 and 28 years (mean: 22.6 years) were recruited. The speakers were born and have spent at least the first 18 years of their lives in either Northern China or Taiwan. They had resided in Canada for less than five years at the time of recording.

2.1.2. Stimuli

The stimuli were monosyllabic Mandarin words, each containing the vowel /ɤ/ with one of the four Mandarin tones, carrying the meaning of "graceful" (/ɤ1/; Tone 1, high-level tone), "goose" (/ɤ2/; Tone 2, mid-high-rising tone), "nauseous" (/ɤ3/; Tone 3, low-dipping tone), or "hungry" (/ɤ4/; Tone 4, high-falling tone), respectively.

2.1.3. Elicitation of plain and clear speech

The elicitation of plain and clear tones followed the procedures developed previously (Maniwa et al., 2009; Tang et al., 2015). A simulated interactive computer speech recognition system was developed using MATLAB (The Mathworks, R2013, Natick, MA, USA), where the program seemingly attempted to recognize a target stimulus produced by a speaker. The speaker was first instructed to read each of the stimuli that was shown on the screen naturally (to elicit plain style productions, e.g., /ɤ4/). Then the program would show its "guess" of the produced token. The software would systematically make wrong guesses due to "recognition" tonal errors (e.g., Did you say /ɤ3/?). The speaker was then requested to repeat the token more clearly (to elicit clear style productions, e.g., /ɤ4/). A total of 96 pronunciations of tone quadruplet words in two speaking styles (plain, clear) were videotaped from each speaker over three recording sessions (4 tones x 2 styles x 12 repetitions). The average duration of the target stimuli was 580 ms (SD = 193 ms) across styles, tones and speakers. In addition to the /ɤ/ word we also recorded /i/ and /u/ words as fillers.

2.1.4. Recording

The data was collected in a sound-attenuated booth in the Language and Brain Lab at Simon Fraser University. The speaker sat approximately three feet from a 15-inch LCD monitor on which the stimulus word was presented. The monitor was positioned at eye-level to facilitate the placement of a front-view video camera, which was placed below the monitor on a desktop tripod. A high-definition Canon Vixia HF30 camera was used to record the front-face of the speaker. The frame rate of the camera is 29 fps. Each speaker was made to sit with their back against a monochromatic green backdrop and was recorded separately. For interaction with the computer display, speakers were instructed in the usage of a video game wireless controller, which offered a comfortable and quiet way to interact with the display with minimal movement required from the speaker and introduced minimal interference with the video and audio recordings.

2.2. Analysis

The analysis followed the tone articulation analysis approach previously developed by our team (Garg et al., 2019). It first involved extraction of articulatory features. Two analyses were subsequently conducted across tones and for each tone. First, discriminative analysis of the extracted motion features in clear and plain styles was conducted via random forest classification (Paul and Dupont, 2015). Random forest tests the features using multivariate analysis to identify which features significantly differentiate plain and clear styles and rank the importance of these features in contributing to the plain-clear differentiation. Second, for each of the features identified by random forest, the extent of movements (e.g., head movement distance) in plain vs. clear speech were compared using mixed-effects modeling to determine which of the features involved a significant difference between the two styles.

2.2.1. Feature extraction

A total of 33 facial articulatory features which were previously identified as tone characterizing features (Garg et al., 2019) were included in this study to examine the plain-clear style differences.

Feature extraction involved the following steps using computer-vision and image processing techniques. First, regions of interest (ROI) on the face such as eyes, nose and lips of the speaker were identified on the first frame of the video and were subsequently tracked in the rest of the video. Briefly, the bounding box on the regions of interest are identified using LBP (Local Binary Pattern) cascade filters and then landmark-outlines are identified using active contour models. Specific keypoints on the landmark outlines such as nose tip, inner corner of the left eyebrow,[2] and cupid's bow on lips are identified for tracking purposes. The Kanade-Lucas-Tomasi (KLT) feature-tracking algorithm was then used to track the aforementioned keypoints after they were found on the first frame of each video token. Next, the 33 features were computed on the motion trajectories of four keypoints identified on the nose tip (proxy for head movement), the left eyebrow, and the midpoints of the upper and lower lips. Then, each set of features was normalized to account for between-speaker differences, by dividing the feature values by a normalization factor computed as the shortest distance between the line joining the two eyes and the nose tip for subsequent analyses.

The absolute mean value was computed for each feature and each style to compare if the magnitude of the movements is different in clear speech than plain speech and when these movements occur during the tone production.

The 33 features can be generally classified into three categories: (1) *distance-based*, characterizing the minimum or maximum total displacement of a keypoint from its initial resting position to a target position; (2) *time-based*, characterizing the time it takes the displacement of a keypoint to reach maximum or minimum distance; and (3) *kinematic*, characterizing the velocity and acceleration of a keypoint at a specific time instance marked by a target event (e.g., instance when velocity reaches a maximum).

Table 1 contains a list of all the features and their descriptions. The distances are measured in pixels and the relative times are measured by the number of video frames divided by the total number of frames. Each feature is normalized to remove the variations due to head size among different speakers. Normalization was done by dividing the feature values by a normalization factor computed as the shortest distance between the line joining the two eyes and nose tip in that particular token. Since the features are normalized, the reported feature magnitudes are unitless. Figure 1 illustrates an example video frame showing the keypoints which are tracked for head, eyebrow and lip movements, and movement trajectories for a sample token in plain and clear speech styles. The distance-based features were calculated as the minimum and maximum distances that each of the tracked keypoints moved from its initial resting state. The positive measurements from the resting state signify an action of rising or opening, whereas negative measurements represent an action of lowering or closing. Velocities were then calculated as rate of change of the curve (i.e., slope). Finally, the acceleration is computed by the rate of change in the velocity curve.

We assessed the physical head size of two randomly selected speakers—one male and one female—in order to relate the derived measures from pixels to physical units (i.e., mm). For distance-based features, each pixel measured to 0.33 mm for male and 0.36 mm for female. For time-based and kinematic features, the videos were recorded at 29 fps and can be used to convert the per frame unit to per seconds. For examples, (1) the average head displacement for the male speaker during head-raising is 1.58 mm (4.75 pixels) and 5.88 mm (16.14 pixels) for the female speaker, and (2) the maximum head velocity during head-raising is 0.32 mm/s (0.98 pixels/s) for the male speaker and 0. 67 mm/s (1.83 pixels/s) for the female speaker. Further information for each feature can be found in Supplementary Table A1 of Garg et al. (2019).

2.2.2. Discriminative analysis via the random forest approach

We adopted the improved random forest approach developed by Paul and Dupont (2015), which was especially appropriate for this study since it enabled us to assess both the significant features and the ranking of these features that differentiate the two styles.

Random forest classification works by training an ensemble of basic decision trees, each of which predicts (or outputs) a class label given an input pool of features, with the final class label being determined by computing the average of the class label predictions from each tree. By using a random subset of samples and a random subset of features to train each tree, randomness is introduced into the system. In order to ensure reproducibility of the experiments, a random seed is set so that in every run the same random numbers are generated.

In our experiments, 1,500 trees were used for random forest classifiers. The style discriminative analysis involved a binary classification of clear vs. plain speech style from the recorded video tokens described above.

2.3. Analysis of plain- and clear-speech comparisons

Our main goal was to examine which of the tone-defining features reported in Garg et al. (2019) could significantly differentiate the two speech styles. To this end, we conducted two types of comparisons: style differences across tones, and style differences within each tone.

2.3.1. Style comparisons across tones

The random forest analysis first provided us with features that differentiate the two styles across tones. All the features from different tones were pooled together and used as a training set for the random forest classifier. Then each feature importance was computed by permuting the samples in that feature and measuring the change in the prediction power. For each feature, if the measured changes were deemed statistically significant, then that feature would be considered important in differentiating the two styles across tones. We then employed the feature importance weights (Paul and Dupont, 2015) to rank the features in decreasing order of importance using leave-one-out cross-validation. The larger the weight, the more important the feature is for style discrimination. The features that were found to be important were further analyzed using linear mixed-effects modeling with style as the independent variable and

[2] Left eyebrow was chosen based on the previous findings that the left relative to right eyebrow is more prominent in prosodic production and is more strongly correlated with prosodic patterns (Cavé et al., 1996; Swerts and Krahmer, 2008). Future research could compare measures of left vs. right eyebrows to further examine how left and right eyebrows contribute to plain-to-clear modifications of Mandarin tones.

TABLE 1 The set of 33 features used to represent tone articulation in each video token (cf. Garg et al., 2019).

ROI	Full feature name	Short term	Type (distance, time, kinematic)
Head	Maximum displacement of the head while head-raising from its starting position	max_vert_head_distance	Distance
Head	Maximum displacement of the head while head-lowering from its starting position	min_vert_head_distance	Distance
Head	Average distance head moved during the utterance	avg_abs_vert_head_distance	Distance
Head	Total distance traveled by head during the utterance	total_abs_vert_head_distance	Distance
Eyebrow	Maximum displacement of the eyebrow while eyebrow-raising from its starting position	max_vert_left_eye_distance	Distance
Eyebrow	Maximum displacement of the eyebrow while eyebrow-lowering from its starting position	min_vert_left_eye_distance	Distance
Eyebrow	Average distance eyebrow moved during utterance	avg_abs_vert_left_eye_distance	Distance
Eyebrow	Total distance eyebrow moved during the utterance	total_abs_vert_left_eye_distance	Distance
Lips	Maximum lip-opening distance	max_lips_distance	Distance
Lips	Maximum lip-closing distance	min_lips_distance	Distance
Lips	Average distance lips moved during utterance	avg_lips_distance	Distance
Lips	Total distance lips moved during the utterance	total_abs_lips_distance	Distance
Head	Relative time at which the displacement of the head while head-raising was maximum	time_max_head_vert_distance	Time
Head	Relative time at which the displacement of the head while head-lowering was maximum	time_min_head_vert_distance	Time
Head	Relative time at which the head velocity was maximum during head-raising	time_max_head_vert_velocity	Time
Head	Relative time at which the head velocity was maximum during head-lowering	time_min_head_vert_velocity	Time
Eyebrow	Relative time at which the displacement of the eyebrow while eyebrow-raising was maximum	time_max_left_eye_vert_distance	Time
Eyebrow	Relative time at which the displacement of the eyebrow while eyebrow-lowering was maximum	time_min_left_eye_vert_distance	Time
Eyebrow	Relative time at which the eyebrow velocity was maximum during eyebrow-raising	time_max_left_eye_vert_velocity	Time
Eyebrow	Relative time at which the eyebrow velocity was maximum during eyebrow-lowering	time_min_left_eye_vert_velocity	Time
Lips	Relative time at which the amount of lip-opening reached maximum	time_max_lips_distance	Time
Lips	Relative time at which the amount of lip-closing reached maximum	time_min_lips_distance	Time
Lips	Relative time at which the lip velocity during lip-opening was maximum	time_max_lips_velocity	Time
Lips	Relative time at which the lip velocity during lip-closing was maximum	time_min_lips_velocity	Time
Head	Maximum head velocity during head-raising	max_head_vert_velocity	Kinematic
Head	Maximum head velocity during head-lowering	min_head_vert_velocity	Kinematic
Head	Maximum absolute acceleration of the head	max_abs_head_vert_acceleration	Kinematic
Eyebrow	Maximum eyebrow velocity during eyebrow-raising	max_left_eye_vert_velocity	Kinematic
Eyebrow	Maximum eyebrow velocity during eyebrow-lowering	min_left_eye_vert_velocity	Kinematic
Eyebrow	Maximum absolute acceleration of the eyebrow	max_abs_left_eye_vert_acceleration	Kinematic
Lips	Maximum lips velocity during lip-opening	max_lips_velocity	Kinematic
Lips	Maximum lips velocity during lip-closing	min_lips_velocity	Kinematic
Lips	Maximum absolute acceleration of the lips	max_abs_lips_acceleration	Kinematic

Head, eyebrows (left eye) and lips are the regions of interest (ROI); "max" in short term represents a raising or opening event and "min" represents a falling or closing event. All the time-related features start with "time" in the short term.

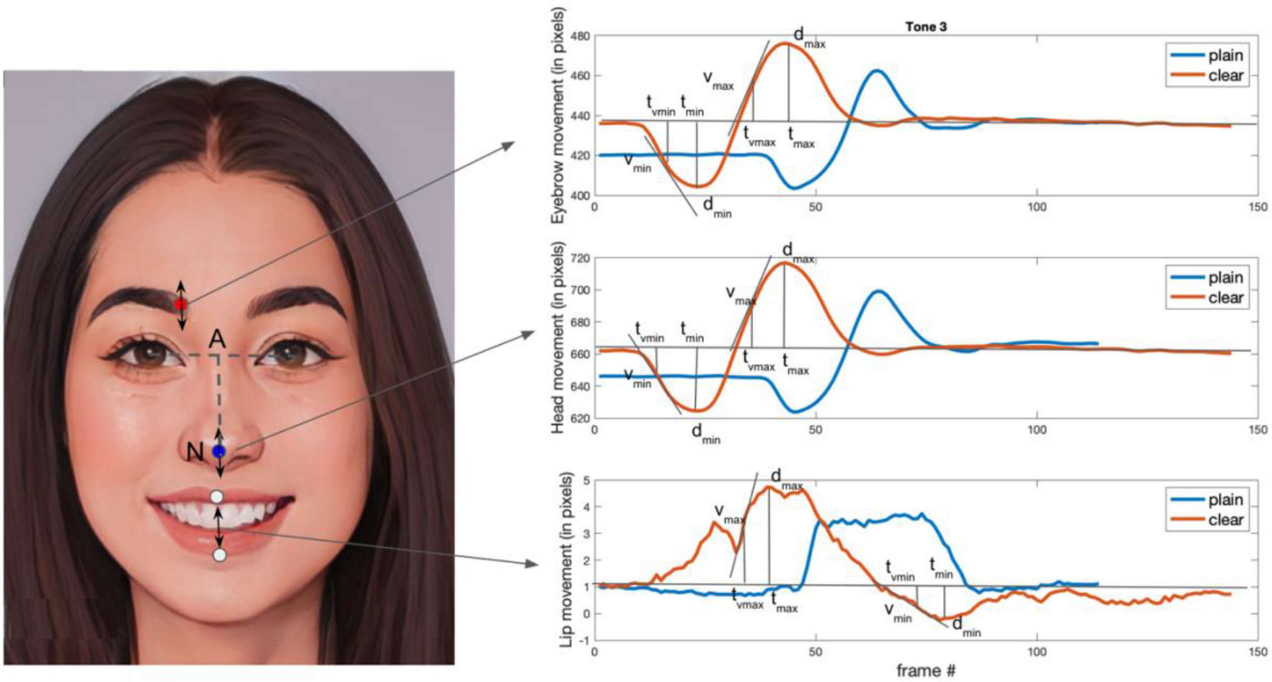

FIGURE 1
The left-side image shows the landmark points (keypoints) at which the trajectory of the vertical movement is measured. A total of four keypoints are selected corresponding to the nose tip (proxy for the head, blue), left eyebrow (red), and lips (white). The right-side plots show the head, eyebrow, and lip movement trajectories for a sample token comparing the clear and plain speech style. Normalization factor used in the analysis is the length of the line segment AN. D, distance; t, time; v, velocity; max, raising/opening; min, lowering/closing (The face is generated by AI.).

the value of each important feature as the dependent variable using the MATLAB *fitlme*. The random intercept and slope of style on speaker were included in the models with the following syntax:

$$\texttt{feature} \sim \texttt{style} + (1 + \texttt{style}|\texttt{speaker})$$

The final set of features that involve a significant style difference as determined by the mixed-effects modeling are considered generic (non-tone-specific) style features.

2.3.2. Style comparisons for each tone

To examine which features were different between the two styles within each tone, we performed similar random forest and mixed-effects modeling analyses as described in Section 2.3.1. for each tone separately. After identifying a set of features that involve a significant difference in style for each tone, we compared these style-characterizing features to those obtained in Garg et al. (2019) that define each tone. We hypothesized that, for a particular tone, any style-characterizing features that overlap with tone-defining features are considered involving tone-specific clear-speech modifications. In contrast, any that overlap with the cross-tone features identified in 2.3.1 should be style-specific only.

3. Results

3.1. General style difference across tones

First, we present the discriminant analysis results on the 33 features using random forest (RF). Using the procedure described in Section 3, thirteen features were found to be significant using RF classifier in differentiating the two speech styles, as shown in Figure 2. The features are arranged in descending order of their importance as determined by the random forest classifier. For each feature, the weight is the increase in prediction error if the values of that feature are permuted across the out-of-bag observations. This measure is computed for every decision tree in RF, then averaged over the entire ensemble and divided by the standard deviation over the entire ensemble. A larger error means that the feature is more important in classifying the style. Among the thirteen features, eight were found to be distance-based and five were related to time. The distance-based features primarily involve changes in the vertical distance of the head, lips and eyebrows, and the time-based features primarily involve changes in the time when the vertical head, lip and eyebrow movements reach maximum velocity. The feature importance ranking further revealed that the “Relative time at which the lip velocity during lip-opening was maximum” was the most differentiating feature to distinguish the two styles followed by the “Maximum displacement of the head while head-lowering from its starting position”, whereas the “Maximum lips velocity during lip-closing” was the least significant factor.

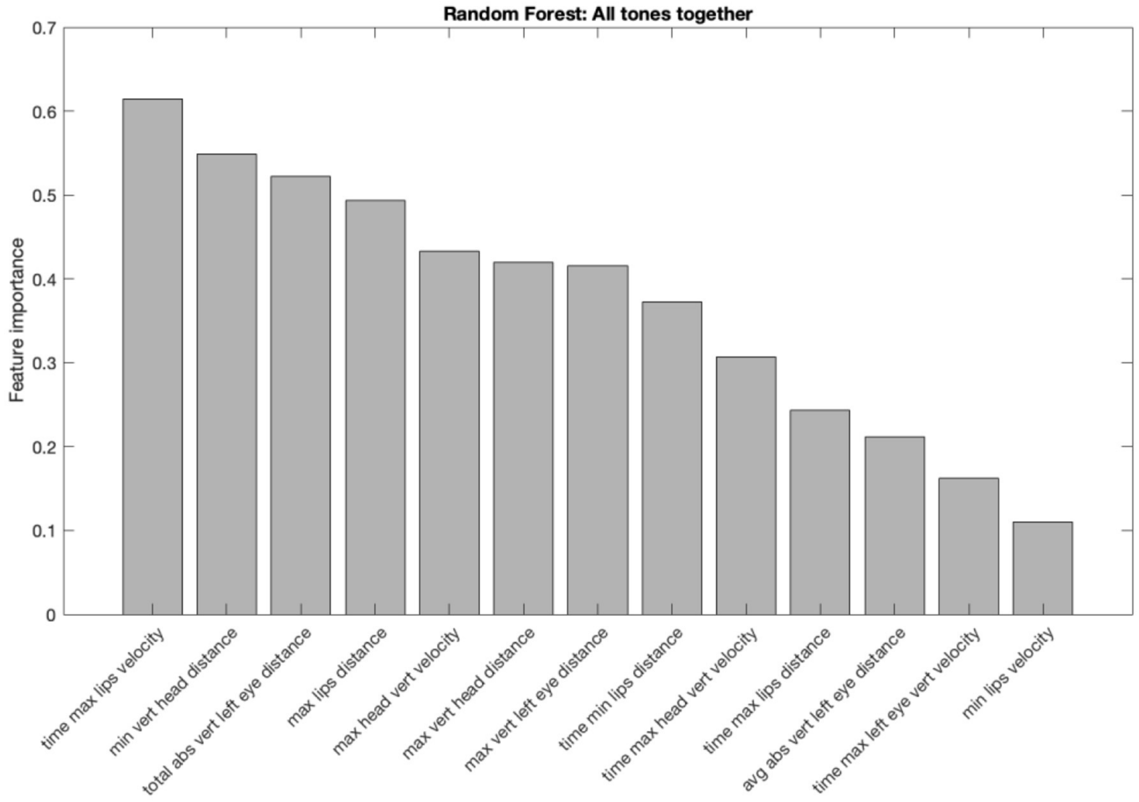

FIGURE 2
Important features identified by the random forest analysis in differentiating the two speech styles (clear and plain). The y-axis shows the feature importance measured by the random forest. The feature importance measures the increase in classification error if the values of the feature are permuted. The greater the error the higher the importance of the feature is.

To further determine the significance of the differences of the speech style for each of the thirteen features identified by the random forest classifier, the mean values of the normalized feature in clear and plain speech were compared using linear mixed-effects analysis as described in 2.3.1. The results, as summarized in Table 2, show that twelve out of thirteen features involve a significant clear-plain difference. Figure 3 displays the clear and plain style comparisons for each feature.

Specifically, the eight features where the magnitude of change is larger in clear than plain style include:

1. Maximum displacement of the head while head-raising from its starting position.
2. Maximum displacement of the head while head-lowering from its starting position.
3. Maximum lip-opening distance.
4. Maximum displacement of the left eyebrow while eyebrow-raising from its starting position.
5. Average distance left eyebrow moved during utterance.
6. Total distance traveled by left eyebrow during the utterance.
7. Maximum head velocity during head-raising.
8. Relative time at which the amount of lip-closing reached maximum.

The four features where the magnitude of change is smaller in clear than plain style are:

1. Relative time at which the head velocity was maximum during head-raising.
2. Relative time at which the lip velocity during lip-opening was maximum.
3. Relative time at which the left eyebrow velocity was maximum during eyebrow-raising.
4. Relative time at which the amount of lip-opening reached maximum.

The above list and Figure 3 shows that eight significant features had a larger movement magnitude in clear speech than in plain speech. These eight features are either distance or time related, including greater maximum distance of head raising or lowering, eyebrow raising and movement, and lip opening from their starting positions in clear than plain style, as well as longer time at which the amount of lip closing reached maximum in clear than plain style. These patterns suggest larger head, eyebrow and lip movements and faster arrival at the movement peak in clear relative to plain tone production. In contrast, four time-related features involved smaller magnitude of change in clear than plain speech, including

TABLE 2 Summary of the mixed-effects linear regression model for each of the features that involves a significant clear-plain difference across tones.

Feature Name	Estimate	SE	*t*-stat	DF	*p*-value
max_head_vert_velocity	0.004	0.001	3.299	1,809	0.001
time_max_head_vert_velocity	−0.035	0.011	−3.241	1,809	0.001
min_vert_head_distance	0.033	0.011	3.137	1,809	0.002
max_vert_head_distance	0.020	0.008	2.410	1,809	0.0160
time_max_left_eye_vert_velocity	−0.029	0.010	−2.915	1,809	0.004
avg_abs_vert_left_eye_distance	0.005	0.002	2.229	1,809	0.026
max_vert_left_eye_distance	0.014	0.007	2.057	1,809	0.040
total_abs_vert_left_eye_distance	1.059	0.368	2.878	1,809	0.004
time_max_lips_velocity	−0.030	0.011	−2.797	1,809	0.005
time_max_lips_distance	−0.023	0.011	−2.155	1,809	0.031
time_min_lips_distance	0.058	0.024	2.428	1,809	0.015
max_lips_distance	0.024	0.011	2.286	1,809	0.022

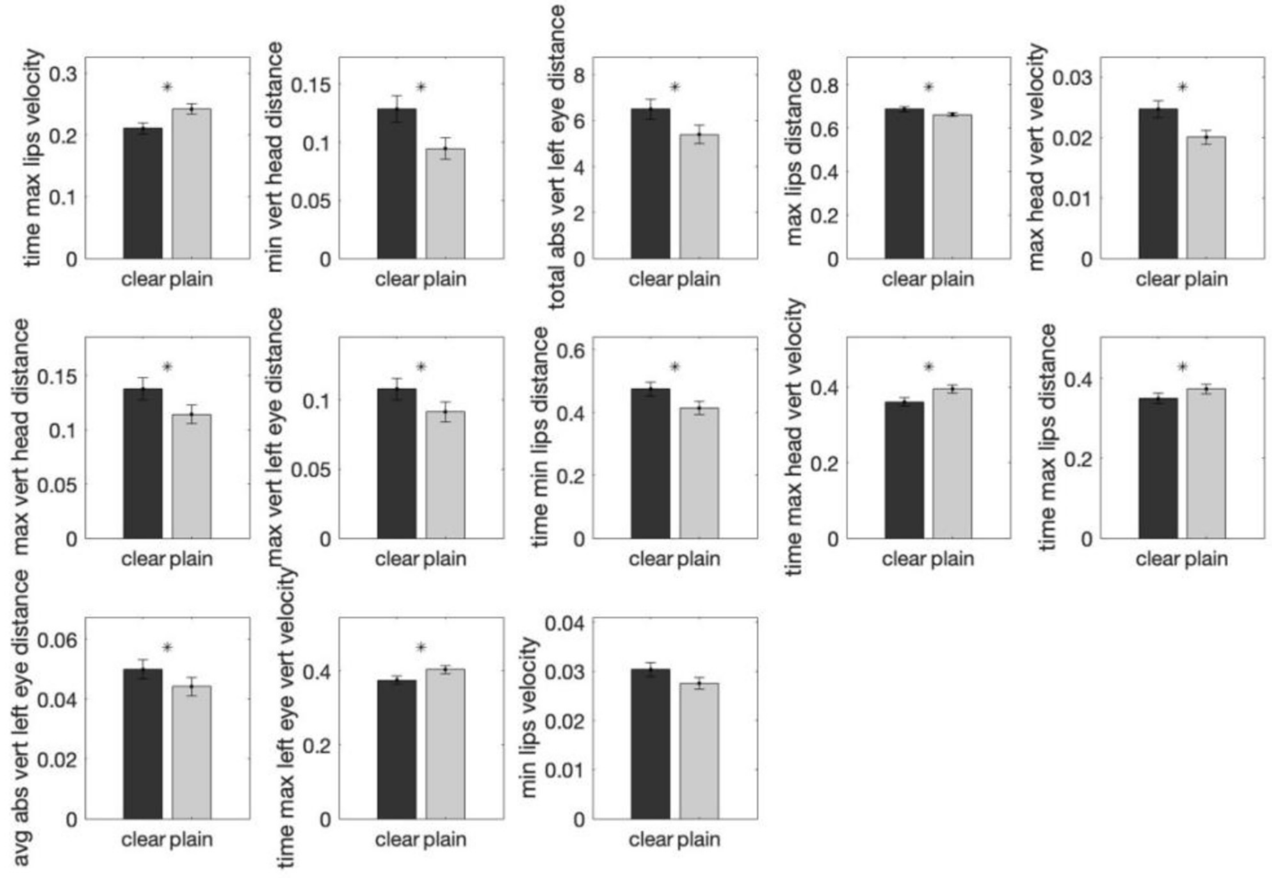

FIGURE 3
Comparisons of the group means of each of the thirteen important features (determined by random forest analyses) in clear and plain style irrespective of the tones. The y-axis shows the average normalized feature value for that particular style group. Twelve features were found to exhibit statistically significant differences between the two speech styles by mixed-effects modeling. The error bar on each of the bar plots shows the standard error. A * above a bar represents a p-value smaller than 0.05.

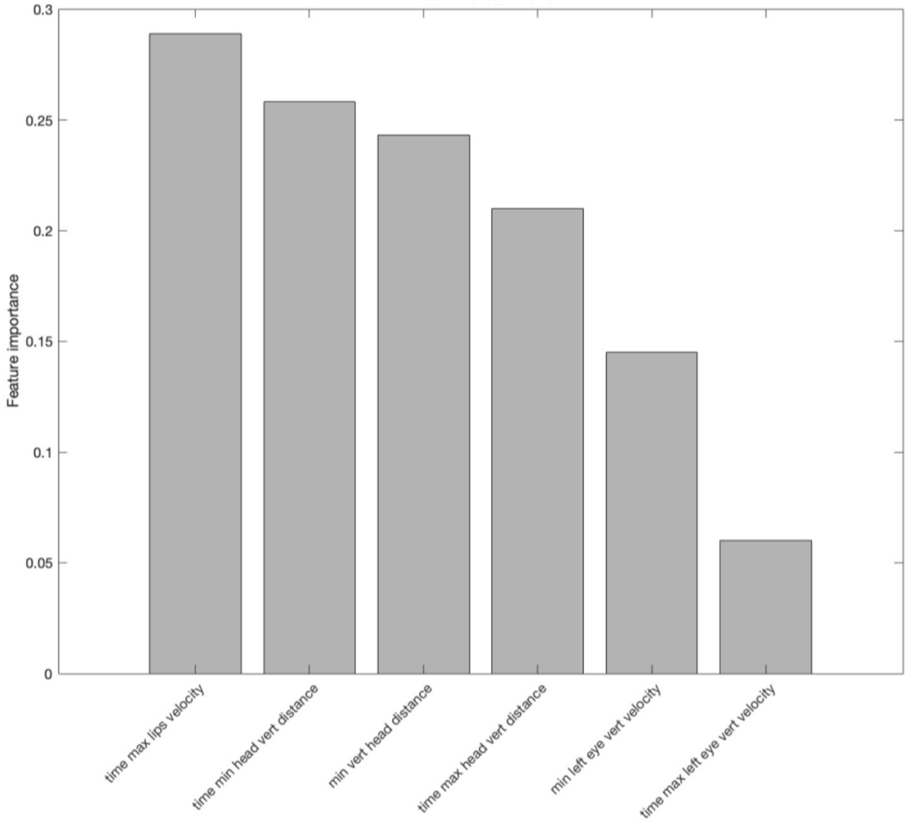

FIGURE 4
Important features identified by the random forest analysis in differentiating the two speech styles (clear and plain) in Tone 1. The y-axis shows the feature importance measured by the random forest. The feature importance measures the increase in classification error if the values of the feature are permuted. The greater the error the higher the importance of the feature is.

shorter time taken for head raising, lip opening and eyebrow raising to reach maximum velocity. These features indicate that movement maxima happened earlier in clear than plain style, suggesting faster arrival at the movement peak in clear relative to plain tone production.

Overall, these patterns consistently reveal larger maximum displacement of head, eyebrow and lips, and faster arrival at these positions in clear than plain tone production, demonstrating exaggerated articulation in clear speech.

3.2. Tone-specific analysis

Next, we analyze each tone separately to identify tone-specific features that can differentiate the two styles. These features are then compared with the set of features characterizing each tone as reported in Garg et al. (2019) to determine the extent to which clear speech modifications adopt tone-intrinsic features.

3.2.1. Tone 1 (High-level tone)

First, the random forest analysis showed six important features differentiating the two speaking styles in Tone 1, listed in Figure 4 in decreasing order of their feature importance values. Four out of six features were time related, with the "Relative time at which the lip velocity during lip-opening was maximum" having the largest feature importance whereas "Relative time at which the displacement of the eyebrow while eyebrow-raising was maximum" having the smallest weight. Apart from these time-related features, "Maximum displacement of the head while head-lowering from its starting position" and "Maximum eyebrow velocity during eyebrow-raising" were also found to be important.

The difference in the magnitude of the mean values was then evaluated between the two styles using mixed-effects modeling. As displayed in Figure 5, for Tone 1, two features were found significant in differentiating the two styles, with 'maximum displacement of the head while head-lowering from its starting position' being larger in clear than plain speech ($\beta = 0.039$, standard error (SE) $= 0.015$, t $(406) = 2.30$, $p < 0.05$) and 'the relative time at which the displacement of the head while head-raising was maximum' being smaller in clear than in plain speech ($\beta = -0.045$, SE $= 0.023$, t $(406) = -1.979$, $p < 0.05$). The first feature regarding head lowering distance has been identified not only as a tone-generic feature (3.1) but also as one of the defining features for Tone 1 in Garg et al. (2019), where head-lowering distance is the smallest in value among all the tones, reflecting that articulation of Tone 1 involves minimal head movement compared to the other tones.

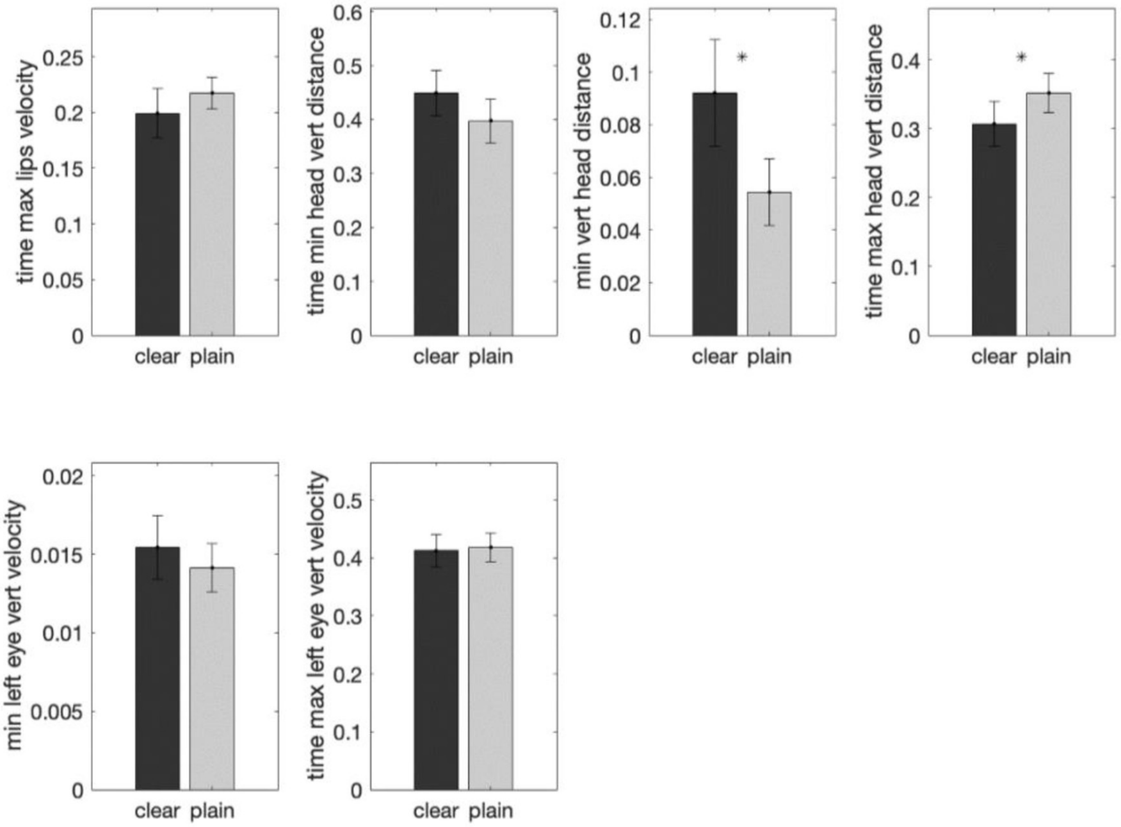

FIGURE 5
Comparisons of the group means of each of the six important features (determined by random forest analyses) in clear and plain style in Tone 1. The y-axis shows the average normalized feature value for that particular style group. Two features were found to exhibit a statistically significant difference between the two speech styles by mixed-effects modeling. The error bar on each of the bar plots shows the standard error. A * above a bar represents a p-value smaller than 0.05.

3.2.2. Tone 2 (High-rising tone)

The random forest analysis identified five out of 33 features as important features in style discrimination for Tone 2, in descending importance ranking (see Figure 6):

1. Relative time at which the lip velocity during lip-opening was maximum.
2. Relative time at which the lip velocity during lip-closing was maximum.
3. Relative time at which the amount of lip-closing reached maximum.
4. Total distance traveled by head during the utterance.
5. Total distance traveled by left eyebrow during the utterance.

Mixed-effects modeling revealed that the two styles were significantly different for all the five features (Figure 7). In both distance-related features clear speech had larger magnitude of movement than plain speech, indicating that the total distance traveled by the head ($\beta = 1.686$, SE $= 0.631$, $t\ (454) = 2.669$, $p < 0.01$) and eyebrow ($\beta = 1.338$, SE $= 0.549$, $t\ (454) = 2.437$, $p < 0.05$) are longer in clear than plain speech. For time-related features, clear relative to plain speech took shorter time for the lip-opening ($\beta = -0.042$, SE $= 0.012$, $t\ (454) = -3.443$, $p < 0.001$) and lip-closing ($\beta = -0.043$, SE $= 0.014$, $t\ (454) = -3.004$, $p < 0.05$) velocity to reach maximum, while the lips took more time to close ($\beta = 0.107$, SE $= 0.040$, $t\ (454) = 2.659$, $p < 0.05$). Thus, although clear speech may involve larger movement and may take longer to complete than plain speech, it tends to reach movement maxima sooner. These patterns are aligned with the overall clear speech features across tones reported above.

Garg et al. (2019) reported that the feature distinguishing Tone 2 from the rest of the tones was that 'relative time at which the displacement of the head while head-raising was maximum' was longer for Tone 2 than for the other tones. The current results show that this feature was not used in the clear-plain speech distinction. Hence, for Tone 2, all the identified features characterizing the clear-plain differences involve style-specific modifications.

3.2.3. Tone 3 (Low falling-rising tone)

For Tone 3, nine features were found to be important based on random forest discriminative analysis (Figure 8). Six out of these nine features are time related and the other three are distance based. The most important feature was "Maximum

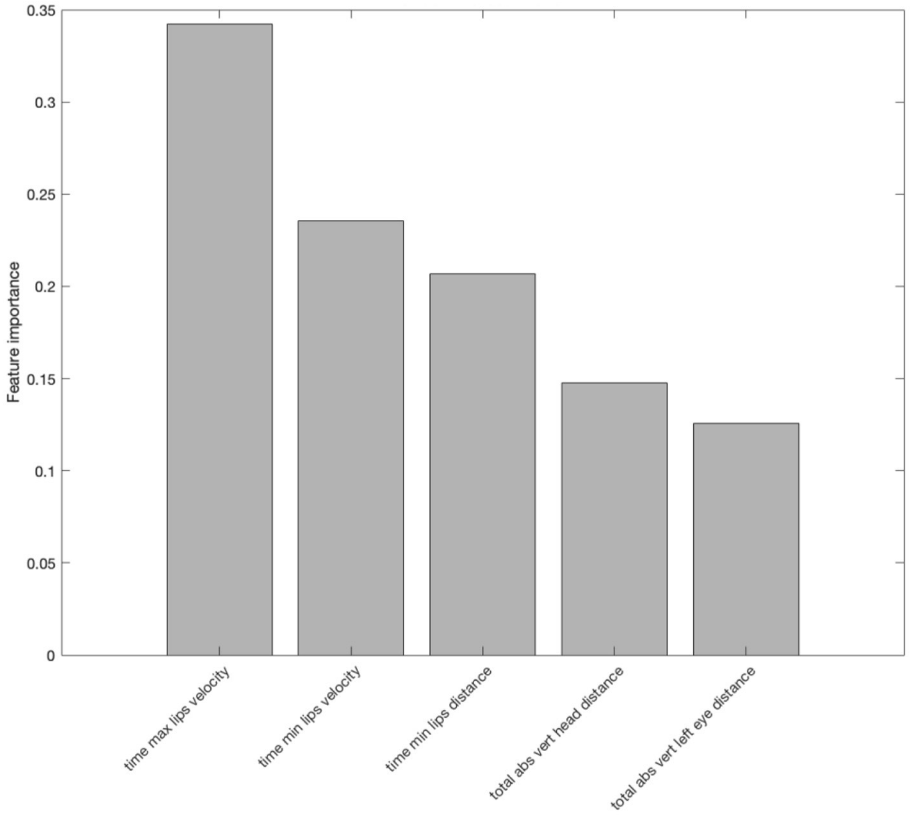

FIGURE 6
Important features identified by random forest in differentiating the two speech styles (clear and plain) in Tone 2. The y-axis shows the feature importance measured by the random forest. The feature importance measures the increase in classification error if the values of the feature are permuted. The greater the error the higher the importance of the feature is.

displacement of the head while head-raising from its starting position" whereas the least weighted feature was "Relative time at which the displacement of the eyebrow while eyebrow-raising was maximum."

Follow-up mixed-effects modeling reveals three features to be significantly different in their magnitude between the two speech styles (Figure 9). Two features are time-related. Specifically, "relative time taken for lip-opening velocity to reach maximum" (β = −0.041, SE = 0.016, t (519) = −2.625, $p < 0.05$) and "relative time taken for head-raising velocity to reach maximum" (β = −0.039, SE = 0.017, t (519) = −2.330, $p < 0.05$) are shorter in clear than in plain speech, indicating a faster approach to target gesture in clear speech. The third feature is distance-related, where "maximum displacement of the head while head-raising from its starting position" (β = 0.035, SE = 0.012, t (519) = 2.863, $p < 0.05$) is larger in clear than in plain speech, indicating larger movements in clear speech style.

Among the three significant style-distinguishing features, the feature involving the "relative time at which the head velocity was maximum during head-raising" was identified as one of the Tone 3-specific features previously (Garg et al., 2019), where it was shorter for Tone 3 relative to the other tones. However, this change is also a universal clear-speech pattern across tones.

3.2.4. Tone 4 (High-falling tone)

For Tone 4, random forest analysis revealed eight features to be important in style distinctions (Figure 10), among which seven are time-based and one is related to distance. The most important feature is the "Relative time at which the displacement of the head while head-lowering was maximum" and the least important feature is the "Relative time at which the amount of lip-closing reached maximum."

Three out of these eight features are shown to involve significant differences between plain and clear speech, as determined by further mixed-effects analysis (Figure 11). Specifically, the "relative time at which the head velocity was maximum during head-raising" was found to be shorter in clear than in plain speech (β = −0.043, SE = 0.016, t (424) = −2.735, $p < 0.05$). The second feature involves "relative time at which the amount of lip-closing reached maximum" (β = 0.092, SE = 0.036, t (424) = 2.581, $p < 0.05$), which occurred later in clear than in plain style, suggesting longer duration in clear-speech production. The third feature is the "total distance traveled by head during the utterance "(β = 1.948, SE = 0.703, t (424) = 2.770, $p < 0.05$), which appears to be larger in clear than plain speech, as expected.

One of these significant features was a Tone 4-specific feature (Garg et al., 2019); that is, the "relative time at which the head velocity was maximum during head-raising" was shorter for Tone

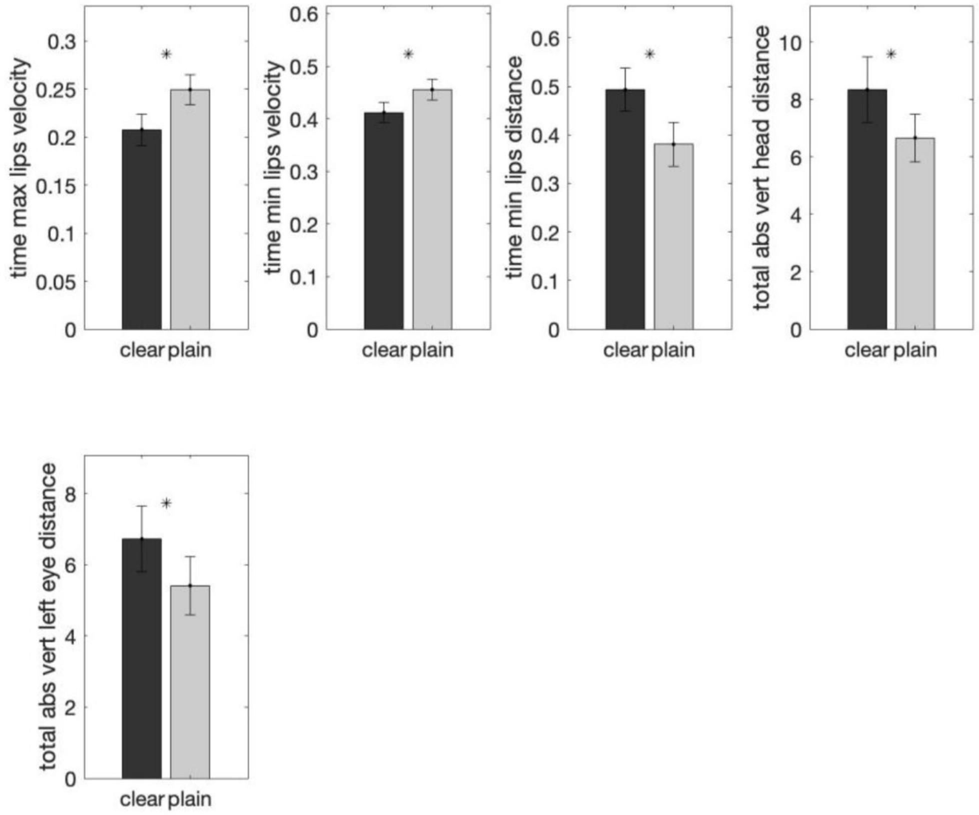

FIGURE 7
Comparisons of the group means of each of the five important features (determined by random forest analyses) in clear and plain style in Tone 2. The y-axis shows the average normalized feature value for that particular style group. All features were found to exhibit a statistically significant difference between the two speech styles by mixed-effects modeling. The error bar on each of the bar plots shows the standard error. A * above a bar represents a p-value smaller than 0.05.

4 relative to most of the other tones, indicating faster return of the head to the resting position after the head lowering gesture for the falling Tone 4. Clear speech, with an even faster head-raising velocity, apparently enhanced this feature.

3.3. Summary of results

In summary, across tones, clear speech demonstrated exaggerated articulation compared to plain speech, with larger maximum displacement of head, eyebrows and lips, and faster arrival at these positions. The analysis of individual tones showed that these general clear-speech enhancement patterns primarily hold for each tone, while certain tone-specific features were also strengthened.

For Tone 1, two features showed a significant difference between plain and clear speech, including one tone-specific feature, namely head lowering distance. However, the direction of this clear-speech modification was in conflict with the Tone 1 intrinsic feature. That is, while Tone 1, as a level tone, was characterized as having smaller head lowering compared to the other tones, the movement was not further restrained in clear speech. Instead, clear speech demonstrated larger head lowering than plain speech, consistent with the tone-general pattern of larger movements in clear relative to plain speech. Similarly, the second significant feature showing a plain-to-clear difference, which involved shorter time taken for head-raising to reach maximum in clear speech, was also in line with the across-tone patterns of faster arrival at the movement peak in clear than plain tone production.

For Tone 2, what significantly distinguished clear and plain styles involved no Tone 2-specific features. Instead, plain-to-clear speech modifications of Tone 2 involved larger head and eyebrow movements and longer (lip-closing) time to complete the production, as well as quicker lip movements to reach target gesture, which were primarily aligned with the overall clear speech features across tones.

Tone 3 clear speech modifications involved one unique Tone 3 feature. The quicker attainment of head-raising velocity maximum in clear relative to plain speech was aligned with the patterns characterizing this tone, where the time taken to achieve maximum head-raising velocity was shorter for Tone 3 than for the other tones. However, this feature is also identified as a tone-universal clear-speech modification (cf. Figure 3). Moreover, the clear-speech modifications of larger head raising and faster lip opening velocity did not involve Tone 3-specific features. Thus, Tone 3 clear-speech

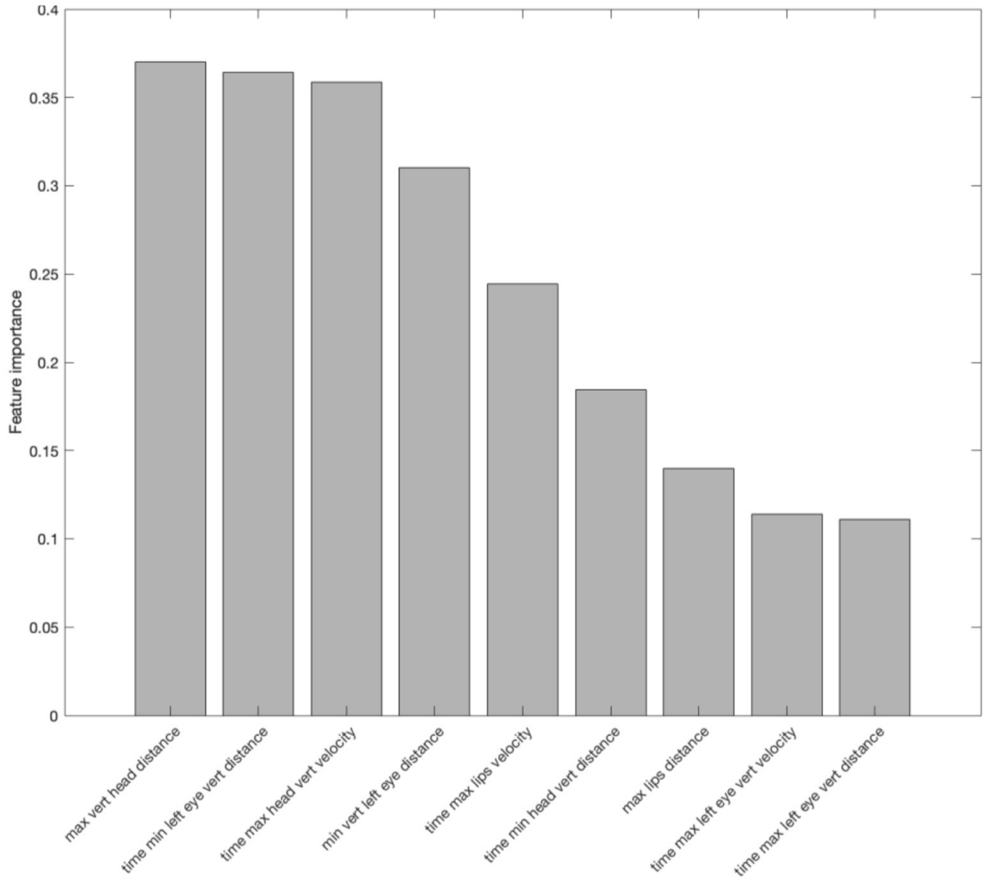

FIGURE 8
Important features identified by random forest in differentiating the two speech styles (clear and plain) in Tone 3. The y-axis shows the feature importance measured by the random forest. The feature importance measures the increase in classification error if the values of the feature are permuted. The greater the error the higher the importance of the feature.

modifications essentially adopt universal features characterizing clear-speech tone.

Tone 4 clear-plain differences made use of one Tone 4-specific feature. That is, the time taken for head-raising velocity to reach maximum, which was shorter for Tone 4 than for most of the other tones, was even shorter in clear than in plain speech, suggesting faster return of the head to the resting position after the head lowering gesture for the falling Tone 4. However, this feature, along with the shorter time taken for eyebrow-raising velocity maximum, was also consistent with the patterns across tones. Additionally, "the time taken for the distance of lip-closing to reach maximum", occurred later in clear than in plain style, suggesting longer duration in clear-speech production.

4. Discussion and concluding remarks

In this study, we examined how visual tonal cues are enhanced in clear speech in the production of Mandarin Chinese tones. As tone production lacks a direct association with vocal tract configurations, it is believed to be less distinguishable through visible articulatory movements. The question thus raised in this study was, in the production of clear-speech tones, whether any modifications of the visual articulatory features strengthen overall visual saliency (signal-based) or augment tone-specific distinctions (code-based). To this end, we compared which visual cues were adopted in clear speech across tones (as evidence of signal-based modifications) and which ones were aligned with the category-defining features for each tone as identified in Garg et al. (2019) (as evidence of code-based modifications).

Through computer vision analyses, this study tracked and quantified 33 facial features associated with head, eyebrow, and lip movements to determine the distance, duration, and kinematic characteristics between each of the keypoints in clear vs. plain tone productions. A 2-step discriminant analysis based on random forest and subsequent mixed-effects modeling was performed, first across tones and then for each tone, to identify the visual features differentiating clear and plain tone productions and rank the importance of these features, and then compare the values of each feature in clear and plain speech to assess if they are significantly different.

The results show differences in visual features between the two speech styles from both cross-tone and within-tone comparisons. Overall, the difference between the two styles lies both in spatial and temporal features as indicated by changes in distance, duration, velocity and acceleration of lip, eyebrow and head movements

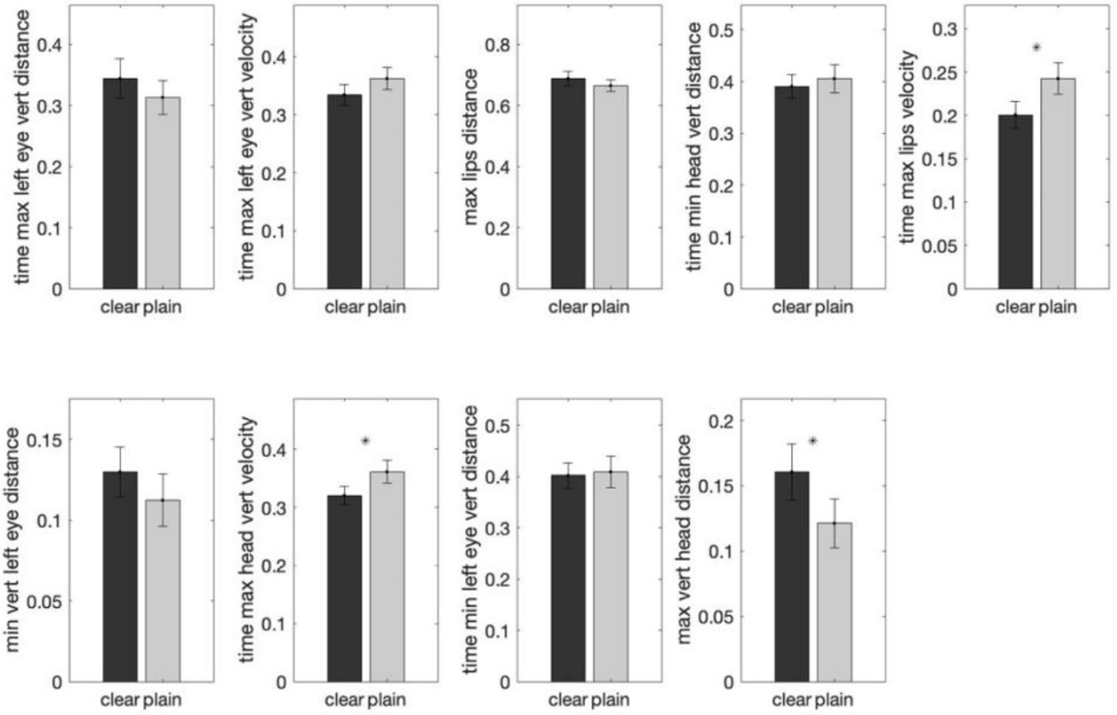

FIGURE 9
Comparisons of the group means of each of the nine important features (determined by random forest analyses) in clear and plain style in Tone 3. The y-axis shows the average normalized feature value for that particular style group. Three features were found to exhibit a statistically significant difference between the two speech styles by mixed-effects modeling. The error bar on each of the bar plots shows the standard error. A * above a bar represents a p-value smaller than 0.05.

associated with clear vs. plain tone productions. The common trend exhibited through these features indicates that signal-based plain-to-clear tone modifications are more dominant than code-based modifications, and are evidenced by both across tone and individual tone results.

Across tones, clear (compared to plain) productions show longer overall duration, larger maximum displacement of the head, eyebrows and lips and faster arrival at these movement peaks. First, the larger displacement maxima and longer duration indicate that clear-speech production of all the tones involves more extended articulatory trajectories, and consequently, takes longer to complete. Such patterns are consistent with previous studies revealing exaggerated articulation in clear speech segments. For example, studies on vowel articulation have consistently revealed longer duration along with greater articulatory movements (involving larger lip and jaw displacement across vowels) for clear relative to plain speech (Kim and Davis, 2014; Tang et al., 2015). Similar exaggerated articulatory activities have also been identified at the suprasegmental level such as long and short vowels (Šimko et al., 2016) as well as lexical tones (Han et al., 2019). Moreover, aside from these spatial features, the current results additionally reveal that, despite the longer distance, clearly produced tones generally reach movement peak positions faster. Such a combination of motion may consequently make the visual cues more prominent, thus enhancing the saliency of tones in clear speech. These findings consistently demonstrate signal-based modifications in clear-speech production across tones through both distance-based and time-based changes with overall enhancement of visual saliency.

Results of individual tone analyses corroborate the patterns across tones, revealing signal-based modifications predominantly. In addition to these general patterns, the current individual-tone findings are particularly noteworthy in that they strengthen the signal-based nature of clear-speech tone modifications in three ways. First, certain tone-general modifications are found to be incompatible with the inherent characteristics of individual tones. For example, the Tone 1 plain-to-clear modification followed the tone-general pattern of larger head lowering. However, as a level tone, Tone 1 inherently involves minimal head and eyebrow movements, presumably attributable to its small variation in pitch (Yehia et al., 2002; Munhall et al., 2004; Kim et al., 2014; Garg et al., 2019). Thus, it appears that signal-based information was adopted in clear Tone 1 modification even when it is in conflict with the intrinsic characteristics of this tone. Second, although some modifications involve tone-characterizing features, they are also aligned with universal clear-speech patterns. For example, for Tone 3, the quicker attainment of head-raising velocity maximum in clear relative to plain speech is aligned with the tone-general patterns as well as being an intrinsic property of this tone. Consequently, such tone-specific adjustments cannot be regarded as code-based alone. Third, significant tone-specific features fail to exhibit changes in clear speech. Notably, Tone 2 and Tone 3, as dynamic tones, have been identified as having multiple category-defining features (Garg et al., 2019). However, most of the crucial

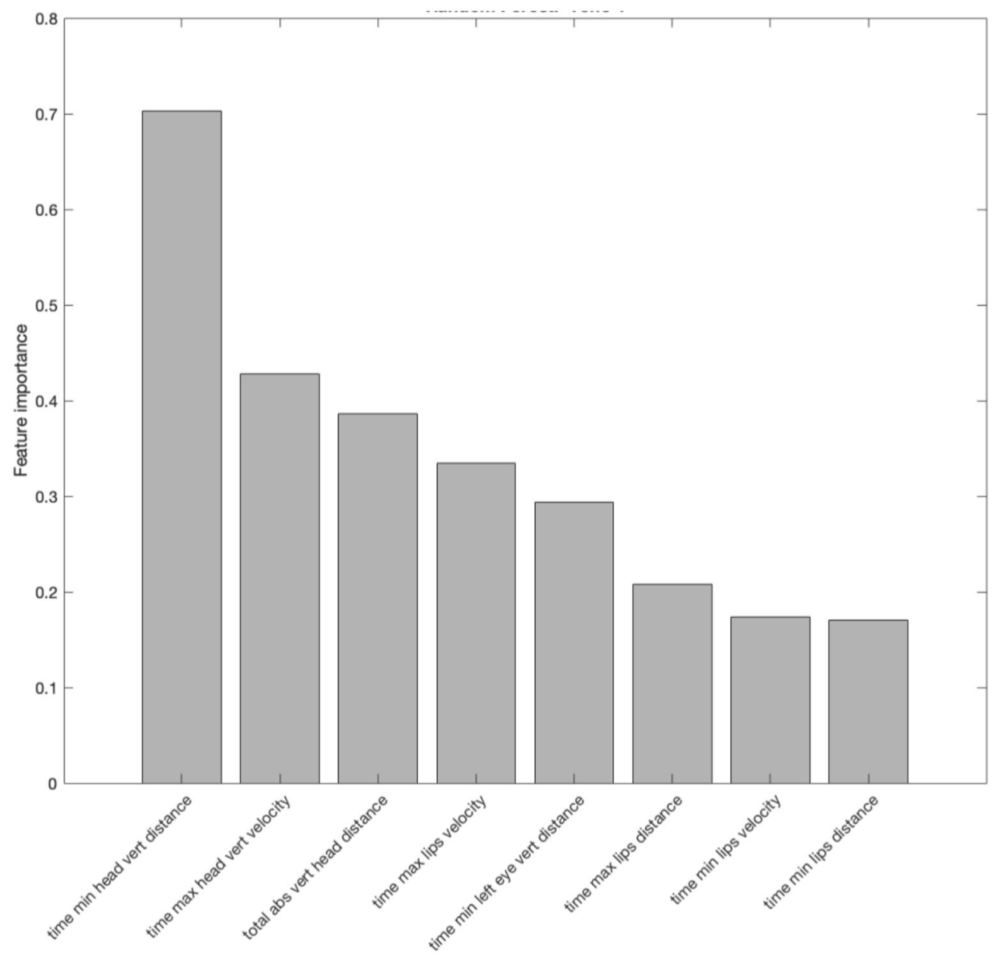

FIGURE 10
Important features identified by random forest in differentiating the two speech styles (clear and plain) in Tone 4. The y-axis shows the feature importance measured by the random forest. The feature importance measures the increase in classification error if the values of the feature are permuted. The greater the error the higher the importance of the feature is.

features of these tones, such as the low tone nature of Tone 3 (associated with head lowering, Garg et al., 2019) or its dynamicity (associated with lip closing and raising, Attina et al., 2010), did not exhibit corresponding modifications in clear speech. Taken together, consistent with the cross-tone results, these individual tone patterns suggest that signal-based cues outweigh code-based ones in clear-speech modification.

Therefore, unlike the patterns found at the segmental level for consonants and vowels showing signal- and code-based clear-speech modification working in tandem (Smiljanić, 2021), the current results suggest that visual clear-speech tone modifications primarily do not rely on code-based, tone-specific cues. Although previous findings on tone articulation indeed suggest alignments of facial movements with spatial and temporal pitch movement trajectories of individual tones (Attina et al., 2010; Garg et al., 2019; Han et al., 2019), most of these cues were not adopted in making tone-specific adjustments in clear speech. One possibility could be that the visual tonal cues, which are shown to be based on spatial and temporal correspondence to acoustic (F0) information rather than a direct association with vocal tract configuration (as is the case for segmental production), are not adequately distinctive (Hannah et al., 2017; Burnham et al., 2022). This is especially true for lip movements, which have been found to be less reliable in differentiating tones (Attina et al., 2010; Garg et al., 2019). Previous segmental studies suggest a trade-off in clear speech production between cue enhancements and maintenance of sound category distinctions (Lindblom, 1990; Ohala, 1995; Smiljanić, 2021). Speakers have been found to refrain from making clear-speech adjustments which would blur category distinctions (Leung et al., 2016; Smiljanić, 2021). In the case of the current study, the speakers may have more readily adopted the universal features that strengthen overall visual saliency since enhancing tone-specific features cannot reliably distinguish different tone categories.

Finally, it is worth noting that the acoustic analysis of the same data set by our research team (Tupper et al., 2021) has also revealed that the speakers primarily utilize signal-based acoustic changes (longer duration, higher intensity) in clear-speech tone modifications rather than code-based F0 changes that enhance the contrast between tones. This may also explain the lack of code-based articulatory modifications in the current study, given the

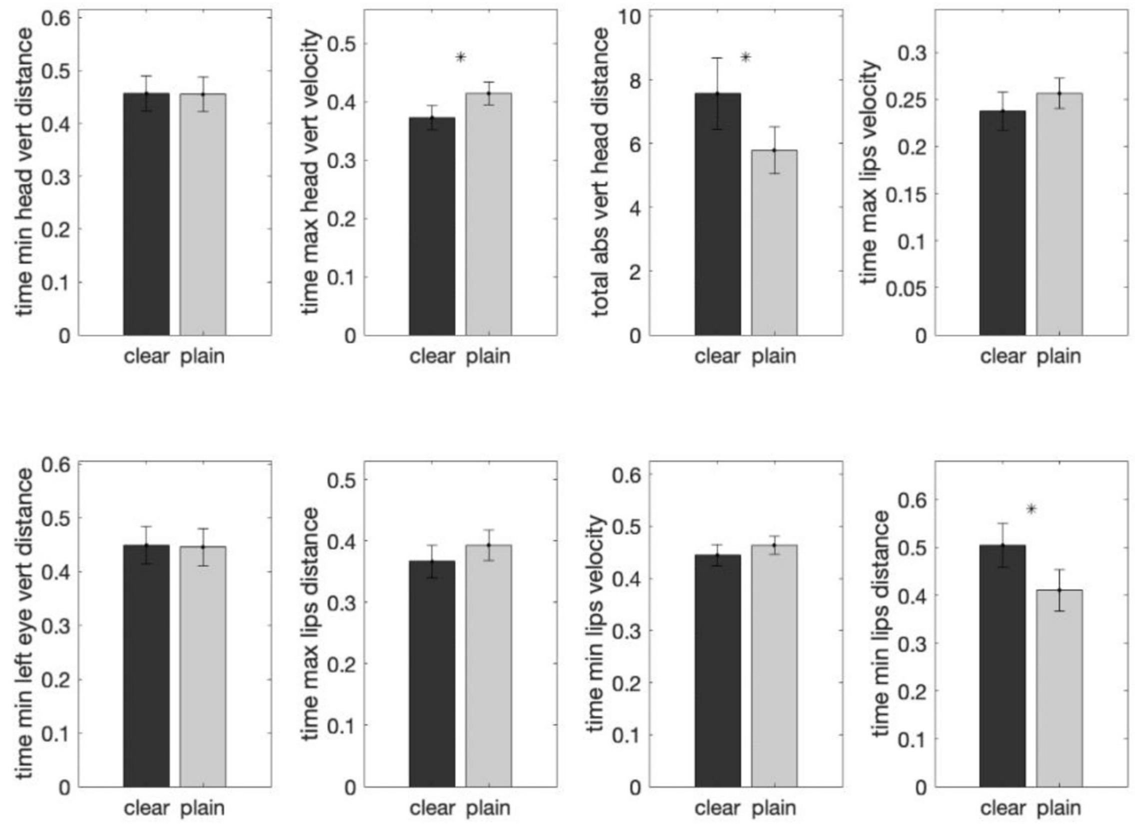

FIGURE 11
Comparisons of the group means of each of the eight important features (determined by random forest analyses) in clear and plain style in Tone 4. The y-axis shows the average normalized feature value for that particular style group. Three features were found to exhibit a statistically significant difference between the two speech styles by mixed-effects modeling. The error bar on each of the bar plots shows the standard error. A * above a bar represents a p-value smaller than 0.05.

presumed audio-spatial correspondence between pitch and visual articulatory movements (Connell et al., 2013; Garg et al., 2019). These findings lead to the subsequent question as to whether these articulatory and acoustic adjustments in clear speech benefit tone intelligibility and whether these universal saliency enhancing cues affect the perception of individual tones differently. The latter could in turn help disentangle the signal- vs. code-based nature of clear tone production.

Data availability statement

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

Ethics statement

The studies involving human participants were reviewed and approved by Office of Research Ethics, Simon Fraser University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

SG wrote the MATLAB code and analysis and helped in preparing the draft. GH provided feedback and suggestions in the analysis and reviewed the analysis and provided computing resources to perform the analysis. JS and AJ helped with the problem question, reviewed the analysis, and provided feedback with the writing. YW helped with the problem question, reviewed the analysis, and wrote the introduction and discussion and helped with the other writing. All authors contributed to the article and approved the submitted version.

Funding

This project has been supported by research grants from the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant No. 2017-05978) and the Social Sciences and Humanities Research Council of Canada (SSHRC Insight Grant No. 435-2019-1065).

Acknowledgments

We thank Lisa Tang for helping with the code and experiments as well as the write-up. We also thank Keith Leung, Jane Jian, Charles Turo, and Dahai Zhang from the SFU Language and Brain Lab for their assistance, as well as WestGrid and Compute Canada for their IT support.

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.

References

Attina, V., Gibert, G., Vatikiotis-Bateson, E., and Burnham, D. (2010). "Production of Mandarin lexical tones: Auditory and visual components," in *Proceedings of International Conference on Auditory-visual Speech Processing (AVSP) 2010*, Hakone.

Bradlow, A. R., and Bent, T. (2002). The clear speech effect for non-native listeners. *J. Acoust. Soc. Am.* 112, 272–284. doi: 10.1121/1.1487837

Burnham, D., Ciocca, V., and Stokes, S. (2001). *Auditory-visual perception of lexical* tone. In, P. Dalsgaard, B. Lindberg, H. Benner, and Z. H. Tan, (eds.), *Proceedings of the 7th Conference on Speech Communication and Technology, EUROSPEECH 2001*, Scandinavia, pp. 395–398. doi: 10.21437/Eurospeech.2001-63

Burnham, D., Reynolds, J., Vatikiotis-Bateson, E., Yehia, H., Ciocca, V., Morris, R., et al. (2006). "The perception and production of phones and tones: The role of rigid and non-rigid face and head motion," In *Proceedings of the International Seminar on Speech Production 2006*, Ubatuba.

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 Commun.* 141, 40–55. doi: 10.1016/j.specom.2022.03.011

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., Espesser, R., et al. (1996). About the relationship between eyebrow movements and F0 variations. *Proceedings of the ICSLP* (pp. 2175–2179), Philadelphia. doi: 10.21437/ICSLP.1996-551

Chen, T. H., and Massaro, D. W. (2008). Seeing pitch: Visual information for lexical tones of Mandarin-Chinese. *J. Acoust. Soc. Am.* 123, 2356–2366. doi: 10.1121/1.2839004

Connell, L., Cai, Z. G., and Holler, J. (2013). Do you see what i'm singing? *visuospatial movement biases pitch perception. Brain and Cognition* 81, 124–130. doi: 10.1016/j.bandc.2012.09.005

Cooke, M., and Lu, Y. (2010). Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. *J. Acoust. Soc. Am.* 128, 2059–2069. doi: 10.1121/1.3478775

Cvejic, E., Kim, J., and Davis, C. (2010). Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Commun.* 52, 555–564. doi: 10.1016/j.specom.2010.02.006

Desai, S., Stickney, G., and Zeng, F. G. (2008). Auditory-visual speech perception in normal-hearing and cochlear-implant listeners. *J. Acoust. Soc. Am.* 123, 428–440. doi: 10.1121/1.2816573

Dohen, M., and Loevenbruck, H. (2005). "Audiovisual Production and Perception of Contrastive Focus in French: a multispeaker study," in *Interspeech/Eurospeech* 2005 (pp. p-2413). doi: 10.21437/Interspeech.2005-49

Dohen, M., Loevenbruck, H., and Hill, H. C. (2006). "Visual correlates of prosodic contrastive focus in French: Description and inter-speaker variability," In *Speech Prosody*, edsR. Hoffmann and H. Mixdorff, 221–224.

Ferguson, S. H., and Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.* 112, 259–271. doi: 10.1121/1.1482078

Ferguson, S. H., and Kewley-Port, D. (2007). Talker differences in clear and conversational speech: Acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research* 50, 1241–1255. doi: 10.1044/1092-4388(2007/087)

Flecha-García, M. L. (2010). Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Commun.* 52, 542–554. doi: 10.1016/j.specom.2009.12.003

Gagné, J. P., Masterson, V., Munhall, K. G., Bilida, N., and Querengesser, C. (1994). Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech. *J. Academy Rehabil. Audiol.* 27, 135–158.

Gagné, J. P., Rochette, A. J., and Charest, M. (2002). Auditory, visual and audiovisual clear speech. *Speech Commun.* 37, 213–230. doi: 10.1016/S0167-6393(01)00012-7

Garg, S., Hamarneh, G., Jongman, A., Sereno, J. A., and Wang, Y. (2019). Computer-vision analysis reveals facial movements made during Mandarin tone production align with pitch trajectories. *Speech Commun.* 113, 47–62. doi: 10.1016/j.specom.2019.08.003

Garnier, M., Ménard, L., and Alexandre, B. (2018). Hyper-articulation in Lombard speech: an active communicative strategy to enhance visible speech cues?. *J. Acoust. Soc. Am.* 144, 1059–1074. doi: 10.1121/1.5051321

Han, Y., Goudbeek, M., Mos, M., and Swerts, M. (2019). Effects of modality and speaking style on Mandarin tone identification by non-native listeners. *Phonetica* 76, 263–286. doi: 10.1159/000489174

Hannah, B., Wang, Y., Jongman, A., Sereno, J. A., Cao, J., Nie, Y., et al. (2017). Cross-modal association between auditory and visuospatial information in Mandarin tone perception in noise by native and non-native perceivers. *Front. Psychol.* 8, 2051. doi: 10.3389/fpsyg.2017.02051

Hazan, V., and Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *J. Acoust. Soc. Am.* 130, 2139–2152. doi: 10.1121/1.3623753

Helfer, K. S. (1997). Auditory and auditory-visual perception of clear and conversational speech. *J. Speech Lang. Hearing Res.* 40, 432–443. doi: 10.1044/jslhr.4002.432

Ishi, C. T., Ishiguro, H., and Hagita, N. (2007). Analysis of head motions and speech in spoken dialogue. *INTERSPEECH* 2007, 8th. *Annual Conference of the International Speech Communication Association* 2, 670–673. doi: 10.21437/Interspeech.2007-286

Kim, J., Cvejic, E., and Davis, C. (2014). Tracking eyebrows and head gestures associated with spoken prosody. *Speech Commun.* 57, 317–330. doi: 10.1016/j.specom.2013.06.003

Kim, J., and Davis, C. (2014). Comparing the consistency and distinctiveness of speech produced in quiet and in noise. *Comp. Speech Lang.* 28, 598–606. doi: 10.1016/j.csl.2013.02.002

Kim, J., Sironic, A., and Davis, C. (2011). Hearing speech in noise: Seeing a loud talker is better. *Perception* 40, 853–862. doi: 10.1068/p6941

Krause, J. C., and Braida, L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *J. Acoust. Soc. Am.* 115, 362–378. doi: 10.1121/1.1635842

Lander, K., and Capek, C. (2013). Investigating the impact of lip visibility and talking style on speechreading performance. *Speech Commun.* 55, 600–605. doi: 10.1016/j.specom.2013.01.003

Leung, K. K., Jongman, A., Wang, Y., and Sereno, J. A. (2016). Acoustic characteristics of clearly spoken English tense and lax vowels. *J. Acoust. Soc. Am.* 140, 45–58. doi: 10.1121/1.4954737

Lindblom, B. (1990). *Explaining phonetic* variation: A sketch of the HandH theory. In W. Hardcastle and A. Marchal (Eds.), Speech Production and Speech Modelling (pp. 403–439). Dordrecht: Springer. doi: 10.1007/978-94-009-2037-8_16

Lu, Y., and Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. *J. Acoust. Soc. Am.* 124, 3261–3275. doi: 10.1121/1.2990705

Maniwa, K., Jongman, A., and Wade, T. (2008). Perception of clear fricatives by normal-hearing and simulated hearing-impaired listeners. *J. Acoust. Soc. Am.* 123, 1114–1125. doi: 10.1121/1.2821966

Maniwa, K., Jongman, A., and Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *J. Acoust. Soc. Am.* 125, 3962–3973. doi: 10.1121/1.2990715

Moon, S. J., and Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *J. Acoust. Soc. Am.* 96, 40–55. doi: 10.1121/1.410492

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., and Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychol. Sci.* 15, 133–137. doi: 10.1111/j.0963-7214.2004.01502010.x

Ohala, J. (1995). Clear speech does not exaggerate phonemic contrast. In Proceedings of the 4th European Conference on Speech Communication and Technology (pp. 1323–1325). doi: 10.21437/Eurospeech.1995-344

Paul, J., and Dupont, P. (2015). Inferring statistically significant features from random forests. *Neurocomputing* 150, 471–480. doi: 10.1016/j.neucom.2014.07.067

Perkell, J. S., Zandipour, M., Matthies, M. L., and Lane, H. (2002). Economy of effort in different speaking conditions. *I*. A preliminary study of intersubject differences and modeling issues. *J. Acoust. Soc. Am.* 112, 1627–1641. doi: 10.1121/1.1506369

Redmon, C., Leung, K., Wang, Y., McMurray, B., Jongman, A., Sereno, J. A., et al. (2020). Cross-linguistic perception of clearly spoken English tense and lax vowels based on auditory, visual, and auditory-visual information. *J. Phon.* 81, 100980. doi: 10.1016/j.wocn.2020.100980

Reid, A., Burnham, D., Kasisopa, B., Reilly, R., Attina, V., Rattanasone, N. X., et al. (2015). Perceptual assimilation of lexical tone: the roles of language experience and visual information. *Attent. Percep. Psychophysics* 77, 571–591. doi: 10.3758/s13414-014-0791-3

Scarborough, R., Keating, P., Mattys, S. L., Cho, T., and Alwan, A. (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. *Lang. Speech* 52, 135–175. doi: 10.1177/0023830909103165

Šimko, J., Benuš, Š., and Vainio, M. (2016). Hyperarticulation in Lombard speech: Global coordination of the jaw, lips and the tongue. *J. Acoust. Soc. Am.* 139, 151–162. doi: 10.1121/1.4939495

Smiljanić, R. (2021). "Clear speech perception: Linguistic and Cognitive benefits," in The *Handbook of Speech Perception*, eds Pardo, J.S., Nygaard, L.C., Remez, R.E., and Pisoni, D.B., 2nd Edition. Wiley. pp. 177-205. doi: 10.1002/9781119184096.ch7

Smiljanić, R., and Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *J. Acoust. Soc. Am.* 118, 1677–1688. doi: 10.1121/1.2000788

Smiljanić, R., and Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Lang. Linguist. Compass* 3, 236–264. doi: 10.1111/j.1749-818X.2008.00112.x

Srinivasan, R. J., and Massaro, D. W. (2003). Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Lang. Speech* 46, 1–22. doi: 10.1177/00238309030460010201

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Swerts, M., and Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *J. Phon.* 36, 219–238. doi: 10.1016/j.wocn.2007.05.001

Swerts, M., and Krahmer, E. (2010). Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions. *J. Phon.* 38, 197–206. doi: 10.1016/j.wocn.2009.10.002

Tang, L. Y., Hannah, B., Jongman, A., Sereno, J., Wang, Y., Hamarneh, G., et al. (2015). Examining visible articulatory features in clear and plain speech. *Speech Commun.* 75, 1–13. doi: 10.1016/j.specom.2015.09.008

Tasko, S. M., and Greilick, K. (2010). Acoustic and articulatory features of diphthong production: a speech clarity study. J. *Speech Lang. Hear*. Research, 53, 84–99. doi: 10.1044/1092-4388(2009/08-0124)

Traunmüller, H., and Öhrström, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *J. Phon.* 35, 244–258. doi: 10.1016/j.wocn.2006.03.002

Tupper, P., Leung, K. K., Wang, Y., Jongman, A., and Sereno, J. A. (2018). Identifying the distinctive acoustic cues of Mandarin tones. *J. Acoust. Soc. Am.* 144, 1725–1725. doi: 10.1121/1.5067655

Tupper, P., Leung, K. W., Wang, Y., Jongman, A., and Sereno, J. A. (2021). The contrast between clear and plain speaking style for Mandarin tones. *J. Acoust. Soc. Am.* 150, 4464–4473. doi: 10.1121/10.0009142

Van Engen, K. J., Phelps, J. E., Smiljanic, R., and Chandrasekaran, B. (2014). Enhancing speech intelligibility: Interactions among context, modality, speech style, and masker. *J. Speech Lang. Hearing Res.* 57, 1908–1918. doi: 10.1044/JSLHR-H-13-0076

Wang, Y., Behne, D. M., and Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *J. Acoust. Soc. Am.* 124, 1716–1726. doi: 10.1121/1.2956483

Yehia, H. C., Kuratate, T., and Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *J. Phon.* 30, 555–568. doi: 10.1006/jpho.2002.0165

Zhao, Y., and Jurafsky, D. (2009). The effect of lexical frequency and Lombard reflex on tone hyperarticulation. *J. Phon.* 37, 231–247. doi: 10.1016/j.wocn.2009.03.002

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 17 May 2023
DOI 10.3389/fcomm.2023.1090030

Check for updates

OPEN ACCESS

EDITED BY
Plinio Almeida Barbosa,
State University of Campinas, Brazil

REVIEWED BY
Donna Mae Erickson,
Haskins Laboratories, United States
Daniel Hirst,
UMR7309 Laboratoire Parole et Langage (LPL),
France

*CORRESPONDENCE
Dafydd Gibbon
✉ gibbon@uni-bielefeld.de

RECEIVED 04 November 2022
ACCEPTED 20 April 2023
PUBLISHED 17 May 2023

CITATION
Gibbon D (2023) Rhythm pattern discovery in Niger-Congo story-telling.
Front. Commun. 8:1090030.
doi: 10.3389/fcomm.2023.1090030

Rhythm pattern discovery in Niger-Congo story-telling

Dafydd Gibbon *

Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany

Description of interactive oral story-telling in West African Niger-Congo languages, "orature" in contrast to "literature", has traditionally been firmly in the domain of anthropological linguistics. The discourse structures of narrator-responder interaction and call-response chanted interludes and their prosody are an open challenge to discourse analysts, linguists and phoneticians. Two orature examples from related Niger-Congo languages, recorded during fieldwork in Côte d'Ivoire, are analyzed from a macrostructural acoustic phonetic perspective in terms of realtime rhythms, and then compared with other related orature examples and examples of reading aloud. In this transdisciplinary methodology, long time domains of up to 5 min in duration and beyond are studied using long-term spectra and spectrograms of the amplitude modulation of speech. Long-distance timing regularities and their variation during story-telling exchanges are analyzed in detail and explained in terms of Rhythm Formant Theory (RFT), a further development of Speech Modulation Theory, and its associated methodology Rhythm Formant Analysis (RFA).

KEYWORDS

Rhythm Formant Theory, Modulation Theory, speech rhythm, orature, story-telling, Ega, Anyi, Niger-Congo

> I would define, in brief, the Poetry of words as *The Rhythmical Creation of Beauty*.
> Edgar Allen Poe, 1850, *The Poetic Principle*.

1. Background

1.1. Orature and its analysis

The purely orally performed narrative traditions of West Africa defy the literary analysis techniques which apply to written literature and its oral performance, and call for a transdisciplinary approach which takes account of the specific realtime rhythmic and melodic features of speech. One of the goals of the present study is exploration of the data with a methodology which facilitates the meeting of the interpretative methods of linguistic and anthropological studies of traditional but partly improvised extempore orature—unwritten poetry, narrative and drama—with phonetic methods, using signal processing coupled with data visualization. The domain is the rhetorical and poetic skills of West African orature, exemplified by rhythm variation and consistency in orature performances.

On the one hand, traditional qualitative approaches to language in the humanities use hermeneutic methods based on native speaker intuition and understanding. These methods range from literary studies of written poetry, narrative and drama and anthropological studies of orature to linguistic analysis of vocabulary and grammar and phonetic analysis of words, sentences and their prosody, often with support from the formal techniques of algebra and logic.

On the other hand, in the physical sciences measurements and numerical methods are used to analyze patterns in physically caused traces of speaking (or writing) events recorded in computer memory. Generalizations are induced with techniques from machine learning and artificial intelligence. In respect of language and speech the main relevant fields range from computational corpus linguistics and phonetic signal analysis to the "Large Language Models" (LLMs) and speech cloning of current language and speech engineering.

Between these two methodological poles of hermeneutic and causal explanation lie several hybrid disciplines, including linguistic phonetics and corpus linguistics, dialectometry, stylometry, experimental psycholinguistics and sociolinguistics, which apply quantitative methods to qualitative categories such as speech sounds and written characters, words, phrases and sentences, and which ground these categories in physical realities of stored or real-time speech and writing. One popular set of hybrid methods from these areas, in which many other disciplines such as geographical information systems and archaeology cooperate, comes together under the umbrella term of "digital humanities". This "metadiscipline" was originally concerned with the digitization and preservation of ancient manuscripts, but in the meantime has undergone syncretistic expansion with absorption of methods from other hybrid disciplines (Ekpenyong and Udoh, 2022).

1.2. Objectives

The topic and methods of the research reported here are loosely related to digital humanities. The disciplines from this area which are most relevant to the present study are dialectometry (Nerbonne and Kretzschmar, 2003) and stylometry (Rybicki and Eder, 2011; Savoy, 2020), which combine corpus linguistic analysis with unsupervised and supervised machine learning in domains which go beyond sentences and investigate, for example, similarities and differences between entire novels and styles. With distance metrics, distance networks and hierarchical clustering, patterns of similarity between literary genres across languages can be classified and authorships of previously anonymous or pen-named works determined, in literary studies, linguistics and forensic text analysis (Juola, 2015).

The present study takes a comparable approach to spoken discourse in the later sections, adding a macrostructural discourse-phonetic dimension. While stylometry typically deals with written texts, except in specific forensic applications, and dialectology deals with both written texts and transcriptions, the present research applies these classification methods directly to physical measurements of spoken language, applying methods from acoustic phonetics and unsupervised machine learning.

The main goal is to provide explanations of perceived realtime rhythm variation in orature, both within a given orature event, and between spoken language events in different languages and in other language varieties such as reading aloud. One common type of causal explanation would focus on relating orature rhythms to learned repetition and rehearsal behavior; another would relate it to speech production processes and memory; a functional explanation might focus on mean opinion surveys (MOS) of perceived rhythmicity; a quantitative approach to explanation would develop correlational models based on neurophonetic measurements. Since no data of these types are available for the orature domain in view in the present study, nor can they be feasibly obtained, these kinds of explanation are not available.

Consequently, other more traditional types of explanation are invoked: functional explanation in terms of hermeneutic understanding of identifiable phases in orature events; structural explanation in terms of a static model of participant roles and a dynamic model of participant interaction; causal explanation in terms of the grounding of realtime rhythms, in terms both of variation in durations of annotated speaker turn events and of acoustic analysis of the low frequency (LF) spectrum of the complete event. The causal line of explanation is accompanied by a falsification chain from simpler and less adequate analyses such as annotation mining to more complex and more adequate approaches such as spectrogram analysis. The falsification chain is an explanatory feature which enables distinguishing between more and less adequate theories.

Specifically, the aim is to discover variability in the rhythmic organization of extempore story-telling dialogues in two Niger-Congo languages of the Ivory Coast, Anyi and Ega, in comparison with read-aloud written stories in other related languages and speech styles (Gibbon et al., 2002). Anyi and Ega are typologically related, and both have similar ranges of complicating factors which condition details of the spoken rhythms of the languages, including lexical and morphosyntactic tone and mildly agglutinating morphology, which they share with other Niger-Congo languages. An additional motivation for the study is to contribute toward insightful description and documentation of these languages (Gibbon et al., 2004) in terms of the acoustic dimension of a multimodal analysis (Rossini and Gibbon, 2011).

1.3. Realtime rhythms

The present concern is not with the syllable, word and sentence structures which have been referred to in the context of metrical phonology as "linguistic rhythm" (Liberman and Prince, 1977).[1] The understanding of rhythm in the present context is more traditional and more diverse: rhythms are understood as sequences of regular beats in speech, music, song and dance, or related to events such as heartbeats, walking, chewing, or to interpersonal events such as handshaking and other bonding interactions. These rhythmic sequences of beats and pulses have a specific tempo and frequency, possibly with several simultaneous rhythms at different frequencies, and a beat sequence count of at least three (Nakamura and Sagisaka, 2011), a "minimal rhythm principle". A duration of at least three seconds for word rhythms, for example, will be referred to as "persistence" and persistence at the same frequency will be referred to as "resonance".

[1] The terms "word" and "sentence" are used in general senses, since grammar is not a prime concern in this context: "word" includes lexical words, phonological words and feet; "sentence" includes clauses and phrases.

In the present context, rhythms are understood in the following general terms:

> A rhythm is perceived, and can be measured, when at least three similarly structured events occur at approximately equal intervals in time and create an expectation of a further similarly structured event after a similar interval, subject to persistence and resonance conditions. Functionally, a rhythm implies an underlying principle of cohesion.

In the case of speech rhythms, the principle of cohesion is metalocutionary marking of a concurrent segment of speech of the same length. The definition is related to the interpretation in Dilley (2005) of pitch accent sequences in English as creating expectations of continuation, that is, lowering of entropy, and then termination, i.e., breaking the expectations. Formally, a speech rhythm is a low frequency (LF) oscillation which can be measured in the LF segment of the speech spectrum below about 10 Hz. The concept is taken up in the sections below on Rhythm Formant Theory[2].

The linguistic correlates of rhythmic beats are in general strong or stressed syllables, or salient words and phrases. The beats contrast syntagmatically with the intervening non-beat intervals, such as weak or unstressed syllables and non-salient words and phrases. Beats and non-beats approximate to the categories of ictus and remiss in poetic meter and in linguistic phonetics.

In modulation theoretic approaches these sequences have been modeled as oscillations between strong and weak states of some audible parameter at temporally regular intervals, possibly with simultaneous oscillations in different frequency ranges (Ohala, 1992; Cummins and Port, 1998; Barbosa, 2002). In an idealized model with completely regular oscillations at a constant frequency, the beats are isochronous (i.e., have identical temporal interval durations). However, this is an idealization. In reality, "equal timing" is relative and involves variation over smaller or larger frequency ranges, depending on the prosodic typology of different languages and on different genres and styles of speaking (Arvaniti, 2009; Kohler, 2009).

1.4. Overview

This exploratory study uses quantitative acoustic phonetic methods for discovering rhythm patterns associated with a domain which is initially described and modeled with interpretative methods. Results are judged *post hoc* on a qualitative case-by-case basis because of the sparseness of the available fieldwork data, not in terms of statistical significance, though the latter is implied by the use of clustering methods. The specific questions addressed in exploring the data are as follows:

1. Which overall utterance duration patterns can be measured in different turn types (narrator, responder, audience) during the narrative?
2. How rhythmical are these patterns?
3. Are there functional interpretations of the differences between rhythm patterns within orature sessions or between sessions in different languages?

The argumentation strategy starts in the present section with the intuition-based qualitative description of language varieties as the *explicandum*, followed by the introduction of the data and then moves toward a valid formal *explicans* via a falsification chain in subsequent sections. First, annotation mining in the time domain is investigated and duration irregularity indices are rejected as incomplete characterizations of rhythm. Second, annotation mining is used to visualize long-term duration patterns and intuitively observable cyclic duration patterns as potential rhythms. Third, the annotation-mining method is replaced by the more precise frequency domain approach of Rhythm Formant Theory (RFT) and its methodology Rhythm Formant Analysis (RFA), starting with analysis of spectral frequency peaks. The spectra turn out to be helpful, but not a complete solution since they lack temporal information. Fourth, the spectra are replaced by spectrograms, i.e., a sequence of shorter term spectra along the time axis, supporting the persistence and resonance properties of rhythm. Fifth, a first step in automatic induction is made, with similarity visualization by means of distance maps. Sixth, the distance maps are replaced by hierarchical clustering rendered in dendrograms, which show exact rhythm-based relations between the language varieties. In this way, the present account of rhythm reaches an explanatory level of theoretically well-founded and methodologically well-grounded causal analyses in addition to further functional explanation through interpretations of visualized spectra and spectrograms.

The data are described in Section 2. Section 3 reports on the results of examining duration patterns in two stories, one each from Anyi and Ega, using the hybrid qualitative-quantitative method of speech annotation mining. In Section 4, Rhythm Formant Theory (RFT) and its associated methodology, Rhythm Formant Analysis (RFA) are introduced and the theory is located within the framework of Speech Modulation Theory. In Section 5 quantitative results in the form of distance maps and dendrograms are constructed by generating vectors of spectral features and using them to compare a total of 11 oral narrative productions in 5 different languages (2 Anyi, 1 Bete, 6 Ega, 1 Ibibio, 1 Côte d'Ivoire French) and in different styles (formal, informal). In Section 6, the causal explications are validated by means of holistic visual comparison of spectrum and spectrogram patterns. Section 7 provides a summary, reviews the conclusions and suggests an outlook for applications and future work.

2. Data

2.1. Language characterization

The two languages represented in the core of the present study are Niger-Congo tone languages: Anyi (Anyin, Agni, ISO 639-3 *any*) and Ega (ISO 639-3 *ega*). Anyi is a Central Tano

[2] The expectation created by a rhythm is conceptually related to the econometric concept of Granger Causality in time series, in that context without necessarily implying a rhythm: "A time series X is said to Granger-cause Y if it can be shown [...] that those X values provide statistically significant information about future values of Y." (Wikipedia; p.c. by A. J. Gibbon).

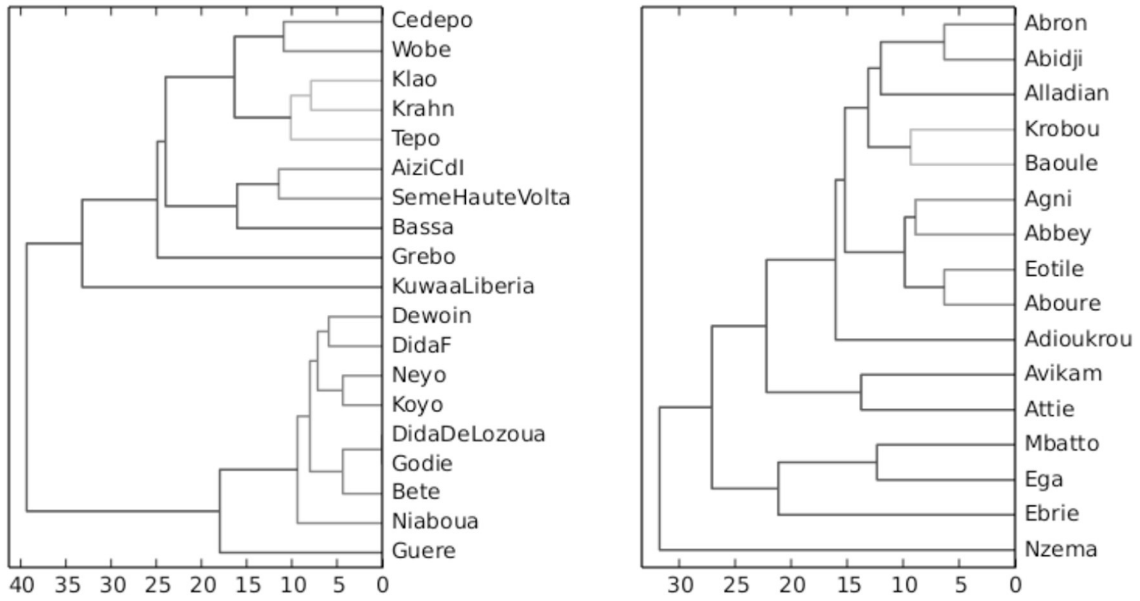

FIGURE 1
Hierarchical clustering of Côte d'Ivoire Kru **(left)** and Kwa **(right)** languages according to differences in phoneme inventories (greater difference shown by edge length and darker shade).

Kwa language spoken in south-eastern Côte d'Ivoire (Ahoua and Adouakou, 2009), and is represented in the study with recordings of two interactive oral narratives. Ega is a putative Kwa isolate spoken in south-central Côte d'Ivoire (Bole-Richard, 1983; Connell et al., 2002; Blench, 2015), enclaved in an Eastern Kru-speaking area by the language Dida (ISO 639-3 *dic*), and is represented with six interactive oral narratives. The recordings in both languages were made during fieldwork in Côte d'Ivoire in 2001 and 2002. In each case, the interactive narrative scenario is the same and is characteristic of story-telling scenarios in the Kwa languages (Berry and Spears, 1991; Ninan et al., 2016; Lô et al., 2020).

In order to provide an initial grounding in terms of linguistic relationships between the Kwa language group and the enclaving Kru language group, hierarchical clusters based on the phoneme inventories of the languages are shown in Figure 1, based on previous studies (Gibbon, 2014, 2016). The sources of the phoneme similarity data are the inventories tabulated in the language atlases for the Kwa (Hérault, 1983) and Kru (Marchese, 1983) languages of Côte d'Ivoire. The dendrograms are constructed automatically from triangular distance tables which are based on pairwise Levenshtein Distance between the phoneme inventories formatted as strings (not unlike triangular distance tables in geographical maps) and are clustered hierarchically using complete linkage (farthest neighbor linkage). The automatically calculated subgroups are highly compatible with the general linguistic classifications in the language atlases.

In addition to Anyi and Ega, a read-aloud narrative in educated Ivorian French (abbreviated "CdIFr") by a native-speaker of Anyi is included, as well as two readings of translations of the benchmark parable used by the International Phonetic Association, *The North Wind and the Sun*, in two further Niger-Congo languages, Daloa Bete (Eastern Kru, Ivory Coast, ISO 639-3 *bev*), closely related to the Dida language which enclaves the Ega-speaking region, and reading in Ibibio (Lower Cross, Nigeria, ISO 639-3 *ibb*). The reason for including the reading genre in non-Kwa Niger-Congo languages is to investigate possible language-transcending genre differences. The read-aloud Bete and Ibibio stories are translations of the Aesop fable *The North Wind and the Sun*. The Ibibio recording is accessible via (Urua, 2004).

The recordings differ in length, which leads to differences in frequency resolution and energy distributions of spectra in the later analyses, but not in the actual frequencies. The durations are shown in Table 1. In the analyses these differences are length-normalized in order to make comparison viable.

2.2. The orature scenario

In the interactive narrative scenario a narrator addresses a responder, turning occasionally to the audience. The narrators are respected for their story-telling skills and may be the *chef du village*, in the case of Ega, or a popular story-teller, in the Anyi example. The responder provides ritual encouragement, skeptical or neutral back-channel feedback for the narrator, and may be either a designated village responder (Figure 2 left) or selected *ad hoc* by the narrator (Figure 2 right). The hand of the narrator is blurred in the photo because she is in the process of selecting her responder by resting her hand on his thigh, a conventional gesture in Kwa communities which signifies a request for support in a communal gathering. The audience provides occasional spontaneous backchannel responses of surprise, horror or amusement at highlights in the story.

The selected Anyi and Ega stories both follow the traditional genre of parable or fable, involving two participants who are

TABLE 1 Durations of recordings.

Anyi interactive narrative:					
Anyi 2:	1 m:53.35 s,	Anyi 17:	4 m:45.24 s		
Ega interactive narrative (one speaker):					
Ega 1:	5 m:00.43 s,	Ega 2:	8 m:33.41 s,	Ega 3:	5 m:30.03 s
Ega 4:	7 m:17.03 s,	Ega 7:	5 m:35.81 s,	Ega 10:	9 m:16.26 s
Readings:					
Bete:	0m:49.82s,	Ibibio:	1m:06.94,	CdIFr:	6m:35.03s

competing or involved in a misunderstanding, and ending with a moral. One Anyi story (conte17) has a classic plot about an elephant who challenges a mouse to a race, and is shocked to find that the mouse can keep up, somewhat reminiscent of the well-known fable about the tortoise and the hare. The moral of the fable is that no-one should underestimate anyone.

One Ega story (conte2) is about a bird who tries to warn a bride, who is walking to meet her new bridegroom in another village (following the custom of exogamous marriage), that her jilted ex-bridegroom is planning to harm her, but the girl does not understand, continues, encounters the ex-bridegroom, is reproached by him, insults him, and meets her end as prophesied.

The moral of the Ega parable is not that the girl or the ex-bridegroom should have behaved more honorably, but that one should learn other languages in order to survive. The ironic point is that the bird's message is not just birdsong but is in Dida, the enclaving Kru language of the bridegroom, which the girl had not yet mastered, and not in the bride's endangered native tongue Ega. This metalinguistic moral was accompanied by a glance and a gesture toward the fieldworkers as learners of the Ega language, whose ambivalent participant-nonparticipant role was thereby underlined.

The narrative development of this parable starts with the exposition by the narrator, who presents an example of the appropriate audience response. There follows a very brief pause and an iterative pattern evolves: either continuation of the narrative or the responder's backchannel interjection, leading back to the narrative. After the narrative-backchannel sequence, the narrator initiates a chanting cycle with a call and a chanted response by the choir which is constituted by the audience. This overall cycle continues until the end of the story and the statement of the moral.

2.3. Modelling orature

The role pattern in the story-telling and recording scenario is shown as a static structure in Figure 3: the narrator-caller, in the center, has the main roles, supported by the responder, and the audience has two main roles, as listeners and as a chanting choir in response to the caller's chants. The observing fieldworkers have an ambivalent role as eavesdroppers, as it were, between participation and non-participation.

The cyclic dynamics of this variety of poetic orature are summarized visually in the iterative transition network depicted in Figure 4. The network formalizes four cycles or iterative principles: the narrative-pause cycle, the narrative-backchannel cycle and the call-response cycle, which together constitute an overall global cycle. This pattern applies both to the Anyi and the Ega narratives.

In this model, the narrator introduces the topic (a move from state S0 to S1) and after a brief rhetorical pause (S1 to to S3), the narrator either continues (a subordinate cyclic move from S3 to S1) and continues with the narrative or receives backchannel feedback from the responder (a different subordinate cyclic move from S3 to S2, then S2 to S0). Alternatively, at this point the narrator chants a call to the audience (S3 to S4), who respond as a choir with a chant (S4 to S5), after which the narrator has the option of another call-response sequence (a subordinate cyclic move S5 to S4), or of continuing via a pause (S5 to S0) with the next phase of the narrative, and starting the next main cycle (from S0). After several main and subordinate cycles the narrator formulates the moral of the story, ending via a pause (at terminal state S3).

The value of the iterative transition network in the present context is that it formally captures the turn interaction pattern and its relations with the long-term discourse-level rhythm and melody patterns.

2.4. Data selection and processing

Two data sets are analyzed. The smaller data subset consists of the Anyi and Ega stories, and the larger data set includes the subset, and contains a total of 11 recordings in five languages (2 Anyi, 6 Ega, 1 Bete, 1 Ibibio, 1 Ivory Coast French) including two further dimensions: two oral genres (interactive extempore narrative and reading aloud) and two styles (formal and informal). The languages are typologically or areally related. The selection is designed to facilitate the pilot analysis without too many typological variables.

The recordings were originally annotated shortly after recording using the TASX-annotator tool (Milde, 2002). The TASX XML format was converted for further processing to a CSV format, and a small number of annotation errors were also corrected. For the recent annotations used in the present study, the Praat phonetic workbench (Boersma, 2001) was used and the TextGrid files were also converted into CSV files for further processing. The signal processing analyses and generation of the figures were implemented in Python using the NumPy, SciPy, MatPlotLib and Tkinter libraries and are provided as open source code in a GitHub public repository[3] in order to enable validation and reproducibility of the present results.

3. A time domain method: annotation mining

3.1. Duration irregularity as dispersion and distance

In linguistic phonetics, a popular method for examining speech timing is annotation mining, a method which originated in statistical language and speech engineering: the assignment of linguistic labels and time-stamps to segments of speech signals,

[3] Available online at: https://github.com/dafyddg/RFAGUI/.

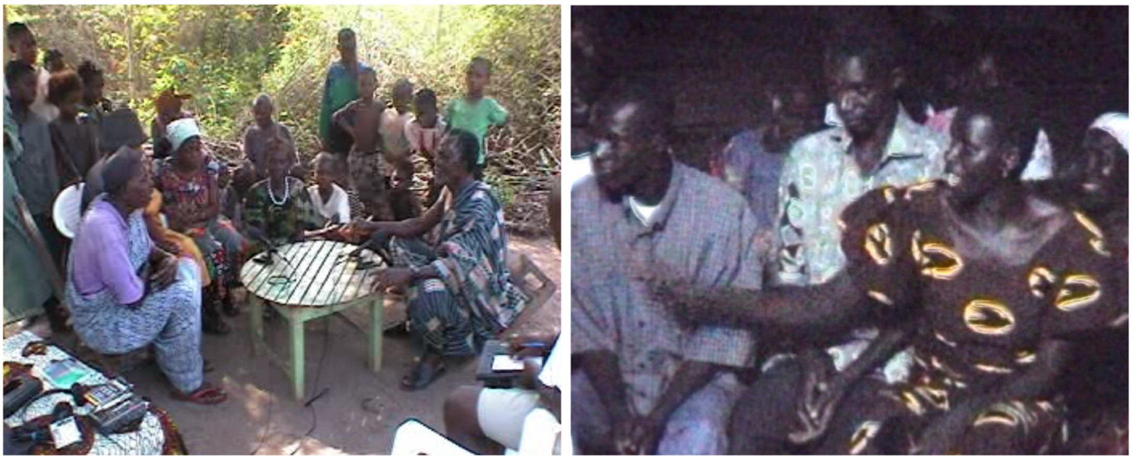

FIGURE 2
(Left panel) Narrator Grogba Gnaoré Marc **(right)** addressing his responder **(left)** and audience (background) during an Ega story-telling session in Gniguédougou village, Côte d'Ivoire, 2001. **(Right panel)** Narrator Kouamé Ama Bié **(right)** selecting her responder **(left)**, with the audience (background), during an Anyi (Anyin, Agni) story-telling evening in Adaou village, Côte d'Ivoire, 2002.

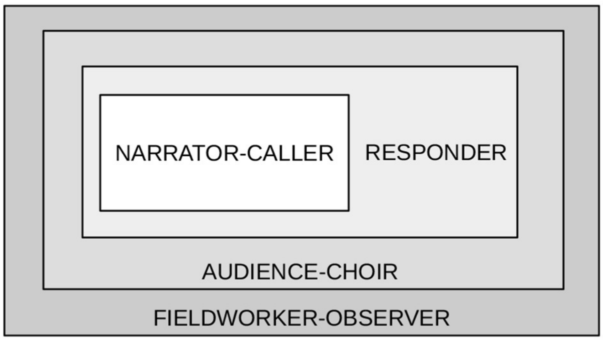

FIGURE 3
Participant role scenario in Kwa interactive narrative orature.

together with the use of irregularity measures to calculate the difference between an assumed ideal of isochrony or equal timing of speech units such as syllables, words or feet and the reality of annotated durations. A number of measures have been used: some are global, based on mean rates or standard deviation for entire utterances, others restrict attention more realistically to the duration differences between neighboring units (Gibbon, 2006; White and Malisz, 2020).

Although often referred to as "rhythm metrics", measures such as standard deviation or covariance are, strictly speaking, not metrics but measures of dispersion.[4] An exception is the Pairwise Variability Index, whose variants (Grabe and Low, 2002) can be derived straightforwardly from the standard metrics Manhattan Distance and Normalized Manhattan Distance.

The more interesting point is that these measures tell only half the story of rhythm, the degree of relative duration irregularity, and have nothing to say about the other half, the alternations and oscillations. The alternations of a rhythm imply (relative) isochrony, but not vice versa, since a series of monotonically increasing durations may show relative isochrony but not rhythmic alternation. Rhythmic duration differences between neighboring syllables alternate between peaks and valleys, e.g., between positive (long-short) and negative (short-long) differences, but the irregularity measures use absolute or squared values, which turn all differences into positive values and thereby not only lose any relation to peak-valley rhythmic alternation but also introduce an ambiguity, a kind of "overkill": the measures do not distinguish between alternating and non-alternating duration sequences [cf. overviews and discussion in Gibbon (2003, 2018, 2021) and Arvaniti (2009)].

A further problem is overgeneralization: by claiming a single irregularity index for a language, the functional variation of rhythm both within utterances and dialogues, and between language varieties and languages, is excluded.

Consequently, in a chain of falsification arguments for rhythm models, the irregularity measures must be regarded as falsified on both formal and empirical grounds, though they remain useful heuristic tools which provide an initial orientation in terms of relative isochrony.

3.2. Plotting syntagmatic turn duration patterns

Annotation mining can also provide useful information beyond irregularity indices, in the form of sequences of duration measurements which show clear syntagmatic duration

[4] Technically, a metric must fulfil the "triangle inequality" condition: for any three points, the sum of the lengths (i.e., distances between the points) of any two sides of the triangle which they constitute must be greater than or equal to the length of the third side. In the case of equality the points are on a straight line.

variation patterns. The recordings of one Anyi narrative and one Ega narrative were annotated at turn level, according to the categories shown in Figures 3, 4, and visualized as *time* × *duration* scatter plots of turn sequencing in Figure 5 (top: Anyi, bottom: Ega).

In the scatter plots the x-axis shows the temporal sequencing of utterance events and the y-axis shows the duration of each event on a logarithmic scale. Log scaling was chosen in order to compress the duration scale and thereby reduce the spread of longer durations along the y axis, permitting easier comparison of the very long narrative utterances with the shorter utterance types. For convenience in comparison, the durations, which are highly variable, are also shown by means of the length and thickness of the utterance event bars. The utterance categories (narrator, responder, call, response) are represented by different colors and listed in the figure legends together with their assigned colors.

Figure 5 demonstrates a number of particularly striking results pertaining to variations in turn duration and therefore tempo:

1. In contrast to many other types of dialogue interaction, the turns do not overlap temporally (this is slightly obscured at some points in the diagram by the use of bar width and length to indicate duration).
2. The Anyi turns have considerably shorter durations than the Ega turns, i.e., a faster turn tempo, perhaps a consequence of the relative informality of the Anyi scenario.
3. The two turn duration plots both show long-term iterative linear patterns of interactively generated long-term discourse rhythms. The discourse cycles are formally related to the linear cyclic prosodic grammars of intonation and rhythm (Pierrehumbert, 1980; Gibbon, 1981, 1987, 2001), and provide a backbone structure for discourse-level prosodic patterns.
4. The Ega duration sequence has a long-term rhythm in the form of periodically accelerating sequences of call-response sequences with rising and falling (i.e., decelerating and accelerating) intervening turn duration patterns. These patterns match the cyclical sequencing of utterance categories, possibly interpretable as a mark of the rhetorically very animated, though formal and authoritative style of the narrator.
5. In the Anyi sequence there is a very long term repetition of call-response sequences, and otherwise shorter rhythms of more diffuse decelerating-accelerating patterns can be seen (for example between about 35 and 90 s, and 110 and 140 s, followed by a call-response sequence). This is possibly another characteristic of the lively narrator and the much more informal style with drums accompanying the chants, much laughter and very much shorter turns.
6. Overall deceleration, in which turn durations become longer as the narrative phases proceed, is evident in both Anyi and Ega.
7. In both Anyi and Ega, the call-response sequences are also visible as conspicuous regular patterns, reflecting the long-term rhythmic expectations predicted by the transition network discourse grammar.

The conclusion from this analysis is that on the one hand interesting rhythmical sequences are shown, in the sense of iterating patterns, but on the other hand it is not yet clear whether these temporal patterns actually have the frequency properties of rhythms.

3.3. Paradigmatic turn clustering with spectral similarity

Anticipating the RFA methodology of the following sections, a cluster analysis of the low frequency (LF) similarities between spectral properties of the turns shown in Figure 6 was calculated. The very short events pause and backchannel are not included.

The calls do not show the same consistency as the response turns or the narrative turns. But the latter form two clear category clusters:

1. The response turns, which are chanted and highly rhythmical, all cluster together, without exception, and are thereby shown to have very similar spectra;
2. Narrative turns also show a tendency toward similar cyclic patterns and most of them cluster together, though they are not chanted, unlike the response turns.

Several distance metrics and clustering criteria were examined for this test. The most plausible result was achieved with the Chebyshev Distance metric (also known as Chessboard Distance or Maximum Value Distance) together with average distance linkage for cluster hierarchy calculation.

The conclusion from this analysis is that there are consistent types of turn event behavior, which would be required for the analysis of rhythms, but it is again not yet clear whether the spectral similarities of turns are in fact rhythmical according to the criteria established in the first section. Since information about rhythmic patterns is crucial for the description and explanation of rhythms and their variation, the turn duration dendrogram model must also be seen as inadequate and, though useful, as partially falsified.

4. A frequency domain method: RFA

4.1. The missing link: modulation and demodulation

Though the time domain methods outlined in the preceding section provide useful insights, there are still missing links when it comes to describing and explaining speech rhythms, and the question arises of how these missing links may be captured. The question also arises of which styles, registers and genres of speaking are more likely to be rhythmical, and thus also isochronous, than others. Intuitively, counting, or rhetorical repetition, poetry reading and chanting are more likely to be more fluent and more rhythmical, while spontaneous speech is likely to be less fluent, less rhythmical and more hesitation-prone.

Rhythm Formant Theory (RFT), a further development of Speech Modulation Theory, together with its Rhythm Formant Analysis methodology, addresses the properties of rhythm which were defined in Section 1: rhythm as oscillation with temporal persistence and frequency resonance. None of these properties are accounted for in isochrony-oriented annotation-mining approaches, though the *time* × *duration* scatter plot approach of Figure 6 provides useful information.

The concept of rhythms as oscillations requires methods for detecting the frequencies of rhythmic modulations of the speech signal as peaks, or rhythm formants, in the speech spectrum. The

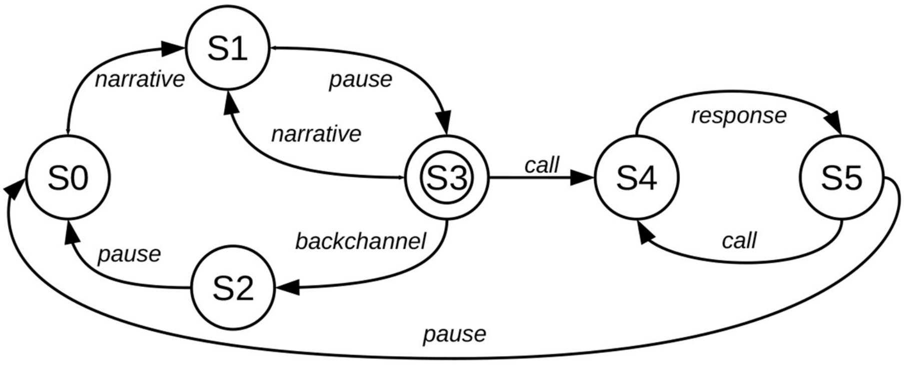

FIGURE 4
Iterative transition network model for interactive spoken narrative.

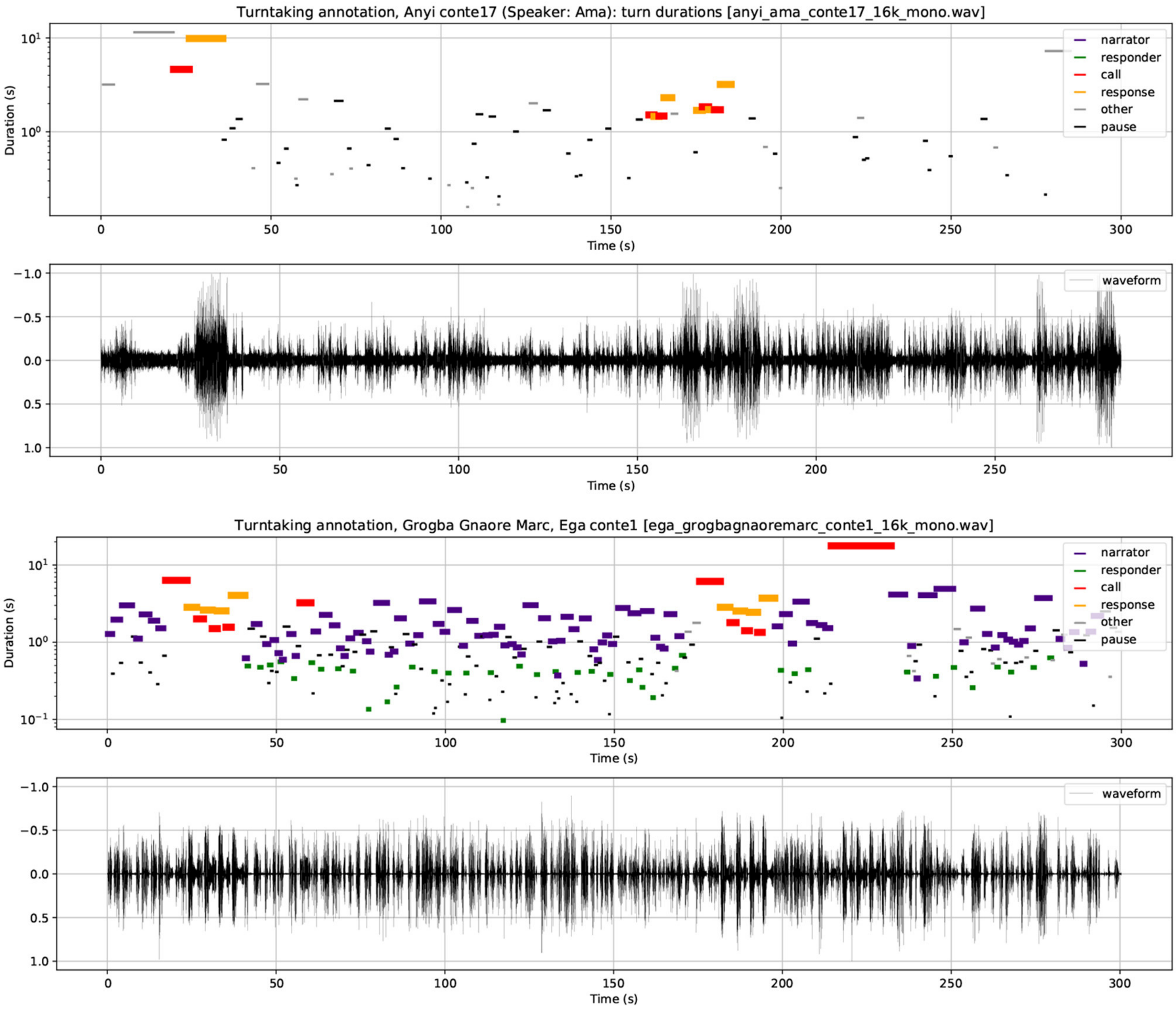

FIGURE 5
(Top panels) Sequential syntagmatic temporal distribution in the Anyi narrative. **(Bottom panels)** Sequential syntagmatic temporal distribution in the Ega narrative.

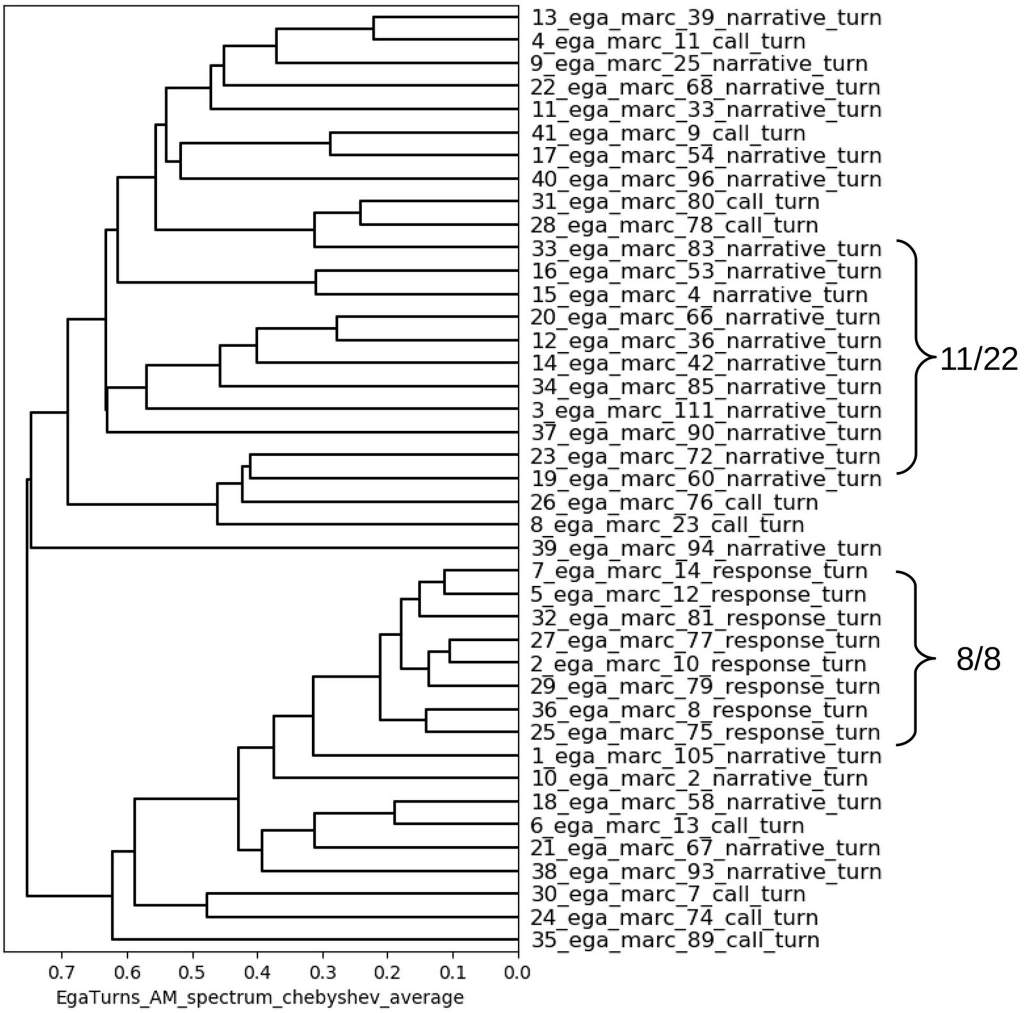

FIGURE 6

Distance dendrogram of turn durations in the Ega fable. Label numbering: 1. file processing sequence (excluding backchannels and pauses); 2. turn sequence in annotation.

appropriate algorithms are essentially the same as the algorithms used for phone formant estimation in acoustic phonetics, but used in a spectral domain below 10 Hz with a much lower frequency range than the range below and above 1 kHz which is relevant for phone formant estimation. The Rhythm Formant Analysis method uses spectral analysis algorithms (Fast Fourier Transform, FFT) to analyze the speech signal, and identifies the frequencies of magnitude peaks in the spectrum as acoustic correlates of rhythms, interpreted as rhythm formants.

The speech signal has three distinct components which are relevant for modulation analysis with RFA (there are interactions between the components, and other modulation types, but these are not immediately relevant here):

1. A carrier signal (F0 with harmonics, generated in the glottis; consonantal noise) which is modulated in two ways by information signals;
2. Frequency modulation (FM) of the carrier signal with a low frequency FM information signal derived from slowly changing tones, pitch accents and intonations, which modulate the frequency (F0) of the carrier signal;
3. Amplitude modulation (AM) of the frequency modulated carrier signal with a low frequency AM information signal pertaining to the syllables, words, phrases, and longer sonority curves, which modulate the amplitude of the carrier signal.

Phonetic analysis (and human perception) of the speech signal amounts to demodulation of the signal, in the exact same technical sense of extraction of information-bearing signals from the modulated carrier signal as in AM and FM radio broadcasting, though with the carrier in very different frequency ranges - around 100 Hz, not, say, 100 MHz:

1. FM demodulation extracts the LF FM information (relating to tones, pitch accents and intonations) and is shown as variations in fundamental frequency (F0, the “pitch track”).
2. AM demodulation extracts the LF AM information (the “sonority curve” of syllable, word, phrase and longer discourse units).

RFA is concerned with the demodulation of the speech signal. Only AM demodulation is considered in the present contribution. The details of rhythms, their frequencies and magnitudes, are obtained by spectral analysis with a Fast Fourier Transform (FFT) applied to long segments of the signal. The resulting spectrum is examined for spectral peaks which are compared with spectral peaks from long segments in other signals from the same speaker or from other speakers, genres or languages.

In a full analysis, for example in traditional forms of speech recognition, phones and phone sequences are also demodulated, but detailed phone demodulation is not of immediate interest. FM tones, pitch accents and intonations, i.e., modulations of F0, also contribute to impressions of rhythm, but are not further considered here (Gibbon, 2021).

4.2. A "clear case" of a rhythmic register: counting

Figure 7 illustrates several features of the RFA method with an acoustic phonetic visualization of counting from one to ten in Ibibio: *kèèd, ìbà, ìtá, ìnààñ, ìtíòn, ìtiòkèèd, ìtíàbà, ìtiàità, ùsúkkèèd, dùòp*. The top panel of Figure 7 shows the waveform and, superimposed, the demodulated LF AM "sonority" envelope. The middle panel shows the LF FM modulation of the male voice with modulations of the carrier wave ranging from 110 to 165 Hz. The bottom panel shows the LF AM spectra of the first and the second halves of the signal.

The morphology of numeral systems, together with the formality of the register of counting, can be expected to provide fairly precise information about basic word rhythms. In English (Gibbon, 2021) the largely simplex monosyllabic numerals below 10 contrast with the disyllabic teens and the trisyllabic twens of the system.

Figure 7 shows an RFA visualization of the base five numeral morphology of Ibibio, in which the morphologically simple numbers from 1 to 5 have a different morphophonological structure from the morphological compounds which constitute the numbers from 6 to 9 (5+1, 5+2, 5+3, 5+4), followed by a new morpheme for 10. The structure of the numerals from 1 to 5 is either monosyllabic (1) or disyllabic (2–5), while the structure of the numerals from 6 to 10 is either monosyllabic (10), trisyllabic (6, 7, 9) or quadrisyllabic (8). Consequently, different rhythmic effects in the two halves are to be expected.

Inspection of the waveform and the AM envelope shows that there are 10 words in 8.865 s, an average interval of about 0.887 s per word, corresponding to a word rate of ≈1.127 words per second. The word rate of 1.127 and the different morphophonology of the sequences 1–5 and 6–9 leads to two predictions. If the words are approximately equally timed, then:

1. There will be a measurable frequency of about 1.127 Hz in the LF spectrum, corresponding to the word rate, and
2. The more complex morphophonology in the second half will necessarily be associated with the faster rhythms of a higher number of component syllables than the simpler morphophonology in the first half.

To test these predictions the signal was divided into two halves: segment A, corresponding to the numerals 1–5, and segment B, corresponding to the numerals 6–10. Separate spectral analyses were made of these two halves. The analyses are shown as separate spectral slices in the bottom panel of Figure 7. Spectral magnitude values rather than energy values are used, with the values rescaled to 0, ..., 1 in order to permit comparison of the spectral shapes rather than of the absolute magnitudes (this also applies to the analyses in the following sections). The results are as follows:

1. The two spectral slices both share a high magnitude rhythm formant at 1.128 Hz, very close to the 1.127 Hz predicted by the rule-of-thumb word count based on the waveform and the AM envelope.
2. In the higher frequency ranges the spectral slices of the two segments differ, with much more diffuse frequencies for Segment B, by virtue of the greater range of morphophonological structures.

Accounting for the individual spectral magnitudes leads too far from the main topic of the present study and requires provision of spectral analyses of far more spectrum parts. However, the results of the counting example already show that RFT can provide a principled modulation-theoretic, empirically well-grounded causal explanation for different speech rhythms in the physical acoustic domain.

4.3. RFA in Speech Modulation Theory

The explanatory theoretical background to Rhythm Formant Theory lies in Speech Modulation Theory. The terminology "rhythm formant" is used because both LF rhythm formants (below 10 Hz) and HF (high frequency) phone formants (between about 300 Hz for F_1 and 2,500 Hz for F_2 and F_3) share the same acoustic definition: magnitude peaks in the spectrum.[5] A more detailed account of Rhythm Formant Theory and its associated Rhythm Formant Analysis methodology is given in Gibbon (2021) with an application to rhetorical speech in Gibbon and Li (2019) and to second language (L2) fluency evaluation (Lin and Gibbon, 2023).

The history of modulation theoretic analyses of rhythm is long but sparse, and ranges from linguistic phonetic and phonological approaches (Ohala, 1992; Dziubalska-Kołaczyk, 2002) to signal theoretic approaches (Todd and Brown, 1994; Traunmüller, 1994; Cummins and Port, 1998; Barbosa, 2002; Tilsen and Johnson, 2008; Tilsen and Arvaniti, 2013; Gibbon, 2021), among others.

[5] The term "rhythm formant" was suggested by Huangmei Liu, Shanghai, and replaces the older term "rhythm zone". The articulatory and perceptual definitions of "formant" differ: in articulatory terms a formant is a resonant filter frequency of the vocal tract, and in perceptual terms a formant is a filter frequency which contributes to a particular sound quality (as between the vowels [i] and [a]). Also, F0, the fundamental frequency, which functions as a carrier signal for FM and AM modulation, can be seen as a formant from the acoustic point of view (which is why it is often labelled "F_0" rather than "f_0' to match the phone formants F_1, F_2, F_3), but not from an articulatory or perceptual point of view as it is a property of a carrier signal source not of a modulator.

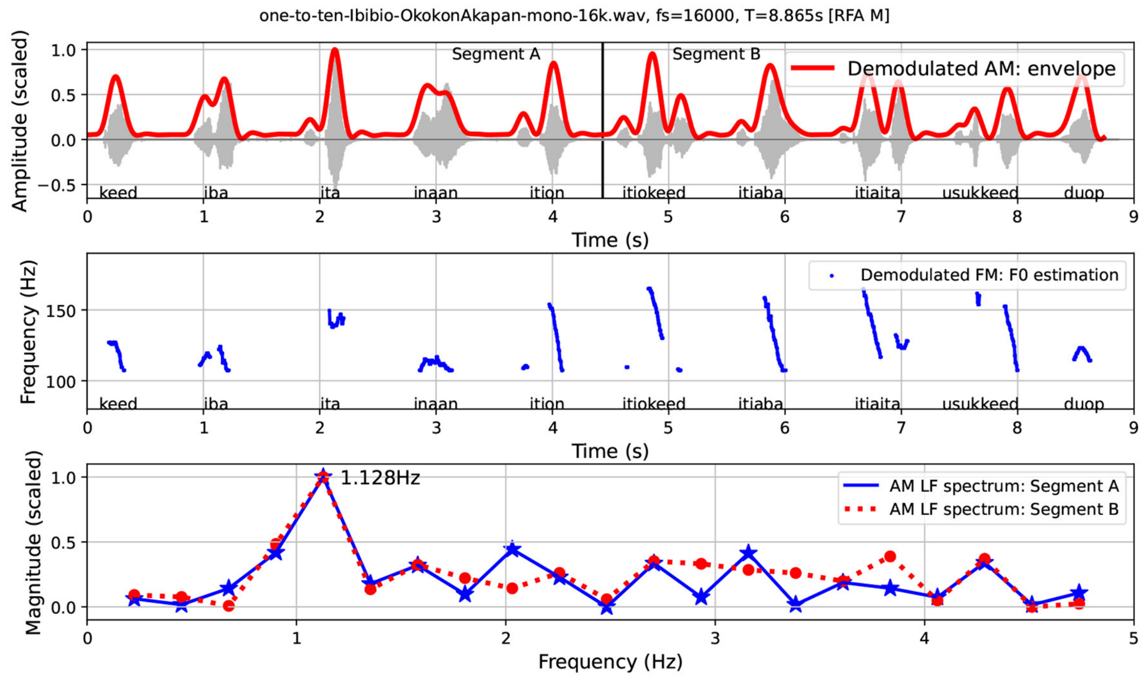

FIGURE 7
Counting from 1 to 10 in Ibibio: waveform, LF AM envelope, LF FM (F0) estimation and LF AM spectrum.

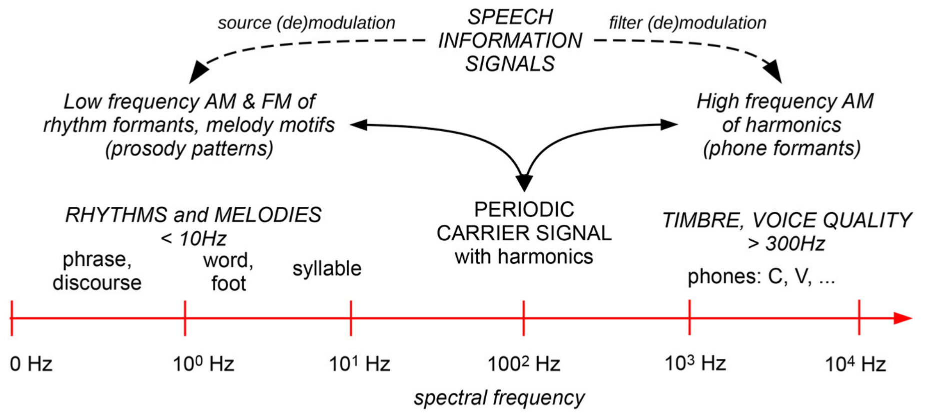
source (de)modulation
SPEECH INFORMATION SIGNALS
filter (de)modulation
Low frequency AM & FM of rhythm formants, melody motifs (prosody patterns)
High frequency AM of harmonics (phone formants)
RHYTHMS and MELODIES < 10Hz
phrase, discourse
word, foot
syllable
PERIODIC CARRIER SIGNAL with harmonics
TIMBRE, VOICE QUALITY > 300Hz
phones: C, V, ...
0 Hz
10⁰ Hz
10¹ Hz
100² Hz
10³ Hz
10⁴ Hz
spectral frequency

FIGURE 8
Speech modulation frequency scale.

The *Speech Modulation Frequency Scale* in Figure 8 shows the domain of RFT at the LF end of the three three main frequency ranges: (1) LF AM modulation by speech rhythms, (2) the carrier signal at mid-frequencies, with LF FM modulation attributable to tones, pitch accents and intonations, and (3) LF AM modulation of HF harmonics by filtering in the oral and nasal tract, attributable to phonemes and their allophones. The three frequency bands for the three components of the speech signal show an interesting exponential pattern: from order of magnitude 10^0 Hz (1 Hz) for word rhythms and 10^1 Hz (10 Hz) for syllable rhythms, pertaining to both LF AM and LF FM, through order of magnitude 10^2 Hz (100 Hz) for the FM modulated carrier frequency (F0), to order of magnitude 10^3 Hz (1,000 HZ) for the LF phone formant modulations of the carrier signal harmonics.

The counting register described in the preceding subsection is a clear case in which rhythm formants can be clearly identified, implying that there are other registers which vary in rhythmicality, as the analysis of interactive orature in following section shows. RFA captures the frequency, persistence and resonance properties of rhythm, the missing links in the annotation-mining approaches. It is predicted that these properties can be used to compare rhythms of different language varieties and registers. This prediction is tested in the following section.

5. Quantitative results: data comparison

5.1. Rhythm vector extraction

The procedures described in the previous section provide the foundation for extracting spectral properties in the form of vectors for comparing languages and language varieties. There are many options for creating vectors, such as the following non-exhaustive list:

1. From the LF spectrum:
   - The entire LF spectrum,
   - The vector of the *n* highest magnitude frequencies or their magnitudes,
   - The *m* highest magnitude spectral peaks or their magnitudes,
   - Magnitudes and frequencies above a certain magnitude level *l*,
2. From the spectral slices in the LF spectrogram:
   - The trajectory through all the spectral slices of the highest magnitude frequencies or their frequencies in each slice,
   - A matrix of trajectories of other choices under the LF spectrum options.

The values of *n*, *m* and *l* in each case are chosen empirically on the basis of their value for interpreting comparisons between language varieties. For the present discussion, the 10 highest magnitude spectral frequencies were chosen ("lfammaxfreqs" in the figure legends). Extraction of the highest magnitude frequencies may result in ignoring lower magnitude peaks, but captures the lower but still high magnitude levels surrounding the highest peaks. These frequencies are important for characterizing the overall shape if the bandwidths of formants associated with the highest peaks are relatively broad, or the peak frequencies fluctuate slightly.

Using the selected vectors, standard distance metrics are used to generate a triangular distance table showing differences (and, by implication, similarities) between vectors. The distance table is created using a standard distance metric, Manhattan Distance (also known as Taxicab Distance, Cityblock Distance and Mannheim Distance), the sum of absolute differences between corresponding positions in the vectors. The intuition behind Manhattan Distance is that a complex terrain cannot always be traversed in a straight line (as with Euclidean Distance), but only via possibly right-angled corners, as a taxi-driver in Manhattan or Mannheim would do. Intuitively, the rhythm formant structure in the LF spectrum appears to be compatible with this assumption. There are many other distance metrics; in similar experiments, Canberra Distance, Euclidean Distance and Cosine Distance were also tried, but turned out to be less useful.

The spectral resolution of the FFT analysis depends on the window interval to which it is applied. Since the recordings differ in length, and the FFT window is the entire recording, the vectors for the different recordings differ in frequency resolution. As the distance metrics require identical lengths, all vectors are dynamically extrapolated to the length of the longest vector in the set.

5.2. Distance networks

The pairwise distances between the length-normalized vectors are rescaled to the range (0,...,1), and the triangular distance matrix is calculated and visualized as a distance network (Figure 9). If all distances are shown, the network is fully connected and the figure does not immediately reveal interesting visual patterns, so the network in Figure 9 uses a maximum distance threshold, in this case = 0.51. The example shows distances between recordings based on the 10 highest magnitude values in the LF spectrum below 10 Hz.

The most conspicuous regions of the network are the following:

1. The cluster Bete, Ibibio and CdIFr recordings of reading aloud (top right), with distances < 0.45;
2. The cluster with all but 2 of the Anyi and Ega extempore recordings (large leftmost cluster), with distances < 0.42;
3. The Ega conte1 node linking these two regions, but closer to the reading group than to the rest of the extempore group, with distances < 0.47 and < 0.487, respectively;
4. The Anyi conte2 node is somewhat isolated, with a distance of 0.485 to its nearest node, which is actually in the reading-aloud cluster, the next nearest being the Ega conte1, the borderline linking node between the readings and the other extempore narratives; the style of Anyi conte2 is interpreted as being more formal than the style of the other Anyi presentation, bringing it closer to the Ega presentations.

It is striking that the first cluster in this list contains precisely the register of read-aloud non-interactive narratives and the second is the register of extempore stories. The similarity between the Ega conte1 recording and the read-aloud recordings may be because it was the first to be recorded in the Ega recording sessions, while the other Ega recordings became less formal as the sessions continued. The speaking style in the Anyi conte17 recording is informal, but the earlier Anyi conte2 recording is more formal, as already noted, possibly as it is also earlier in the recording series.

5.3. Distance hierarchies

A more complete overall picture of relationships between the recordings is given by hierarchical clustering and can be rendered as a dendrogram (Figure 10). The clustering is calculated according to the same criterion as the distance network, Manhattan Distance, with the addition of the complete (farthest neighbor) distance linkage clustering method, a robust criterion which takes the farthest distance between any two elements in the clusters in order to fuse the clusters into a superordinate cluster. The linkage criterion thus adds an additional layer of inductive generalization to the analysis.

The hierarchical clustering dendrogram shows the relations between reading and non-reading groups in more detail than the

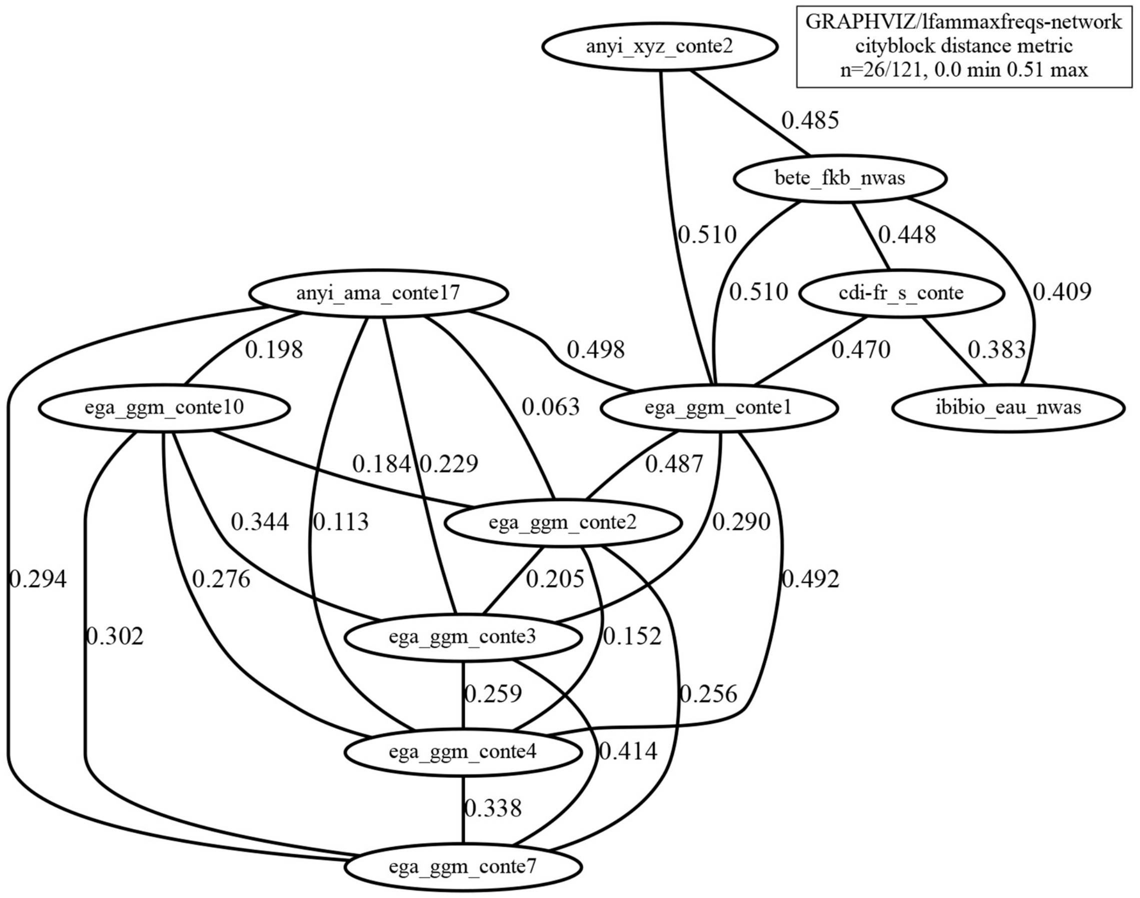

FIGURE 9
Distance network based on the vector of the 10 highest magnitude frequencies in the spectrum, relating recordings with pairwise distances < 0.51 (max. distance normalizlknow ed to 1.0).

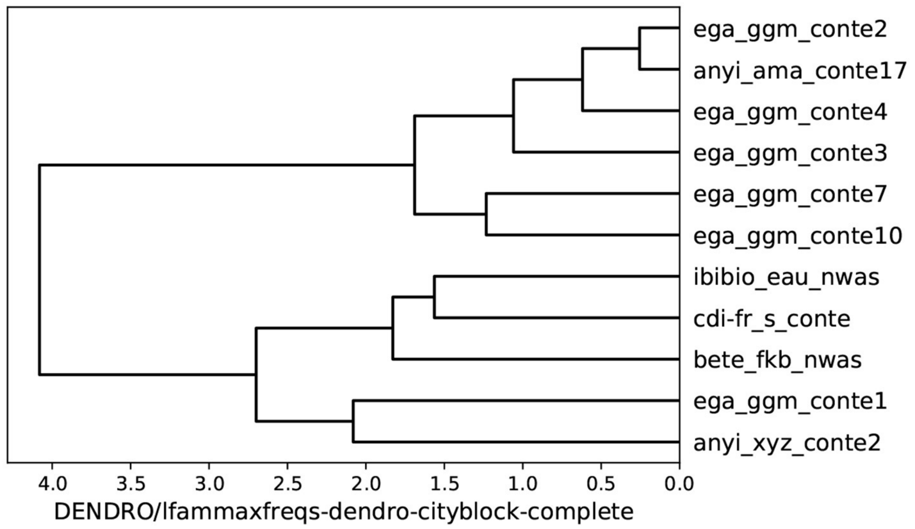

FIGURE 10
Hierarchical clustering of similarity relations between the recordings, with Manhattan Distance and farthest neighbor linkage, based on the vector of the 10 highest peaks in the spectrum.

distance networks. The read-aloud recordings (*ibibio_eau_nwas*, *cdi-fr_s_conte*, *bete_fkb_nwas*) cluster together, and at the next hierarchical cluster level they are joined by the two more formal Ega and Anyi cases (*ega_ggm_conte1*, *anyi_xyz_conte2*). The more informal *anyi_ama_conte17* clusters closely with the remaining 5 Ega recordings. As expected, the relations in Figure 9 are reflected in Figure 10. The pairwise distance comparison thus yields the predicted result: the read-aloud stories are distinguished from the interactive oral narratives, as predicted.

Automatically arranging languages or styles in a network or dendrogram is not the end of the story, however, though it is already very useful. The phonetician or linguist clamors for an explanation of why this classification works, not only a representation or a description. This need is addressed in the following section, first by means of holistic interpretation of long-term spectra of the data and then by means of holistic interpretation of spectrograms.

6. Interpretative results: a holistic perspective

6.1. Frequency: the LF spectrum

The quantitative analyses in the preceding section provide useful insights but they need explicit interpretation, as algorithmic correctness or statistical significance alone do not imply relevance or importance. Two interpretation steps are taken in this section. The first step is to examine and understand the spectral shapes which are extracted from the AM envelopes by applying an FFT to the entire event of story-telling or reading, and were illustrated in Figure 7. The holistic analysis is applied whether the event is a few seconds long, as in the Ibibio counting example, one minute, as with the read-aloud stories, or in the case of the extempore stories, five minutes or more.

Further quantitative analysis of spectral shapes are a legitimate but different issue. Holistic interpretation of the holistic LF spectra is not concerned with quantitative details of individual spectra but with the "gestalt" of the spectrum, i.e., the visual shape. Consequently, the individual subfigures in Figure 11 are deliberately kept together in a very compact format in order to encourage holistic interpretation. The sources of the spectra are identified in the caption. Ideally, the holistic descriptions would be performed by an independent panel of judges but this is not feasible within the current research environment.

Interpretation:

1. Ega (rows 1 and 2): The lowest salient peaks cover approximately the same range, 0.1 to 0.5 Hz, indicating discourse rhythms which correspond to intervals between 10 and 2 s. The first Ega story (row 1: 1) has a broad formant over this range while the other Ega stories have similar patterns of two or three peaks. These stories are narrated by the same male speaker.
2. Anyi (row 3): The two stories have quite different spectral patterns, and are narrated by different female speakers. The first story is narrated in a very restrained and somewhat monotonous style and the LF spectrum is very much unlike the Ega LF spectra; it has a broad salient region with a peak at about 0.6 s. The second story is narrated in a lively style, with drums and chants, and the spectrum is similar to the Ega stories in having a salient peak at about 0.1 Hz, corresponding to intervals of 10 s.
3. North Wind and Sun readings (row 4: 1 and 2): The Bete (row 4: 1) and Ibibio (row 4: 2) readings of *The North Wind and the Sun* are in a fairly neutral formal reading style and have very similarly shaped spectra, with a slope up to a peak around 0.5 s (Bete) and 0.6 s (Ibibio). The shapes of the Bete and Ibibio LF spectra and the spectrum of the first Anyi story are very similar, perhaps indicating a shared feature of formality. The durations of the Bete and Ibibio spectra are very similar, while the first Anyi story is almost twice as long as these two, but has less than half the durations of the other Anyi story and the Ega stories.
4. CdI French reading (row 4: 3): The CdIFr LF spectrum has a similar overall shape to the Bete and Ibibio readings and the first Anyi story, again presumably indicating a feature of formality which may be associated with reading aloud, and also a feature of less lively oral narrations.

Although the long-term spectra show distinctive patterns which are explainable in terms of categories and structures of a discourse grammar, the spectrum has formal and empirical limitations: the spectrum contains no temporal information about how long the rhythms last, and therefore cannot capture the persistence and resonance criteria for rhythms. In fact, even one or two periods, which would not actually count as a rhythm, will be registered in the spectrum.

Consequently, a spectrum only permits formulating tentative hypotheses about rhythms, since it is an abstraction from the essential temporal information required for a full characterization of rhythm. The following section is concerned with providing the missing temporal dimension by means of interpretation of LF spectrograms.

6.2. Time: the LF spectrogram

LF spectrograms are calculated for each of the data items in order to capture temporal properties of rhythms. Each spectrogram consists of a sequence of spectral slices, i.e., spectra which are calculated from FFT analyses with shorter windows. However, "short' is relative: in order to capture the very low frequencies involved, the windows for the LF formants have to be very much longer than the 5 ms or so which are used to capture HF phone formants. In order to cover the low frequencies of the rhythm formants, the window interval is set at 8 s.

The 8 s moving window means high frequency resolution but low temporal resolution, a practical case of Heisenberg's "principle of uncertainty". The lost temporal resolution is regained by overlapping the windows almost completely and moving forward in very short strides of the order of 50 ms. In order to ensure that the final 8 s are not ignored in the final window step, the recordings are lengthened by 8 s of blank audio (which creates small artifacts at the end of the spectrograms).

The spectrograms are rendered in traditional heatmap format with vertical spectral slices in a 3-dimensional representation

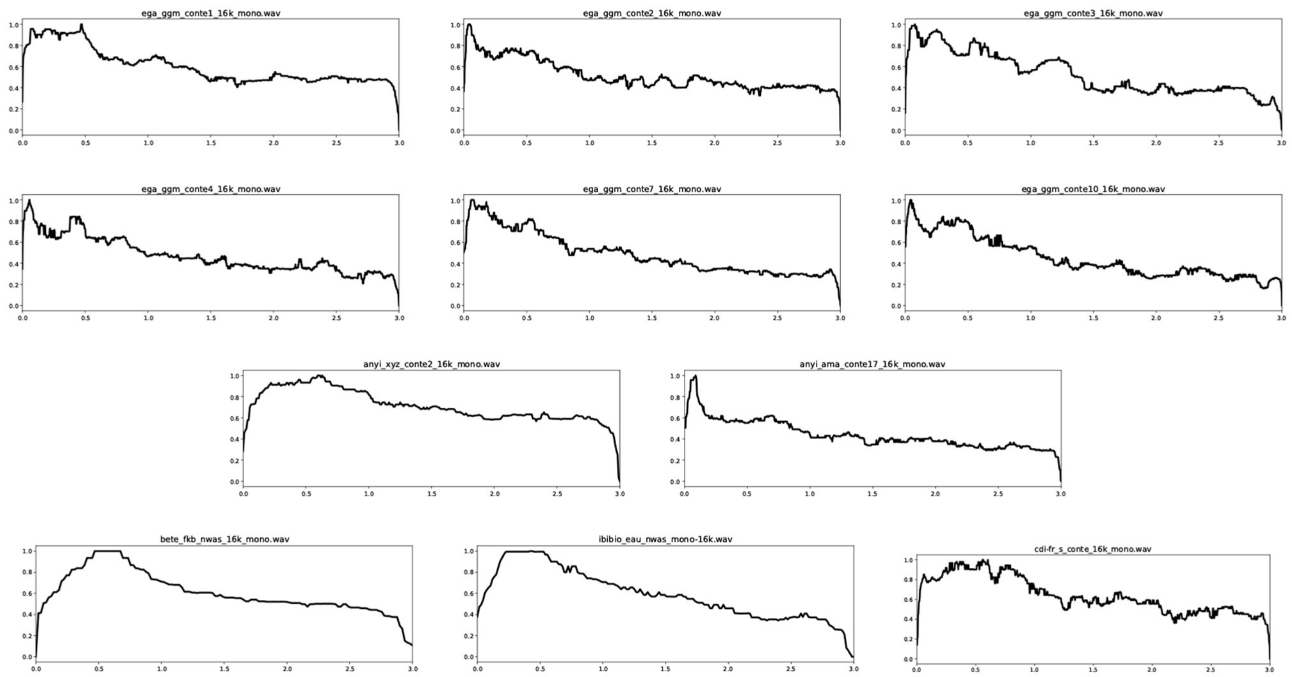

FIGURE 11
LF spectra of the data examples, smoothed with a moving median window and compactly formatted for holistic interpretation: Row 1: Ega conte1, Ega conte2, Ega conte3. Row 2: Ega conte4, Ega conte7, Ega conte10. Row 3: Anyi conte2, Anyi conte17. Row 4: Bete; Ibibio; CdIFr. (readings).

time × *frequency* × *magnitude*, with *time* on the *x*-axis, *frequency* on the *y*-axis, and *magnitude* as color or gray shading (cf. Figure 12) with darkest meaning strongest. Rhythmic sequences at a given frequency appear as dark horizontal bars and provide evidence for the persistence and resonance criteria for rhythm.

The spectrograms can be interpreted as follows:

1. Ega (rows 1 and 2): Pairs of dark horizontal bars are evident in the first Ega story (row 1: 1), at two frequencies, and correspond to the rhythmical chanting periods, which appear in the waveform oscillogram as regularly spaced vertical bars. The third story (row 1: 3) is somewhat similar, and the second story (row 1: 2) also contains a highly rhythmical section close to the beginning, with similar horizontal bars. In the other stories horizontal bars can also be oberved, though they are shorter and more scattered.
2. Anyi (row 3): The two Anyi spectra are quite different from each other, partly due to the much shorter duration and greater formality of the first story, where the horizontal bars are very prominent. In the second story, the horizontal bars are more scattered, as in the majority of the Ega stories.
3. Bete, Ibibio (row 4): The Bete (row 4: 1) and Ibibio (row 4: 2) readings have approximately the same duration, and the horizontal bars are very prominent, as expected in a more formal style (both native speakers are university professors). The very clear multiple rhythms at different frequencies in the Ibibio example are particularly interesting, indicating temporally regular enunciation of smaller speech units such as words and syllables.
4. CdIFr: The Ivorian French reading is more similar to the Ega stories and the second Anyi story, but with a greater concentration of lower frequencies in the spectrogram.

The detailed temporal information in the spectrograms supports the results of the holistic spectral analysis. In effect, they tease apart the formant peaks in the LF spectrum and assign them to different times. The horizontal bars are most salient in the readings which had already been interpreted as more formal, in particular in the shorter Bete and Ibibio readings. This might be expected from a reading aloud speech activity whose rhythmicity might be predicted to be quite high. The fact that these recordings have durations of around 1 min, rather than the 2–9 min of extempore narrations enhances the visual impression of the relatively long resonant phases in the spectrograms.

The many details which have not been discussed call for further investigation of the temporal patterning of small speech units within the story, in addition to the clear rhythmical patterns; this is beyond the scope of the present study, however.

In conclusion, the issue mentioned at the end of the previous section, that spectral analysis can register even a single period, is now resolved, since temporal information is available in the spectrogram. The interpretation of horizontal bars in the spectrogram is closely related to the principle that a rhythm must have at least three beats, expressed as the principles of persistence (a minimum rhythm sequence length) and resonance (persistence of the same frequency).

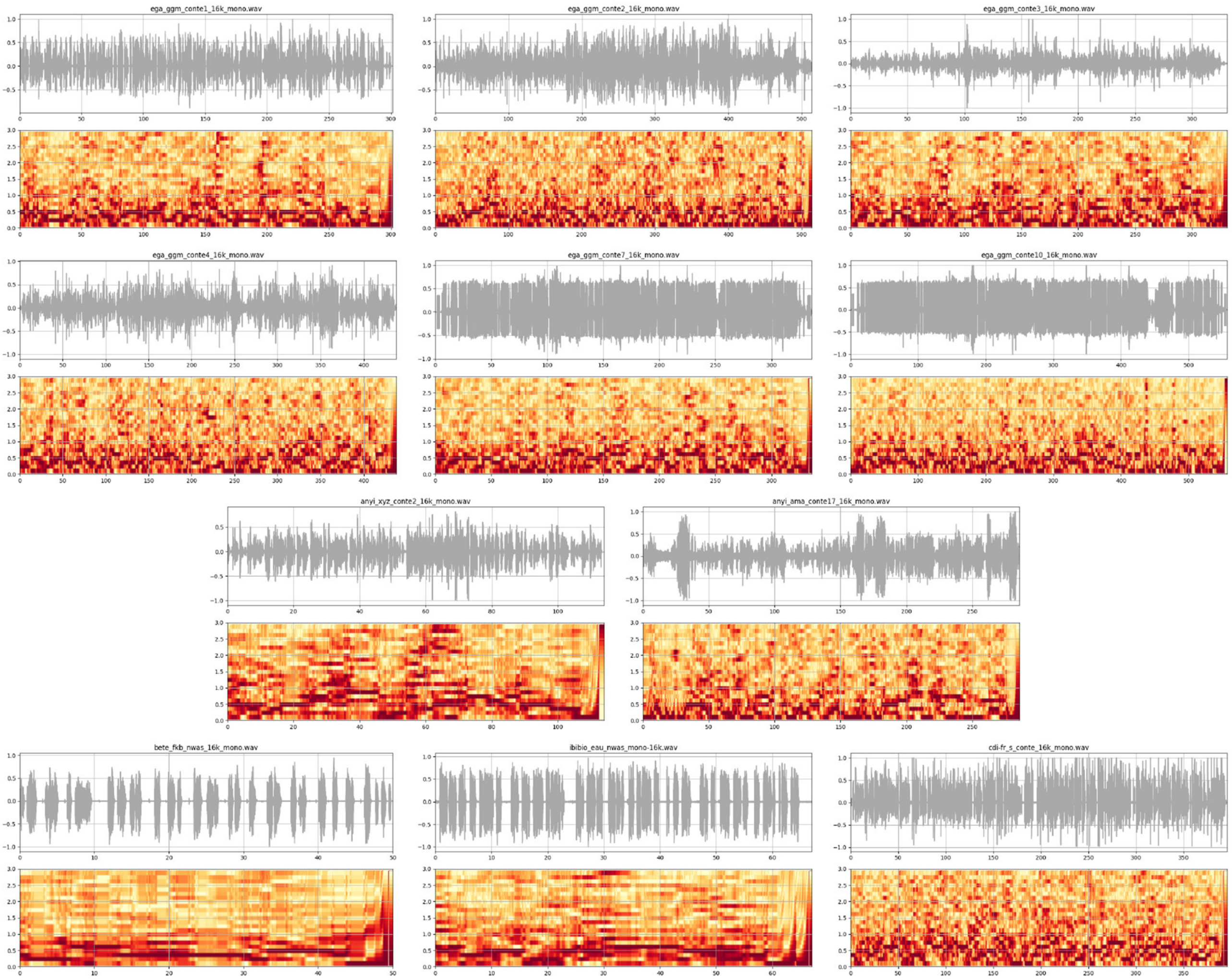

FIGURE 12
LF spectrograms of the data examples, filtered to emphasize the higher frequencies, and compactly formatted for holistic interpretation and comparison: Row 1: Ega conte1, Ega conte2, Ega conte3. Row 2: Ega conte4, Ega conte7, Ega conte10. Row 3: Anyi conte2, Anyi conte17. Row 4: Bete; Ibibio; CdIFr. (readings).

7. Summary, conclusions and outlook

The present transdisciplinary study explores a new methodology which combines functional, structural and causal explanations, and models rhythms and rhythm variation by means of acoustic phonetic analysis. The method is applied to interactive narratives in two West African languages, Anyi and Ega, in comparison with reading aloud in other West African languages.

The languages concerned were introduced, together with a brief account of the background to the narration events, the content of the narrative and the roles of the participants, and the dynamic narrative development pattern was summarized formally in a transition network. An initial discussion of time domain patterns in the narratives was conducted using a tier of turn annotations of two of the narratives, one each in Anyi and Ega, with interpretation of rhythms using visualization of a discourse turn duration pattern, and by relating a hierarchical classification of turn durations to turn types.

Rhythm Formant Theory (RFA) and its associated methodology of Rhythm Formant Analysis (RFA) were introduced and applied initially to rhythm formants in the low frequency spectrum and to the clustering of data samples. In order to provide time domain information on temporal rhythm variation, low frequency spectrograms were introduced for visual interpretation.

At the outset, the following questions were raised and answered in the course of the discussion:

1. Which overall utterance duration patterns are observed in different turn types (narrator, responder, audience) during the narrative? Answer: concurrent rhythms at different frequencies are empirically established using both annotation mining and RFA.
2. How rhythmical are these patterns? Answer: In order to complete the definition of rhythm as oscillation, additional criteria of persistence in time and resonance (persistence in frequency) were introduced.

3. Are there functional interpretations of the differences between rhythm patterns within orature sessions or between sessions in different languages? Answer: Functional interpretations of rhythm patterns were related to turn interaction types, speech registers and social structures, for example with the *chef du village* as narrator, compared with a popular and more informal narrator without this status, to different narration styles and to the genre difference between extempore interactive narration and more formal read-aloud narratives. Rhythm variation was initially demonstrated using the register of counting aloud.

The claim is that the Rhythm Formant Theory and Rhythm Formant Analysis approach is empirically better grounded the rhythm index approaches and also than approaches which rely only on the analysis of the LF spectrum rather than the LF spectrogram. It was demonstrated that the definition of rhythm provided from the start has inherent explanatory value in a number of ways. First, RFT/RFA provides criteria for distinguishing between more and less adequate descriptions, which was shown in the falsification chain of falsified and improved models. Second, RFT/RFA relates easily to functional explanation based on qualitative discourse analysis criteria. Third, RFT/RFA provides a structural explanation in the form of a formal explication for rhythm patterns in the form of a cyclical transition network. Fourth, RFT/RFA provides a causal theory, analysis method and interpretability criteria for realtime rhythm and its relations with the forms and structures of language.

Data availability statement

The datasets presented in this article are not readily available because the dataset consists of audio tracks of fieldwork recordings and the participants are identifiable. Therefore, the recordings are not publicly available but can be used provided that a request is made to the author. The two photos used in the article are by permission of the subjects. Requests to access the datasets should be directed to DG.

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 data was collected as part of fieldwork projects carried out in 2001 and 2002, with approval from the University of Abidjan and the relevant Président de la Préfecture and Chef du Village. Written informed consent for participation and publication of identifiable data/images was not required for this study in accordance with the national legislation and institutional requirements. The participants' oral consent for both participation and the publication of identifiable data/images was documented in the digital recordings. The procedure was approved at the time by the funding agencies VW Foundation and German Academic Exchange Service.

Author contributions

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

Acknowledgments

Special thanks are due to Kouamé Ama Bié and the late Grogba Gnaoré Marc, the narrators of the Anyi and Ega fables; to the villagers of Adaou and Gniguédougou, Côte d'Ivoire; to Firmin Ahoua as co-PI and to Sandrine Adouakou, as Agni researcher; to Eno-Abasi Urua, Moses Ekpenyong, and Mfon Udoinyang for comprehensive support with Ibibio; to Sophie Salffner and Nadine Grimm, both as student co-fieldworkers and, together with Karin Looks and Alexandra Thies, as language documentation experts and annotators. I am very grateful to the late Nicla Rossini as co-author of a study on gesture in Ega orature; to Jolanta Bachan for discussions on finite state dialogue modeling; to the late Wiktor Jassem and the late Grzegorz Dogil for discussions of spectral analysis and segmentation; to Jue Yu and Huangmei Liu, Shanghai, and to Xuewei Lin and Peng Li, Guangzhou, for their support in developing RFT and RFA. Financial support from the DAAD, the VW Foundation and the DFG for fieldwork funding is also acknowledged. I am especially grateful to the reviewers and to Rosemarie Tracy for many insightful queries and comments, and for their insistence on paying more attention to reader uptake.

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.

References

Ahoua, F., and Adouakou, S. (2009). *Parlons agni indénié: Côte d'Ivoire*. Paris: L'Harmattan.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica* 66, 46–63. doi: 10.1159/000208930

Barbosa, P. (2002). "Explaining cross-linguistic rhythmic variability via a coupled-oscillator model for rhythm production," in *Proceedings of the 1st International Conference on Speech Prosody*, eds B. Bel, and I. Marlien (Aix-en-Provence: Laboratoire Parole et Langage), 163–166.

Berry, J., and Spears, R. (1991). *West African Folktales*. Evanston, IL: Northwestern University Press.

Blench, R. (2015). *The Ega language of Côte d'Ivoire: Etymologies and Implications for Classification*. Available online at: https://www.academia.edu/33800011/The_Ega_language_of_Cote_dIvoire_how_can_it_be_classified (accessed October 12, 2022).

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot Int.* 5, 341–345.

Bole-Richard, R. (1983). "Ega," in *Atlas Linguistique des Langues Kwa*, ed G. Hérault (Abidjan: Institut de Linguistique Appliquée, Agence de Coopération Culturelle et Technique), 359–401.

Connell, B., Ahoua, F., and Gibbon, D. (2002). Illustrations of the IPA: Ega. *J. Int. Phonet. Assoc.* 32, 99–104. doi: 10.1017/S002510030200018X

Cummins, F., and Port, R. (1998). Rhythmic constraints on stress timing in English. *J. Phonet.* 26, 145–171.

Dilley, L. C. (2005). *The Phonetics and Phonology of Tonal Systems* [thesis (Ph.D.)]. Cambridge, MA: Massachusetts Institute of Technology, Dept. of Linguistics and Philosophy.

Dziubalska-Kołaczyk, K. (2002). *Beats-and-Binding Phonology*. Berne: Peter Lang.

Ekpenyong, M. E., and Udoh, I. I. (eds.). (2022). *Current Issues in Descriptive Linguistics and Digital Humanities. A Festschrift in Honor of Professor Eno-Abasi Essien Urua*. Singapore; Springer Nature.

Gibbon, D. (1981). "A new look at intonation syntax and semantics," in *New Linguistic Impulses in Foreign Language Teaching*, eds A. James, and P. Westney (Tbingen: Narr), 171–198.

Gibbon, D. (1987). "Finite state processing of tone systems," in *Proceedings of the Third Conference of the European Association for Computational Linguistics (EACL)* (Copenhagen: European Association for Computational Linguistics), 291–297.

Gibbon, D. (2001). "Finite state prosodic analysis of African corpus resources?," in *EUROSPEECH 2001, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH* (Aalborg), 83–86.

Gibbon, D. (2003). "Computational modelling of rhythm as alternation, iteration and hierarchy?," in *Proceedings of the International Congress of Phonetic Sciences, volume III* (Paris: Barcelona), 2489–2492.

Gibbon, D. (2006). "Time types and time trees: Prosodic mining and alignment of temporally annotated data," in *Methods in Empirical Prosody Research*, eds S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, et al. (Berlin: Walter de Gruyter), 281–209.

Gibbon, D. (2014). *Visualisation of Distances in Language Quality Spaces: DistGraph, an Online Teaching Tool for Language Typology Data Mining*. Available online at: https://wwwhomes.uni-bielefeld.de/gibbon/DistGraph/.

Gibbon, D. (2016). "Legacy language atlas data mining: mapping Kru languages," in *Proceedings of the Language Resources and Evaluation Conference (LREC)*. ELRA/ELDA.

Gibbon, D. (2018). "The future of prosody: it's about time," in *Proceedings of the 9th International Conference on Speech Prosody* (Poznañ: SProSIG), 1–9.

Gibbon, D. (2021). The rhythms of rhythm. *J. Int. Phonet. Assoc.* 53, 233–265. doi: 10.1017/S0025100321000086

Gibbon, D., Bow, C., Bird, S., and Hughes, B. (2004). "Securing interpretability: the case of Ega language documentation," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)* (Paris: ELRA), 1369–1372.

Gibbon, D., Gut, U., Adouakou, S., and Urua, E.-A. (2002). "Rhythm in West African tone languges: a study of Ibibio, Anyi and Ega," in *Typology of African Prosodic Systems, volume 1 of* Bielefeld Occasional Papers in Typology (Bielefeld: Bielefeld University), 159–165.

Gibbon, D., and Li, P. (2019). "Quantifying and correlating rhythm formants in speech," in *Proceedings of Linguistic Patterns in Spontaneous Speech (LPSS)* (Taipei. Academia Sinica), 1–6.

Grabe, E., and Low, E. L. (2002). "Durational variability in speech and the rhythm class hypothesis," in *Laboratory Phonology 7*, eds C. Gussenhoven, and N. Warner (Berlin; New York, NY: De Gruyter Mouton), 515–546.

Hérault, G. (1983). *Atlas Linguistique des Langues Kwa*. Abidjan: Institut de Linguistique Appliquée, Agence de Coopération Culturelle et Technique.

Juola, P. (2015). The Rowling Case: a proposed standard analytic protocol for authorship questions. *Digit. Scholar. Human.* 30(Suppl. 1), 100–113. doi: 10.1093/llc/fqv040

Kohler, K. (2009). Editorial: whither speech rhythm research? *Phonetica* 66, 5–14.

Liberman, M., and Prince, A. (1977). On stress and linguistic rhythm. *Linguist. Inq.* 8, 249–336.

Lin, X., and Gibbon, D. (2023). "Distant rhythms: computing fluency," in *Proceedings of the International Congress of Phonetic Sciences*. Prague: Charles University.

Lô, G., de Boer, V., and van Aart, C. J. (2020). Exploring West African folk narrative texts using machine learning. *Information* 11, 236. doi: 10.3390/info11050236

Marchese, L. (1983). *Atlas linguistique Kru*. Abidjan: Institut de Linguistique Appliquée, Agence de Coopération Culturelle et Technique.

Milde, J.-T. (2002). "The TASX-environment: an XML-based toolset for the creation of multimodal corpora," in *COLING-02: The 2nd Workshop on NLP and XML (NLPXML-2002)* (Stroudsberg, PA: Association for Computational Linguistics), 1–6.

Nakamura, S., and Sagisaka, Y. (2011). "A requirement of texts for evaluation of rhythm in English speech by learners," in *17th International Congress of Phonetic Sciences*, ed I. P. Association (Hong Kong: International Phonetics Association), 1438–1441.

Nerbonne, J., and Kretzschmar, W. (2003). "Introducing computational techniques in dialectometry," in *Computers and the Humanities, Computational Methods in Dialectometry*, eds B. Miller, A. Lieto, R. Ronfard, S. G. Ware, and M. A. Finlayson, (Philadelphia: Kluwer) 245–255.

Ninan, O. D., Ajíbádé, G. O., and Odéjobí, O. A. (2016). "Appraisal of computational model for Yorùbá folktale narrative," in *Proceedings of the 7th Workshop on Computational Models of Narrative (CMN 2016)*, eds B. Miller, A. Lieto, R. Ronfard, S. G. Ware, and M. A. Finlayson (Kraków), 1–14.

Ohala, J. (1992). "Alternatives to the sonority hierarchy for explaining segmental sequential constraints," in *Papers from the Parasession on the Syllable* (Chicago, IL: Chicago Linguistics Society), 319–338.

Pierrehumbert, J. B. (1980). *The Phonology and Phonetics of English Intonation*. [Thesis (Ph.D.)]. Cambridge, MA: Massachusetts Institute of Technology, Dept. of Linguistics and Philosophy.

Rossini, N., and Gibbon, D. (2011). "Why gesture without speech but not talk without gesture?," in *Gesture and Speech Interaction Conference (GESPIN 2011)* Bielefeld.

Rybicki, J., and Eder, M. (2011). Deeper Delta across genres and languages: do we really need the most frequent words? *Liter. Linguist. Comp.* 26, 315–321.

Savoy, J. (2020). *Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling*. Cham: Springer.

Tilsen, S., and Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. *J. Acoust. Soc. Am.* 134, 628–639. doi: 10.1121/1.4807565

Tilsen, S., and Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. *J. Acoust. Soc. Am.* 124, 34–39. doi: 10.1121/1.2947626

Todd, N. P. M., and Brown, G. J. (1994). "A computational model of prosody perception," in *Proceedings of the International Conference on Speech and Language Processing* (Yokohama), 127–130. doi: 10.21437/ICSLP.1994-35

Traunmüller, H. (1994). Conventional, biological, and environmental factors in speech communication: a modulation theory. *Phonetica* 51, 170–183.

Urua, E.-A. E. (2004). Illustrations of the ipa: Ibibio. *J. Int. Phonet. Assoc.* 34, 105–109.

White, L., and Malisz, Z. (2020). "Speech rhythm and timing," in *The Oxford Handbook of Language Prosody*, eds C. Gussenhoven, and A. Chen (Oxford: Oxford University Press), 167–182.

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 22 May 2023
DOI 10.3389/fcomm.2023.1115360

Check for updates

OPEN ACCESS

EDITED BY
Sandra Madureira,
PUCSP, Brazil

REVIEWED BY
Wellington Da Silva,
Sensedia, Brazil
Kaisa Laitinen,
University of Jyväskylä, Finland

*CORRESPONDENCE
Kerstin Fischer
✉ kerstin@sdu.dk

RECEIVED 03 December 2022
ACCEPTED 10 April 2023
PUBLISHED 22 May 2023

CITATION
Fucinato K, Niebuhr O, Nørskov S and Fischer K (2023) Charismatic speech features in robot instructions enhance team creativity. *Front. Commun.* 8:1115360. doi: 10.3389/fcomm.2023.1115360

Charismatic speech features in robot instructions enhance team creativity

Karen Fucinato[1], Oliver Niebuhr[2], Sladjana Nørskov[3] and Kerstin Fischer[1]*

[1]Department of Design and Communication, University of Southern Denmark, Kolding, Denmark, [2]Department of Mechanical and Electrical Engineering, Centre for Industrial Electronics, University of Southern Denmark, Sonderborg, Denmark, [3]Department of Business Development and Technology, Aarhus University, Herning, Denmark

This study examines whether a social robot can enable team creativity and increase team performance depending on its speaking style. The aim is to provide insight into human teams' creativity and performance when exposed to different ways of speaking by a social robot, that is, when the robotic creativity facilitator is using different acoustic-prosodic features. In one condition, participants received their instructions from the robot in a "charismatic" speaking style, in the other, the robot used a less engaging way of speaking. The results show that when the robot's speech is based on charismatic speech characteristics, it is significantly better at enhancing team creativity and performance than when its speech uses fewer charismatic speech characteristics. The robot's speaking style thus influences its effectiveness as team creativity facilitator.

KEYWORDS

charismatic speech, human-robot interaction, team creativity, robot speech, speech prosody

1. Introduction

Social robots are increasingly used in contexts in which they inform or instruct children and adults, for example, as museum guides (Shin et al., 2016), tutors (Han et al., 2008), or collaborators (Barrett et al., 2012). There is a growing interest in how robots can shape the social dynamics of teams, often with the goal to improve collaborative outcomes (De Visser et al., 2020). Teams are the most common form of collaboration and play a crucial role in enhancing organizational innovation and performance (Salas et al., 2008; Fay et al., 2015; McDowell et al., 2016). Research on teams has received extensive attention from, *inter alia*, organizational, psychology, sociology, communication and information science scholars. Across these disciplines, teams are typically defined as consisting of two or more individuals, having a common goal, performing tasks that interdependent in terms of workflow, goals and outcomes, and being embedded in an organizational context (Alderfer, 1977; Hackman, 1992; Salas et al., 1992; Kozlowski and Bell, 2003). Another important characteristic of teams is that they interact socially, face-to-face or virtually, and can be collocated, geographically dispersed or hybrid (Hinds and Bailey, 2003; Cousins et al., 2007).

Compared with individuals, teams are better at facilitating and improving the creative process (Leavitt, 1974), (Puranam, 2018). As teams are so central to creativity and collaboration, research has started to inquire into how machines, such as virtually and physically embodied robots, may affect teamwork (Sebo et al., 2020), (Yan et al., 2020). We define creativity in terms of outcomes, namely as "the production of novel and useful ideas in any domain" (Amabile et al., 1996, p. 1155) or more simply put as "novel, potentially useful ideas" (Seibt et al., 2020, p. 934). Many organizations are dependent on developments driven by creative ideas, designs, solutions, products, and services (Kozlowski and Bell, 2003). In these contexts, robots should enhance, rather than hinder, human performance.

While research on individual creativity has been investigated extensively, research on team creativity lags behind (Gilson et al., 2015). This is, to a certain extent, attributed to the fact that investigating creativity at the team level is more complex as team creativity is not warranted even in the presence of enabling conditions as other factors might be at play (Gilson et al., 2019). Nonetheless, team processes such as communication and coordination are essential for team performance, including team creativity (Leavitt, 1974; LePine et al., 2008). Indeed, the way team members communicate with each other may motivate them to share knowledge with the team and help them develop positive team relationships (Camden and Kennedy, 1986; Inglis, 1993; Hinds and Pfeffer, 2003). This study thus examines whether such effects of communication extend to team creativity, or, more specifically, whether and how speaking styles of a robotic facilitator affect team creativity.

While human-robot interaction research has demonstrated that robots may be successful as creativity facilitators (e.g., Kahn et al., 2016), a robot design choice that has hitherto received little attention is its speech characteristics. Verbal communication is one of the most natural ways for humans to communicate (Crumpton and Bethel, 2015). Similarly, it plays an important role in many human-robot interaction scenarios and has been shown to have a substantial effect on the persuasiveness and impressionistic evaluation of robots as well (Fischer et al., 2020). Despite this important role, robot designers have paid relatively little attention to the voices used on robots (McGinn and Torre, 2019), and specifically not in relation to teamwork and creativity.

There is however reason to assume that charismatic speech may have pragmatic and cognitive effects that lead to improved team creativity; in particular, there is empirical evidence that charismatic speech increases the speaker's perceived competence, self-confidence, and passion. For instance, Niebuhr and Michalsky (2019) found that speaking rate and silent pause duration are mostly related to perceived speaker competence, while pitch range and emphatic-accent frequency are related to passion and a higher level of arousal, which in turn can lead to heightened attention and hence more engagement (cf. Niebuhr, 2021). These features could thus also prove beneficial for increasing team creativity. Moreover, charismatic leadership is known to have many positive effects on their listeners, and charismatic leadership has been found to promote task performance (Holladay and Coombs, 1994; Cho and Cho, 2014) and team creativity (Murphy and Ensher, 2008; Sørensen, 2013).

In this study, we therefore examine how a robot's speech characteristics can influence the creativity of human teams in order to inform social robot design choices in the future. This leads us to the research question which guides the present study: How does a social robot's charismatic vs. non-charismatic speaking style influence team creativity and performance?

2. Previous work

Previous work concerns studies addressing how robots can enhance team creativity, and work on the effect of charismatic speaking styles.

2.1. Robots as creativity facilitators

Many studies have shown that robots can function as creativity facilitators; for instance, Kahn et al. (2016) show that a robot leads to more creative results than a self-paced presentation, for instance. Alderfer (1977), Ali et al. (2019) show that social robots are principally suited to facilitate creativity in children. Similarly, Alves-Oliveira et al. (2020) had a robot engage children in joint storytelling activities and found that verbal creativity levels increased, that children produced more objective ideas, and that they elaborated on their ideas more. Concerning adults, Geerts et al. (2021) found no difference between human and robot creativity facilitators on participants' productivity in a brainstorming task. Thus, previous work shows that robots are generally suited to engage people in creativity tasks, and that they perform better than other technologies and not worse than human facilitators.

Concerning the determining factors of creativity facilitation, Elgarf et al. (2021) found that the degree to which a robot is creative does not influence the extent to which children are creative during a joint storytelling task. Hu et al. (Howell, 2010) address factors that may impact the effects of robot behaviors on people's creativity. In their study with adults, they found both internal (e.g., personality, knowledge and motivation) and external factors (e.g., cultural differences) to influence the extent to which a robot can enhance creativity. The role of speech characteristics of the robot facilitator have not yet been addressed, and the studies have so far concentrated on investigating creativity at the individual level, not on team creativity.

In contrast, much previous work investigates whether robots can influence the performance of teams. For instance, Jung et al. (2015) explore whether robot interventions can positively influence conflict dynamics by addressing interpersonal violations that occur during a problem-solving task. They found that the robot's interventions increased the groups' awareness of conflict, thus acting against the tendency to suppress the conflict. These findings suggest that robots can help a team by the management of occurrences such as conflict. In a recent study by Rosenberg-Kima et al. (2020), a social robot facilitated a small collaborative group activity of students in higher education to examine the effects of facilitation on attitudes toward the activity facilitation, the group activity, and the robot. Students perceived the robot positively, as friendly and responsive, even though the robot did not directly respond to the students' verbal communications. Furthermore, the robot was perceived to have advantages over a human facilitator, such as better time management, objectivity, and efficiency. Similarly, De Visser et al. (2020) suggest that robots may influence trust in human–robot teams through "relationship equity". Thus, robots have been shown to be able to influence team performance positively in general.

2.2. Robots and charismatic speech

Previous work on robot speech highlights the importance in considering robot speech characteristics and specifically the benefits of prosodic-expressive speech. Nass and Brave (2005), for instance, find that people prefer voice interfaces that match the listener in personality and emotional expression. Jung (Jung, 2017) suggests that robots need to align their emotional expression with users in order to participate in socially acceptable interactions with humans. Furthermore, the prosodic features of robot utterances may influence the robot's persuasiveness. Specifically, Fischer et al. (2020) show that robots are more persuasive when using speech characteristics similar to those used by charismatic speakers. They examine the effects of manipulating robot speech melody and other prosodic features on robot persuasiveness in three different studies. The robots' speech was manipulated to correspond to the speech characteristics of two famous CEO's, Steve Jobs or Mark Zuckerberg. Overall, the three studies demonstrate that robots that use Steve Jobs' speech characteristics, participants followed the respective robot's requests and implicit advice more often, and the robots were rated as more charismatic. Similarly, Fischer et al. (2021) find that the same manipulations in robot voices make students perform better in an (unrelated) foreign language learning task. Similarly, Westlund et al. (2015) used a more expressive speaking style produced by an actress, which led to better learning outcomes. Thus, a robot using a charismatic speaking style can encourage listeners to do more work or to do it more thoroughly.

These findings suggest that speech characteristics should be taken into account in robot speech creation and that they may potentially be quite powerful. At the same time, a recent study by Velner et al. (2020) suggests that simply copying general human speech characteristics may not lead to the expected results, so case by case analyses of the effects of speech characteristics in human-robot interaction seem warranted. Furthermore, in the studies reported on above, a broad range of acoustic-prosodic features was employed to manipulate the robot voices. While obviously effective, it is not clear from these studies which features contribute to the perception of the robot as charismatic. In the current study, we therefore focus on a minimum of manipulations to be able to better track the effects of particular speech characteristics.

3. The present study

Our study investigates the role of the robot's speech characteristics in promoting better team performance and creativity in a task that asks participants first to come up with associations about a given picture and then to jointly apply those associations to brainstorming ideas for a new product. The independent variable tested is therefore the robot's speaking style. The dependent variables concern robot perception, team performance and team creativity results from student teams completing a divergent-thinking task. Based on our literature review above, we expect:

H1: A robot acting as a creativity facilitator will be perceived more positively, as more charismatic and as having more interactive social capacities when it uses charismatic speech, compared to a robot that uses a less charismatic speaking style.

H2: Team performance will be perceived more positively when participants are exposed to a robot facilitator that uses charismatic speech, compared to a robot that uses a less charismatic speaking style.

H3: Creative outcomes will improve when group members are exposed to a robot facilitator that uses charismatic speech, such that they produce more, more original, more flexible, and more elaborate ideas, compared to a robot that uses a less charismatic speaking style.

4. Method

The study was carried out as a between-subject experiment, in which teams of students were instructed in a set of robot videos how to carry out a creativity task, where the robot used one speaking style in one condition and another in the other condition.

4.1. Participants

The participants of this study were students from five different Master and Bachelor programs in Philosophy, User Studies, Innovation Management, Communication and Design, based at four different universities in Denmark and Switzerland (convenience sample). The experiments were carried out during the COVID-19 pandemic, when all courses were held online. Team creativity was relevant in some way or other to the content of all of these courses, and the respective teachers had invited us in because they believed the activities involved to be interesting to their students. Altogether, there were 100 student participants in five different online classes; further demographic information was not elicited, in line with their identity as course participants. Each experiment session took place during their respective class lecture times *via* Zoom. The whole experiment from start to finish took about 30–40 min.

4.2. Experiment procedure

The experiment was presented to the student participants as a "Creativity Workshop". The participants were not told anything about social robots, nor that a robot would be guiding their task; instead, they were told that the study was looking into how groups can work together creatively online.

In order to administer this Creativity Workshop, an online collective whiteboard Miro was used (see Figure 1). This allowed us to prepare a journey for them with different steps they had to perform, as well as the student groups to ideate together simultaneously using digital post-it notes and live mouse cursers that indicate a user's positions, movements, or contributions. This platform also helped the students with idea consolidation, where all ideas can be captured in one place (Brem, 2019).

After a brief introduction to the "creativity workshop," participants were divided into groups of 3–4 randomly using Zoom's breakout room feature, which moves users automatically into separate meeting rooms. Participants were then directed to Google Slides where they were given three sets of instructions: fill out a consent form, write their names down with their other group members next to their corresponding group number, and click on their designated Miro board link. Furthermore, we asked them to record their Miro board screen. The informed

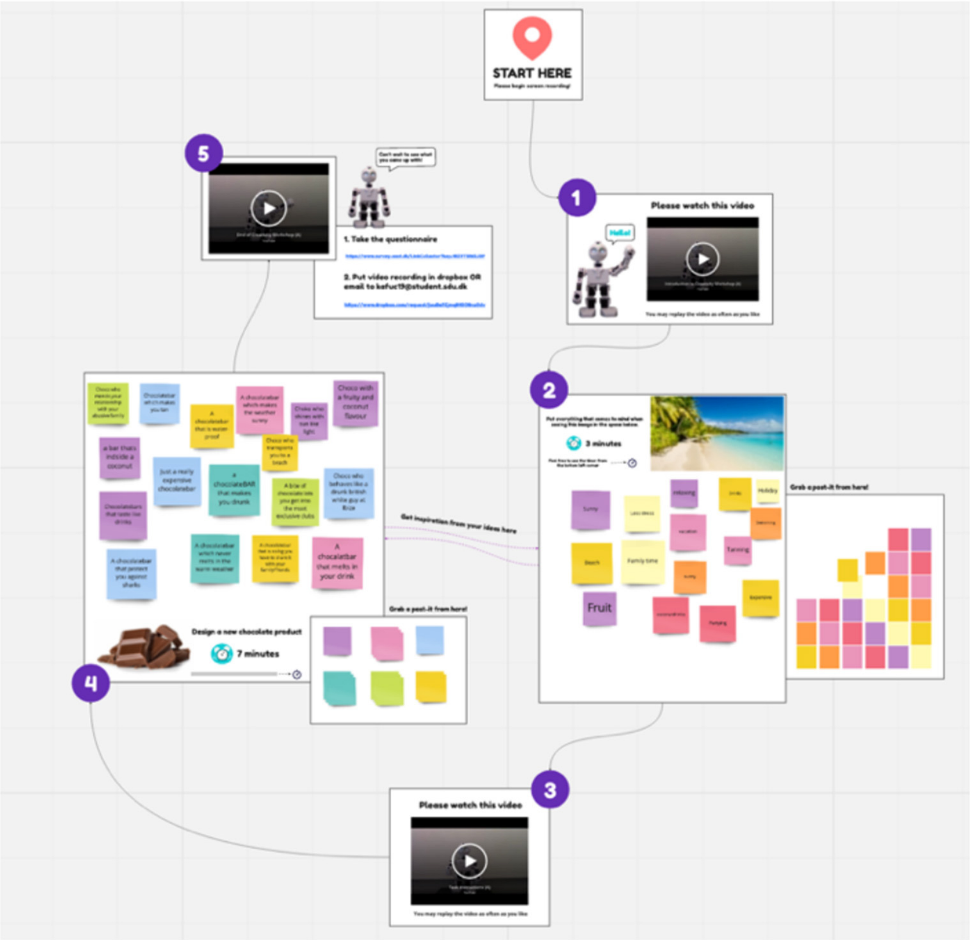
START HERE
Please begin screen recording!
1
Please watch this video
Hello!
You may replay the video as often as you like
2
3 minutes
Fruit
Grab a post-it from here!
Get inspiration from your ideas here
Design a new chocolate product
7 minutes
Grab a post-it from here!
3
Please watch this video
You may replay the video as often as you like
4
5
Can't wait to see what you came up with!
1. Take the questionnaire
2. Put video recording in dropbox OR email to kafuc19@student.sdu.dk

FIGURE 1
Experiment set-up on the miro board (3 robot videos, brainstorming task and idea generation task).

consent form was based on GDPR guidelines and stated that both participation and recording were voluntary (many groups chose not to record themselves); IRB evaluation was not provided since IRB evaluation is provided only for newly externally funded projects in the country in which the experiment was carried out. However, ethical considerations were discussed with an ethical advisor, the participants were provided with consent forms that fully conform with GDPR, and they were extensively debriefed.

When participants arrived at their designated Miro board, they began screen recording with video and audio. The boards were identical in both conditions, but different with respect to the videos provided. Each step of the task was numbered to help the student groups navigate the board.

The set-up of the experiment was created based on a creativity workshop technique described in Brem and Brem (2019) called visual synectics (see below), consisting of: idea collection and idea generation.

The first step was to watch the first video. In this video, the robot introduced itself and gave instructions on the first part of the task, with examples given to help avoid confusion. Participants were told to work together as a team and that there were no restrictions or evaluations on their ideas. They had 3 min to complete the first part of the task. The groups were provided with post-it notes on the side and a designated space for their idea generation. After the idea collection, participants were guided to the second video where the robot gave instructions for the second part of the task. Similar to the first video, groups were provided with examples, and they were explicitly told to be creative, to come up with as many ideas as possible and that there were no stupid ideas. This was done to prevent any possible limitations or fears of evaluation for the students. The groups had seven minutes to complete this portion of the task. When both parts of the task were completed, participants watched a final video in which the robot congratulated the group on their hard work and asked them to take the questionnaire. Then, participants were directed to complete the questionnaire. There were two identical questionnaires administered to the participants, one for those who were exposed to the first (charismatic) condition and the other for the second (non-charismatic) condition. Lastly, the groups were reminded to send a video of their screen recordings after the completion of the questionnaire.

All participants were then called back to the main session, where we asked them informally about their experience and about the robot. Finally, they were debriefed about the additional goals of the experiment (the official goal was to introduce them to a novel creativity technique, which it did).

4.3. Creativity task

A two-part divergent thinking task was developed for participants to complete based on a technique called Visual Synectics, which is a divergent thinking idea generating technique in which pictures are used as a visual stimulus to create useful ideas (Asanowicz, 2008). Successful creation depends on the ability to see metaphoric meaning, through a process called Conceptual Blending. Elements and relations between things are "blended" in a subconscious process through everyday thoughts and language. Insights gained from this process of blending promotes creative thinking and emphasizes the role of unpredictable associations and metaphors in creativity (Asanowicz, 2008).

In our study, each group was first given a photo of a beach as their visual element. They are then instructed to write down anything that comes to mind when seeing this image, such as attributes or feelings that arise in relation to the photo. For example, when seeing the image of the beach below, one may write down "palm trees", "Hawaii", or "family vacation". After that, the groups were asked (in the second robot video) to develop a new chocolate product to be sold at grocery stores for the second part of the task. They were instructed to generate their ideas off of the descriptions they used for the first part of the task as inspiration. For example, one could think of "a chocolate bar in the shape of a palm tree" or "a chocolate bar wearing a lei in Hawaii". It was encouraged that the group members should work together as a team when creating ideas in both parts of the task. This process is comparable to group brainstorming, because it can stimulate people to consider categories or thoughts that they might not otherwise think of Paulus et al. (2012).

The social robot acted as a creative workshop facilitator, where it explained the two-part task that participants in a group needed to complete. The social robot provided directions to the creativity task and some examples, and it explicitly encouraged the team members to be creative and work as a team together through direct and unambiguous instructions.

4.4. Experiment stimuli

The participants in each group watched three separate videos during the experiment. In the first video, the EZ-robot (see Figure 2) gave instructions for the first part of the task. In the second video, it provided the participants with instructions and examples for the second part of the divergent-thinking task. In the third video, the robot thanked the group for their hard work and reminded participants to fill out the questionnaire and to send their screen recording videos. A set of videos were made using the charismatic speech and another set of videos used the non-charismatic speech.

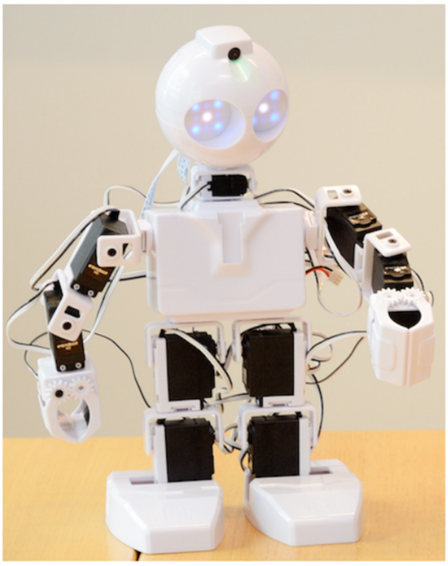

FIGURE 2
EZ-robot.

For the robot videos, the JD Humanoid robot "EZ-robot" was used, a small, stylized humanoid robot[1] with 16 degrees of freedom and 16 servos, which can be operated by a simple interface. Its "eyes" consist of nine LEDs each. In the videos we recorded, the robot moved its arms and head slightly to match the content of the message in the corresponding manipulated audio file. Three videos were created, which were subsequently matched with two different versions of synthesized speech.

Two different versions of the robot's utterances were created from the same raw synthesized speech file called [en-us] Jack Bailey—S from a free Text to Speech Software (TTS) called wideo.co, which sounds like a male, adult American English speaker. Then, using the speech analysis software *praat* (Boersma and Weenink, 2001), the overall relationship between formants and pitch height was adjusted using the gender manipulation function in *praat*. For the more charismatic speaking style, the acoustic-prosodic features of the robot utterances were manipulated in the direction of a female speaker because this corresponds to a shorter vocal space. This in turn is related to smiling and higher arousal and hence engagement (Niebuhr, 2021). The manipulation was only adjusted by 5%, which is too little to make the male voice sound female. Rather, the result is a smiling-sounding male. For the non-charismatic speaking style, the manipulation was adjusted by 5% in the other direction. In other words, the voice is more male and less smiling sounding. The pitch level and pitch range were also changed, such that the charismatic voice is 15 hertz higher and the range is 3 semitones larger. In contrast, for the non-charismatic voice, the pitch height was lowered by 15 hertz and the range by 3 semitones. These manipulations are again related to speaker engagement and to higher flexibility (Berger et al., 2017).

[1] https://www.ez-robot.com/learn-robotics-getting-started-humanoid-robot-kit.html

Furthermore, local manipulations of accents (i.e., emphasized syllables) in the charismatic prosody were done to increase the number of accents to 4–5 per minute by means of lengthening of voiced consonants, again for the sake of increasing the impression of engagement and flexibility. The accents in the synthesized voice were reduced to 0 for the non-charismatic prosody. It is important to note that the manipulated speech features in both conditions did not change the speaker's voice or voice quality itself. Also, both versions of the speech manipulation were clear, pleasant and understandable to users even though we refer to this condition as "non-charismatic", as confirmed in the informal discussions with the participants after the experiments.[2]

4.5. Post-experimental questionnaire

In order to test the three hypotheses and determine if the charismatic prosody of a robot can influence creativity and team performance, previously validated scales were used in a questionnaire that participants took after watching the videos of the robot and completing the divergent thinking task. The measured variables used in the questionnaire include questions about charisma-related characteristics and the perceived sociality of the robot, and self-reported team performance. Most of the questions given to measure each variable used Likert-scales.

4.5.1. Subjective evaluation of the robot

The first part of the questionnaire asked participants to rate the robot with charisma-related questions as well as its perceived sociality. The charisma-related questions asked participants to rate the robot on a scale from 1 to 7 on the following adjectives, where 1 corresponds to "not at all" and 7 to "very much". The adjectives used were: Enthusiastic, Charming, Convincing, Engaging, Boring, and Passionate. The scale from Fischer et al. (2020), was used, which in turn relied on measures used by Rosenberg and Hirschberg (2009) to study charismatic speech in humans. The adjectives "Self-confident", "Uninspiring", and "Charismatic" were added to the list of attributes to allow a more detailed analysis of the specific expected effects of the manipulations in speaking styles made.

4.5.2. Perceived sociality of the robot

The scale of perceived sociality measures the social robotic properties that can be related to the types of human experiences and interactive dispositions. The perceived sociality scale used in this study is based on previous research by Seibt et al. (2020), who argue that social interactions with robots are not always the result of anthropomorphizing, but of "sociomorphing", i.e., the perception of actual non-human social capacities. The set of questions regarding perceived sociality asks what the participant anticipates from the social robot such as expectations of certain interactive capacities.

The participants rated six statements on a scale from 1 (strongly disagree) to 5 (strongly agree). The two statements, "The robot will understand what I feel" and "The robot will understand how I reason" ask about different types of expected empathy. The question, "The robot will want to be with me and close to me," explores whether any capacity to experience and enjoy sociality is attributed to the robot. Whether participants attribute the social robot to the capacity of counter factional imagination is asked with the question, "The robot and I could try to imagine together what it would be like to be able to fly". The statement, "If the robot were to walk off in all directions, I would keep it on a leash" is expected to indicate the relationship between capacities ascribed and moral dignity, which in turn helps to distinguish anthropomorphizing and sociomorphing (Seibt et al., 2020).

4.5.3. Perceived team performance

Team performance variables were measured using scales from previous work that relate to different team performance constructs: outcomes of teamwork as well as the team's process. Perceptions of outcomes of teamwork were measured using a 5-point Likert scale asking to rate their team's performance where 1 is "strongly disagree" and 5 is "strongly agree". This scale used four subscales of teamwork outcomes: team effectiveness, team efficiency, team psychological safety, and team creative performance. Effectiveness refers to the degree to which team members meet the requirements and expectations of work quality as well as team collective goals (Velner et al., 2020). The items of team effectiveness were selected and slightly adapted from previous research on group effectiveness (Inglis, 1993; Velner et al., 2020). The questions include, "My team was effective in getting the task done" and "My team completed the task successfully". Efficiency refers to the team's ability to adhere to means of producing satisfactory results with minimum inputs as well as the ability to complete tasks quickly (Velner et al., 2020). Efficiency-related questions include "Overall, my team did our task in a time-efficient way" and "My team completed this task smoothly".

The subscale of team psychological safety is a construct that addresses a shared belief held by team members that the team is safe for interpersonal risk taking (De Visser et al., 2020). The term psychological safety is meant to suggest a sense of confidence that the team will not embarrass, reject, or punish someone for speaking up. Psychological safety is one of the key factors influencing team creativity performance (De Visser et al., 2020). Questions for psychological safety in this study include, "It was difficult to ask other members of this team for help", "It is safe to take a risk on this team," and "No one on this team would deliberately act in a way that would undermine my efforts."

To measure a team's process, collective engagement was assessed using an Engagement Questionnaire (Salanova et al., 2003). It was slightly adapted for use in work groups. Collective engagement is central for creativity and occurs when group members stimulate one another's divergent thinking and their

[2] Links to the videos: Condition 1 (charismatic prosody).
Video 1: https://youtu.be/x5wvRzulbzo.
Video 2: https://youtu.be/dk-eEv849pM.
Video 3: https://youtu.be/ecEQw9CNmCY.
Condition 2 (non-charismatic prosody).
Video 1: https://youtu.be/MLsXNImwmuM.
Video 2: https://youtu.be/CqszFYzSCJI.
Video 3: https://youtu.be/G3eR-mzIPbI.

individual ideas are combined for the group's creative output (Harvey and Kou, 2013). The collective engagement measure is scored in three subgroups: Collective Vigor (seven items; e.g., "During the task, my group felt full of energy"), Collective Dedication (four items; e.g., "My group felt very motivated to do a good job"), and Collective Absorption (seven items; e.g., "Time was flying when my group was working"). The variables of the measure used a 5-point Likert scale asking to rate the questions from 1 to 5, where 1 is "strongly disagree" and 5 is "strongly agree".

A selection of relevant questions based on a scale developed by Shin et al. (Shalley et al., 2004) was used to measure each team's self-perception w.r.t. their creative performance (Brem and Brem, 2019). There were two questions related to group perception: "My team members' work was original, adaptive, and practical," and "My team members generated creative ideas". Additionally, subjective creative perception questions were asked after they had worked with their teams in order to gauge their perceived creativity compared to the creativity task scoring. These two questions were, "How novel are your group's ideas," and "How valuable are your group's ideas?". Lastly, a 5-point slider question was presented on participants' perception of the extent to which they contributed to the group result, compared to their teammates.

4.5.4. Creativity analysis

Participants' productions were scored using Guilford's (Gilson et al., 2019) categories of divergent thinking perspectives: Fluency, Flexibility, Originality, and Elaboration. Fluency is the quantity of ideas a group generates. Therefore, if the group gave five responses, that would equal a score of 5. Flexibility is the number of different categories of relevant responses or the breadth of categories they cover with the ideas. Thus, if there were two categories identified, flexibility would be scored 2. Originality is the relative novelty and uniqueness of each answer. Each response is compared to the total amount of responses from all of the groups who participated in the task. Responses that were given by only 1% of the group are considered unique (1 point). Responses that were only given by that particular group are considered one of a kind (2 points). Higher total scores for a group indicate more original thinking. Elaboration is the amount of detail in the responses. For example, "a blue chocolate" = 0 points, but "a blue chocolate wearing sunglasses" = 1. An additional point would be given for any further details given, such as what the chocolate tastes like or details in the design. Since coding is highly mechanistic, in line with previous work [e.g., 26], we did not deem it necessary to calculate intercoder reliability.

Another indicator of divergent thinking, participation, was added for this study to see how much the group was collaborating together as a team. Participation was calculated by measuring the total amount of time participants spent talking compared to how long it took them to complete the two parts of the task. This was taken into account because more evenly distributed conversational turn-taking in a group can predict better performance, as well as a more active discussion leads to added innovative group performance (Tennent et al., 2019). The more time a group spent talking during the task, the better team performance is considered. The English subset of the screen recorded videos taken during the experiment were analyzed concerning the percentage of time that group members spent talking to one another in order to find their participation levels.

5. Results

The 100 student participants who attended the experiment formed 30 different groups who completed the task together as a team, and 76 students filled out the questionnaire completely; 40 students were exposed to condition 1 (C1, i.e., charismatic speech) and 36 students were exposed to condition 2 (C2, i.e., non-charismatic speech). In the inferential statistics below, M_{C1} andM_{C2} refer to the mean values of the two speaking-style or charisma conditions.

The questionnaire and behavioral data were analyzed using multivariate analyses of covariance (MANCOVAs). Three MANCONAs were performed, one per set of obtained rating data, i.e., creativity performance, subjective robot evaluation, and perceived sociality/team performance. Each MANCOVA aimed to test the effects of our independent variable, i.e., the two-level fixed factor Condition (speaking styles C1 vs. C2, see above), on the dependent variables represented by the rating scales in each set.

In order to reduce in these tests the statistical noise that has been introduced by distributing the overall experiment across several classes and subgroups of students, we made Class (the different courses in which the experiments were conducted) and Group (the small groups of 3–4 students in each class) covariates of the analysis. That is, Class and Group were random effects in our experimental procedure and we addressed them as such in the statistical tests. To that end, we split up Class and Group into categorical dummy variables, following the procedure suggested in the statistical-methods handbook by Howell (Howell, 2010, p. 600), and we defined both covariates as being categorical (non-continuous) in each MANCOVA. Note that this test design did not include interaction effects between the covariates and our dependent variables (ratings), but neither did we hypothesize any such interactions nor were relevant to the questions of our study.

With regard to the creativity performance (see Figure 3), we found a significant main effect of Condition ($F_{[4,69]} = 6.893$, $p < 0.001$, $\eta_p{}^2 = 0.286$). The main effect is based on influences of speaking style on originality (M = 26.63, SD = 12.54 in the charismatic and M = 18.44, SD = 18.04 in the non-charismatic condition; $F_{[1,72]} = 9.889$, $p = 0.002$, $\eta_p{}^2 = 0.121$) and elaboration (M = 9.95, SD = 5.82 vs. M = 7.33, SD = 4.15; $F_{[1,72]} = 4.628$, $p = 0.035$, $\eta_p{}^2 = 0.067$). Both criteria scored 35–45% higher in connection with the robot's charismatic tone of voice. The MANCOVA also yielded significant main effects of both Class ($F_{[4,69]} = 2.871$, $p = 0.029$, $\eta_p{}^2 = 0.143$) and Group ($F_{[4,69]} = 7.758$, $p < 0.001$, $\eta_p{}^2 = 0.310$), however, these did not concern originality and elaboration. Classes differed in terms of flexibility ($F_{[1,72]} = 13.019$, $p = 0.035$, $\eta_p{}^2 = 0.153$) and Groups in terms of Fluency ($F_{[1,72]} = 4.127$, $p = 0.046$, $\eta_p{}^2 = 0.054$), with mainly one class or group standing out against the others.

Regarding the subjective evaluation of the robot (see Figure 4), Condition, i.e., speaking style, came out as the only significant main effect on participants' perception of the robot ($F_{[9,58]} = 3.054$, $p = 0.005$, $\eta_p{}^2 = 0.321$). That is, ratings did not differ significantly as a function of the two covariates Class and Group. More specifically, the charismatic voice was found to make the robot sound more enthusiastic (M = 4.72, SD = 1.30 vs. M = 3.88, SD = 1.45; $F_{[1,66]}$ = 10.257, p = 0.002, $\eta_p{}^2$ = 0.135) and, above all, more passionate (M = 4.10, SD = 1.48 vs. M = 2.90, SD = 1.16; $F_{[1,66]}$ = 13.622, $p < 0.001$, $\eta_p{}^2$ = 0.171). Also, in line with these impressions, further statistical trends emerged showing that the robot with the charismatic speaking style tended to be perceived as less uninspiring (M = 2.9, SD = 0.26 vs. M = 3.54, SD = 0.29; $F_{[1,66]}$ = 3.908, p = 0.052, $\eta_p{}^2$ = 0.056) and as more charming (M = 4.6, SD = 0.29 vs. M = 3.9, SD = 0.32; $F_{[1,66]}$ = 4.502, p = 0.079, $\eta_p{}^2$ = 0.041).

For the perceived sociality and perceived team performances, there was a main effect of Condition on four task- or output-oriented questions: "The robot will understand how I reason" (M = 2.1, SD = 0.21 vs. M = 2.6, SD = 0.23; $F_{[1,51]}$ = 3.991, p = 0.047, $\eta_p{}^2$=0.073), "My group felt very resilient during the task" (M = 3.05, SD = 0.93 for the charismatic and M = 3.47, SD = 0.84 for the non-charismatic condition; $F_{[1,51]}$ = 4.087, p = 0.045, $\eta_p{}^2$ = 0.074). Note that robot empathy and perceived team resilience were rated slightly better (higher) with the non-charismatic robot. In contrast, in the two questions on the perceived quality of ideas, it was the charismatic robot that caused the better (higher) team-performance ratings: "How valuable are your group's ideas" (M = 3.3, SD = 0.15 vs. M = 3.0, SD = 0.17; $F_{[1,51]}$ = 4.270, p = 0.044, $\eta_p{}^2$ = 0.077) and "How novel are your group's ideas" (M = 3.38, SD = 0.91 for the charismatic vs. M = 3.06, SD = 0.79 for the non-charismatic robot; $F_{[1,51]}$ = 6.990, p = 0.011, $\eta_p{}^2$ = 0.12).

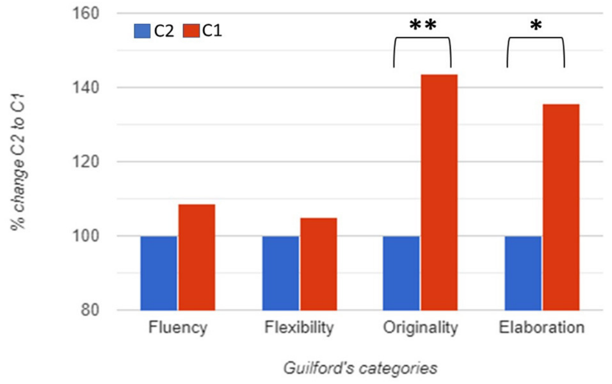

FIGURE 3
Results of the creativity analysis (C1: charismatic vs. C2: non-charismatic speech).

Like for the creativity performance, team performance also differed significantly with respect to the two covariates, but only on one scale per covariate and again not along the same lines as for the factor Condition. The two affected scales concern group dynamics, i.e., what the team members thought about the group or the other group members. In the case of Class, differences emerged along the scale "My team members generated creative ideas" ($F_{[1,51]}$ = 10.550, p = 0.002, $\eta_p{}^2$ = 0.177). For Group, the affected scale was "My team completed this task smoothly" ($F_{[1,51]}$ = 4.750, p = 0.034, $\eta_p{}^2$ = 0.085).

We carried out a *post hoc* power analysis, based on the observed effect sizes of our main variable of 0.28 and 0.32; given our 76 participants, the estimated power is.99 for the creativity analyses and the subjective evaluation of the robot. However, given effect sizes of 0.07–0.12 for the perceived sociality and team performance questions, the study must be considered underpowered.

The analysis of the videos provided showed considerable differences between groups in terms of teamwork; some groups carried out the tasks silently and individually while others collaborated intensely, such that they built on each other's ideas and jointly discussed possible product innovations.

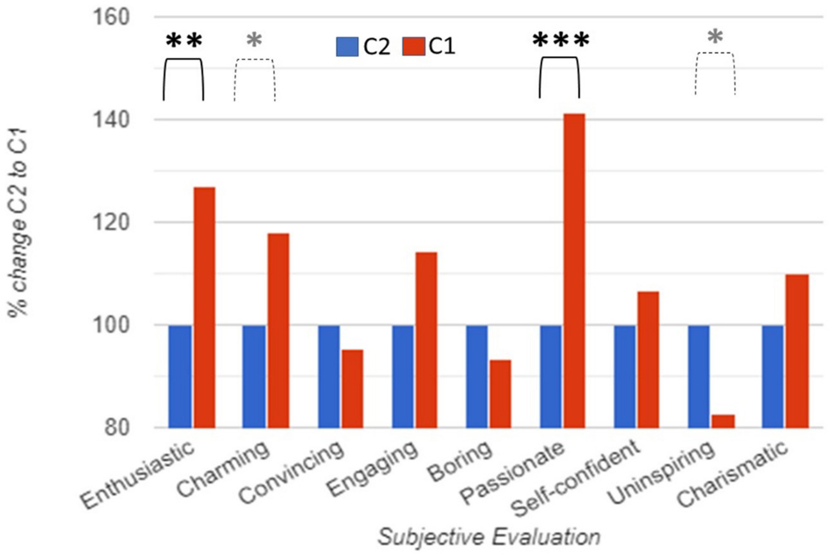

FIGURE 4
Relative differences in the subjective evaluation of the robot (C1: charismatic vs. C2: non-charismatic speech).

6. Discussion

In this study, the behavioral effects of a social robot's charismatic speech on human team members have been examined to see if creativity and team performance can be positively impacted. The results support our hypothesis (H3) that charismatic instructions by a robot improve team creativity concerning two of the four measures of creativity: "Originality" and "Elaboration." Thus, the groups exhibited overall better creative processes and creative outcomes than groups that were facilitated by a non-charismatic speaker.

Similarly, if the social robot used a charismatic speaking style, it was perceived positively and was seen as having interactive social capacities, supporting the hypothesis (H1) that charismatic prosody leads to more positive attributions. Regarding the evaluation of the robot, participants viewed the robot that used the charismatic speech as more enthusiastic and passionate, and to a weaker extent also as less uninspiring and more charming. Remarkably, these attributions were consistent across courses and groups, confirming that charisma is "easy to sense but hard to define" (Niebuhr and Michalsky, 2019). The traits attributed to the more charismatic robot are likely to have an effect on the team members' creativity (see Niebuhr, 2021) by suggesting higher arousal levels and more engagement (the 'smiling voice'), as well as more flexibility (more emphasized syllables and higher pitch range).

Similarly, regarding perceived sociality, team performance and perceived psychological safety, there are significantly better ratings for the charismatic robot, despite more (and significant) variation between courses and groups. All of these results point to the same conclusion, that when the robot uses charismatic speech characteristics, it is rated more positively with respect to traits that are associated to charisma than the robot with a non-charismatic prosody.

The only exceptions to the positive effects of the more charismatic voice of the robot concern the extent to which the participants perceived how resilient their teams were and how empathic the robot was. That is, what makes a meeting subjectively more successful, namely discipline, perseverance and focus, was more strongly triggered by the less charismatic robot, while what makes a creative meeting objectively more successful, namely commitment, a change of perspective, and finding new ideas, was supported by the more charismatic robot. Here we can only speculate what causes this effect; one possibility may be that when a team leader is less charismatic, the team will stand more together. The finding is related to a similar observation made in Niebuhr et al. (2021) on "the sound of successful meetings". Based on a corpus of 70 idea-generation meetings of teams of 3–4 US high-school students, the authors investigated whether and how the interlocutors' prosody correlates with the subjective and objective performance indicators of such creativity-oriented meetings. They find that subjectively more successful meetings have a "sound" that is, amongst other things, characterized by an overall lower voice-pitch level, a significantly smaller pitch range and less pitch variability (e.g., due to fewer accent-related pitch movements). That is, the "sound" is more matter-of-factly and less charismatic. However, the actually *objectively* more successful meetings (in terms of the number generated ideas/solutions and their level of feasibility) were characterized by a livelier, more charismatic intonation in the sense of higher voice-pitch levels and larger and more frequent pitch movements. Why exactly this is so and why subjective and objective perceptions and indicators correlate and diverge in this way and with voice is a completely new type of question that must be addressed in future studies.

Overall, the impressions of the social robot made by participants imply that charismatic speech characteristics should be taken into account for robot synthesis in the future. This is in line with work on human communication, where speaking style is considered an essential component of speech communication (Camden and Kennedy, 1986). Furthermore, our results are consistent with findings by Fischer et al. (2020), that speech characteristics of robot speech may influence both the persuasiveness and impressionistic evaluation of the robot, and with the findings by Fay et al. (2015), that charismatic robot speech may lead to better student performance.

The results furthermore demonstrate that very few prosodic parameters can have significant effects on how the respective speaker is perceived and on how effective the robot is as a facilitator of team creativity tasks. While previous studies that demonstrated significant effects of robot speaking styles have relied on a large range of prosodic features, in this study, only pitch height, pitch range and number of accents were manipulated, i.e., how high the voice was, how large its range was, and how often the speech melody indicated that there was an important word. Thus, the study results can contribute to narrowing down which speech features the effects of charismatic speaker are actually determined by.

A possible limitation of the study is the fact that not all student groups were equally affected by the robot's speaking style as far as the objective creativity performance results are concerned. That is, we found that the courses and groups, in which the experiments were conducted, differed significantly with respect to some of the measures; specifically, classes differed in terms of flexibility and groups in terms of fluency, which means that the range of different ideas created was significantly different between students who study philosophy, design, communication, innovation management and user studies. Similarly, classes differed with respect to the extent to which they thought that they produced creative ideas. In contrast, groups differed concerning the number of ideas they produced and with respect to the extent to which they thought that they fulfilled the task smoothly. We do not know what these differences are caused by, whether they are intercultural differences, age difference, interindividual or just coincidental differences; for instance, in one class that was held as a block seminar, students reported to be extremely exhausted already before the experiment. Moreover, given their different disciplinary backgrounds, the students were used to both creativity techniques and team work to different degrees, and it turned out in the discussions after the experiments, that they also had experience with filling out such questionnaires to very different degrees. For instance, in one class, students reported that they thought that we were pulling their legs with the questionnaire; that is, they were completely unfamiliar with the questionnaire items that came from standardized questionnaires in psychology and related disciplines. Thus, potentially many different factors may have caused the differences between classes and groups observed; as noted by Gilson et al. (2019), even in the presence of the conditions that are essential for team creativity, other variables may influence the outcome, as team creativity is a complex phenomenon. It may have thus been beneficial to include some additional individual- and team-level factors, such as communication competence of the members, team history, team members' attitudes toward teamwork, etc. to improve the reliability of the findings. Furthermore, the fact the study relied on self-reports to tap into the team process may pose limitations to the validity of these findings. Nonetheless, variables such as psychological safety are meant to be assessed subjectively. Future work will have to reveal what those differences between the five classes are likely to be caused by. Nevertheless, the subjective evaluation of the robot with the charismatic speaking style as more or less enthusiastic and passionate was similar across the different classes or groups.

One may also object that participants were exposed to the robot only briefly and in non-interactive videos; thus we do not know to what extent long-term exposure to the robot's speaking style and a more interactive encounter with the robot might have. However, given the effects of human charismatic leaders and charismatic teachers on team performance (Nass and Brave, 2005; Niebuhr et al., 2016), we can expect that with more exposure or interaction, the effect will rather increase than decrease. Furthermore, studies that compare the effects of creativity techniques online and onsite report little effects of the setting; for instance, Cho and Cho [15x] show that students are more satisfied with offline collaboration and perceive offline collaboration as more effective, but they find no significant difference in student performance online and offline. We can therefore assume that charismatic speech characteristics will have similar effects in onsite creativity teamwork at least in terms of results as in the online teamwork described in this study.

7. Conclusion

Our study has shown that a charismatic speaking style may impact team creativity; specifically, very small changes in mean pitch, number of accents and pitch range can have a considerable effect on how original and elaborate team members' ideas are in a visual synectics task. A team creativity facilitor's speaking style may thus play an important role in the creative results of such teamwork in terms of originality and elaboration of the design ideas. Furthermore, the facilitator's speaking will impact how they are perceived as well as how valuable or novel the team members understand their results to be. Speaking style should thus be taken into consideration when designing robots as creativity facilitators.

Furthermore, we can conclude that by using a robot facilitator, we were able to uncover that speaking style may potentially influence creativity teams in general, and it may be useful to consider speaking style also as relevant variable in team creativity exercises in general.

Data availability statement

The raw survey 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

Under the supervision of KFi and KFu designed the text for the stimuli and the Miro board and carried out the investigation. KFu also analyzed the students' productions with respect to their creativity. ON created the speech stimuli and carried out the statistical analysis. SN, KFi, and KFu jointly developed the experimental study design and the questionnaires. All authors contributed to the writing of the paper.

Funding

The research reported here is part of the SFACTOR project (Grant Nr. 105081) supported by the Nordforsk Foundation.

Conflict of interest

ON is CEO of the speech technology company AllGoodSpeakers ApS. A full statement in this regard can be found at https://oliverniebuhr.com/conflict-of-interest.html.

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.

References

Alderfer, C. P. (1977). "Group and intergroup relations," in J. R. Hackman and J. L. Suttle eds., *Improving the quality of work life* (Palisades, CA: Goodyear), p. 227–296.

Ali, S., Moroso, T., and Breazeal, C. (2019). "Can Children Learn Creativity from a Social Robot?," in *Proceedings of the 2019 on Creativity and Cognition (CandC '19)* Association for Computing Machinery, (New York, NY, USA), p. 359–368. doi: 10.1145/3325480.3325499

Alves-Oliveira, P., Arriaga, P., Cronin, M. A., and Paiva, A. (2020). "Creativity encounters between children and robots," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (p. 379–388). doi: 10.1145/3319502.3374817

Amabile, T. M., Conti, R., Coon, H., Lazenby, J., and Herron, M. (1996). Assessing the work environment for creativity. *Acad. Manage. J.* 39, 1154–1184. doi: 10.2307/256995

Asanowicz, A. (2008). "How to Find an Idea?—Computer Aided Creativity," *Architecture in Computro [26th ECAADe Conference Proceedings]*, p. 735–742. http://papers.cumincad.org/cgi-bin/works/Show?ecaade2008_025 doi: 10.52842/conf.ecaade.2008.735

Barrett, M., Oborn, E., Orlikowski, W. J., and Yates, J. (2012). Reconfiguring boundary relations: robotic innovations in pharmacy work. *Organiz Sci.* 23, 1448–1466. doi: 10.1287/orsc.1100.0639

Berger, S., Niebuhr, O., and Peters, B. (2017). "Winning over an audience—a perception-based analysis of prosodic features of charismatic speech," in *Proceedings of the 43rd Annual Conference of The German Acoustical Society* (p. 1454–1457).

Boersma, P., and Weenink, D. J. M. (2001). PRAAT, a system for doing phonetics by computer. *Glot Int.* 4, 341–345.

Brem, A. (2019). Creativity on demand: how to plan and execute successful innovation workshops. *IEEE Eng Manage Rev.* 47, 94–98. doi: 10.1109/EMR.2019.2896557

Brem, A., and Brem, S. (2019). *Die Kreativ-Toolbox für Unternehmen.* Ideen generieren und innovatives Denken fördern. Stuttgart.

Camden, C. T., and Kennedy, C. W. (1986). Manager communicative style and nurse morale. *Hum. Commun. Res.*, 12, 551–563. doi: 10.1111/j.1468-2958.1986.tb00091.x

Cho, J. Y., and Cho, M. H. (2014). Student perceptions and performance in online and offline collaboration in an interior design studio. *Int J Technol Des Educ* 24, 473–491. doi: 10.1007/s10798-014-9265-0

Cousins, K. C., Robey, D., and Zigurs, I. (2007). Managing strategic contradictions in hybrid teams. *Eur J Inform Syst.* 16, 460–478. doi: 10.1057/palgrave.ejis.3000692

Crumpton, J., and Bethel, C. L. (2015). A survey of using vocal prosody to convey emotion in robot speech. *Int. J. Soc. Robot.* 8, 271–285. doi: 10.1007/s12369-015-0329-4

De Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., et al. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *Int. J. Social Robot.* 12, 459–478. doi: 10.1007/s12369-019-00596-x

Elgarf, M., Skantze, G., and Peters, C. (2021). "Once upon a story: can a creative storyteller robot stimulate creativity in children?," in *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents* (p. 60–67). doi: 10.1145/3472306.3478359

Fay, D., Shipton, H., West, M. A., and Patterson, M. (2015), Teamwork and organizational innovation. *Creativity Innovat. Manage.* 24, 261–277. doi: 10.1111/caim.12100

Fischer, K., Niebuhr, O., and Alm, M. (2021). Robots for foreign language learning: speaking style influences student performance. *Front. Robot.* AI, 273. doi: 10.3389/frobt.2021.680509

Fischer, K., Niebuhr, O., Jensen, L. C., and Bodenhagen, L. (2020). Speech melody matters—how robots profit from using charismatic speech. *ACM Trans. Hum. Robot Interact.* 9, 1–21. doi: 10.1145/3344274

Geerts, J., de Wit, J., and de Rooij, A (2021). Brainstorming with a social robot facilitator: better than human facilitation due to reduced evaluation apprehension?. *Front. Robot.* AI, 8, 156. doi: 10.3389/frobt.2021.657291

Gilson, L. L., Lee, Y. S. H., and Litchfield, R. C. (2019). *Advances in Team Creativity Research. Oxford Research Encyclopedia of Business and Management.* Oxford: Oxford University Press. doi: 10.1093/acrefore/9780190224851.013.171

Gilson, L. L., Lim, H. S., Litchfield, R. C., and Gilson, P. W. (2015). "Creativity in teams: A key building block for innovation and entrepreneurship," in C.E. Shalley, M.A. Hitt, and J. Zhou (Eds.), The Oxford handbook of creativity, innovation, and entrepreneurship (Oxford University Press), p. 177–204.

Hackman, J. R. (1992). "Group influences on individuals in organizations," in M. D. Dunnette and L. M. Hough, eds. *Handbook of industrial and organizational psychology* (Palo Alto, CA: Consulting Psychologist Press), p. 199–267.

Han, J-. H., Jo, M-. H., Jones, V., and Jo, Jun-H. (2008). Comparative study on the educational use of home robots for children. *J. Inform. Process. Syst.* 4, 159–168. doi: 10.3745/JIPS.2008.4.4.159

Harvey, S., and Kou, C-. Y. (2013). Collective engagement in creative tasks. *Administrat. Sci.Quarterly*, 58, 346–386. doi: 10.1177/0001839213498591

Hinds, P., and Pfeffer, J. (2003). "Why organizations don't "know what they know": Cognitive and motivational factors affecting the transfer of expertise," in M. Ackerman, V. Pipek, and V. Wulf, eds., Beyond knowledge management: Sharing expertise (Cambridge, MA: MIT Press), p. 3–26.

Hinds, P. J., and Bailey, D. E. (2003). Out of sight, out of sync: understanding conflict in distributed teams, organization. *Science.* 14, 615–632. doi: 10.1287/orsc.14.6.615.24872

Holladay, S. J., and Coombs, W. T. (1994). Speaking of visions and visions being spoken an exploration of the effects of content and delivery on perceptions of leader charisma. *Manage. Commun. Quarterly*, 8, 165–189. doi: 10.1177/0893318994008002002

Howell, D. C. (2010). *Statistical Methods for Psychology*. Belmont: Cengage Wadsworth

Inglis, M. (1993). The communicator style measure applied to nonnative speaking teaching assistants. *Int. J. Intercultural Relat.* 17, 89–105. doi: 10.1016/0147-1767(93)90014-Y

Jung, M. F. (2017). "Affective grounding in human-robot interaction," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. doi: 10.1145/2909824.3020224

Jung, M. F., Martelaro, N., and Hinds, P. J. (2015). "Using robots to moderate team conflict: the case of repairing violations," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 229–36). doi: 10.1145/2701973.2702094

Kahn, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Shen, S., Ruckert, J. H., et al. (2016). "Human creativity can be facilitated through interacting with a social robot," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. doi: 10.1109/HRI.2016.7451749

Kozlowski, S. W. J., and Bell, B. S. (2003). "Work groups and teams in organizations," In W.C. Borman, D. R. Ilgen, and R. J. Klimoski, eds. *Handbook of psychology, Vol. 12. Industrial and organizational psychology* (London: Wiley), p. 333–375. doi: 10.1002/0471264385.wei1214

Leavitt, H. J. (1974). *Suppose We Took Groups Seriously…Prepared for Western Electrics Symposium on the Hawthorne Studies*, pp. 1–20. https://eric.ed.gov/?id=ED103291.

LePine, J.A., Piccolo, R.F., Jackson, C.L., Mathieu, J.E., and Saul, J.R. (2008). A meta-analysis of teamwork processes: tests of a multidimensional model and relationships with team effectiveness criteria. *Pers. Psychol.* 61, 273–307. doi: 10.1111/j.1744-6570.2008.00114.x

McDowell, T., Agarwal, D., Miller, D., Okamoto, T., and Page, T. (2016). "Organizational design: The rise of teams," in *Global Human Capital Trends 2016, Deloitte Insights*.

McGinn, C., and Torre, I. (2019). "Can you Tell the Robot by the Voice? An Exploratory Study on the Role of Voice in the Perception of Robots," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. doi: 10.1109/HRI.2019.8673305

Murphy, S. E., and Ensher, E. A. (2008). A qualitative analysis of charismatic leadership in creative teams: the case of television directors. *Leadership Quarterly* 19, 335–352. doi: 10.1016/j.leaqua.2008.03.006

Nass, C., and Brave, S. (2005). *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. Cambridge, MA: MIT Press.

Niebuhr, O. (2021). "Advancing Higher-Education Practice by Analyzing and Training Students' Vocal Charisma: Evidence from a Danish Field Study," in *7th International Conference on Higher Education Advances, Spain*. p. 743–751. doi: 10.4995/HEAd21.2021.12827

Niebuhr, O., Böck, R., and Allen, J. R. (2021). "On the Sound of Successful Meetings: How Speech Prosody Predicts Meeting Performance," in *ACM Proc. 23rd International Conference on Multimodal Interaction* (Montreal, Canada), p. 1–10. doi: 10.1145/3461615.3485412

Niebuhr, O., Brem, A., Novák-Tót, E., and Voße, J. (2016). "Charisma in business speeches: A contrastive acoustic-prosodic analysis of Steve Jobs and Mark Zuckerberg," in J. Barnes, A. Brugos, S. Shattuck-Hufnagel, and N. Veilleux (Eds.) *Proceedings of the 8th International Conference of Speech Prosody* (Speech Prosody Special Interest Group. Speech Prosody), p. 79.

Niebuhr, O., and Michalsky, J. (2019). "Computer-generated speaker charisma and its effects on human actions in a car-navigation system experiment—or how steve jobs' tone of voice can take you anywhere," in *International Conference on Computational Science and its Applications* (Cham: Springer), p. 375–390. doi: 10.1007/978-3-030-24296-1_31

Paulus, P. B., Dzindolet, M., and Kohn, N. W. (2012). Collaborative creativity—group creativity and team innovation. *Handbook Organiz. Creativity* 327–357. doi: 10.1016/B978-0-12-374714-3.00014-8

Puranam, P. (2018). *The Microstructure of Organizations.* Oxford, England: Oxford University Press. doi: 10.1093/oso/9780199672363.001.0001

Rosenberg, A., and Hirschberg, J. (2009). Charisma perception from text and speech. *Speech Commun.* 51, 640–655. doi: 10.1016/j.specom.2008.11.001

Rosenberg-Kima, R. B., Koren, Y., and Gordon, G. (2020). Robot-supported collaborative learning (rscl): social robots as teaching assistants for higher education small group facilitation. *Front. Robot.* 6, 148. doi: 10.3389/frobt.2019.00148

Salanova, M., Llorens, S., Cifre, E., Martínez, I. M., and Schaufeli, W. B. (2003). Perceived collective efficacy, subjective well-being and task performance among electronic work groups. *Small Group Res.* 34, 43–73. doi: 10.1177/1046496402239577

Salas, E., Cooke, N. J., and Rosen, M. A. (2008). On teams, teamwork, and team performance: discoveries and developments. *Human Factors J. Hum. Factors Ergonomics Soc.* 50, 540–547. doi: 10.1518/001872008X288457

Salas, E., Dickinson, T. L., Converse, S. A., and Tannenbaum, S. I. (1992). "Toward an understanding of team performance and training," in: R.W. Swezey and E. Salas, eds. *Teams: Their training and performance* (Norwood, NJ: Ablex), p. 3–29.

Sebo, S., Stoll, B., Scassellati, B., and Jung, M. F. (2020). Robots in groups and teams: a literature review. *Proceed. ACM Hum. Computer Interact.* 4, 1–36. doi: 10.1145/3415247

Seibt, J., Vestergaard, C., and Damholdt, M. F. (2020). Sociomorphing, not anthropomorphizing: towards a typology of experienced sociality. *Front. Artif. Intell. Appl.* 335, 51–67. doi: 10.3233/FAIA200900

Shalley, C. E., Zhou, J., and Oldham, G. R. (2004). The effects of personal and contextual characteristics on creativity: where should we go from here? *J. Manage.* 30, 933–958. doi: 10.1016/j.jm.2004.06.007

Shin, Y., Kim, M., and Lee, S-. H. (2016). Reflection toward creativity: team reflexivity as a linking mechanism between team goal orientation and team creative performance. *J. Bus. Psychol.* 32, 655–671. doi: 10.1007/s10869-016-9462-9

Sørensen, L. S. (2013). *How to grow an apple: Did Steve Jobs speak Apple to success?-An analysis of Steve Jobs' rhetorical and linguistic development in relation to Apple's organizational performance*. Masterthesis. Aalborg University.

Tennent, H., Shen, S., and Jung, M. (2019). "Micbot: a peripheral robotic object to shape conversational dynamics and team performance," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. doi: 10.1109/HRI.2019.8673013

Velner, E., Boersma, P. P., and de Graaf, M. M. (2020). "Intonation in robot speech: does it work the same as with people?," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (p. 569–578). doi: 10.1145/3319502.3374801

Westlund, J. K., Gordon, G., Spaulding, S., Lee, J. J., Plummer, L., Martinez, M., et al. (2015). *Learning a second language with a socially assistive robot*. Almere, The Netherlands.

Yan, B., Lewis, K., Figge, P., Hollingshead, A., Alexander, K. S., Kim, Y. J., et al. (2020). Intelligent Machines and Teamwork: Help or Hindrance? in *Proceedings of the Academy of Management Annual Meeting*. doi: 10.5465/AMBPP.2020.140

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 27 July 2023
DOI 10.3389/fcomm.2023.1114447

Check for updates

OPEN ACCESS

EDITED BY
Oliver Niebuhr,
University of Southern Denmark, Denmark

REVIEWED BY
Isabella Poggi,
Roma Tre University, Italy
Wen Liu,
Shandong University, China

*CORRESPONDENCE
Sandra Madureira
✉ sandra.madureira.liaac@gmail.com

RECEIVED 02 December 2022
ACCEPTED 22 June 2023
PUBLISHED 27 July 2023

CITATION
Madureira S and Fontes MAS (2023) Multimodal impressions of voice quality settings: the role of vocal and visual symbolism.
Front. Commun. 8:1114447.
doi: 10.3389/fcomm.2023.1114447

Multimodal impressions of voice quality settings: the role of vocal and visual symbolism

Sandra Madureira* and Mario A. S. Fontes

Philosophy and Language Sciences Department, Pontifical Catholic University of São Paulo, São Paulo, Brazil

This study considers instances of voice quality settings under a sound-symbolic and synesthetic perspective, focusing on the auditory impressions these settings might have on listeners' attributions of meaning effects and associations between vocal and visual features related to emotional expression. Three perceptual experiments are carried out. The first experiment examined the impressionistic effects of eight voice quality settings characterized by differences in pitch. The second experiment examined the impressionistic effects of seven voice quality settings characterized by productions with the presence or absence of turbulent airflow, irregularity, and tenseness. The third experiment investigated associations between facial expressions of basic emotions and voice quality characteristics. Data are considered in terms of acoustic (fundamental frequency values), articulatory (reduced or expanded length of the vocal tract), perceptual impressions of size (big/small), strength (strong/weak), brightness (dark, clear), and distinctiveness (muffled/distinct), and visual features (facial expressions of the basic emotions sadness, happiness, anger, disgust, fear, and neutrality). The results provide corroborating evidence of existing links between sound and meaning and are discussed in relation to the frequency, production, sirenic biological codes, phonetic metaphors, and the vocal and facial gestures involved in emotional expression.

KEYWORDS

speech expressivity, vocal and visual prosodies, multimodal analysis, voice quality, perceptual analysis speech expressivity, perceptual analysis

1. Introduction

Speech is a powerful means of communication due to its expressive phonetic quality. Issues related to the expressive nature of phonetic quality are addressed by Laver (1976), who argues that phonetic quality can be considered from a semiotic point of view as part of a more comprehensive human communicative system. Phonetic quality, considered in this way, transcends the strict linguistic functions of information meaning and includes the paralinguistic and indexical information that Laver refers to as the whole physical, psychological, and social profile of the speakers, as revealed by the quasi-permanent auditory coloring of their voice quality.

Laver emphasizes that the concurrent features of voice quality, those that are under the speaker's volitional control, together with those due to physiologically intrinsic, non-controlled voice characteristics can be abstracted from the speech flow.

The semiotic perspective on phonetics that Laver advances not only sheds light on speaker identity construction in human communication but also raises issues related to the paralinguistic sound symbolic uses of the long-term phonetic characteristics of voice quality settings.

Sound symbolism reflects direct links between sound and meaning. It is referred to as a pre-semantic phenomenon that evolved from distant preverbal stages in language evolution (Fónagy, 2001; Westbury, 2005; Sučević et al., 2013). It also reflects sound synesthetic interactions between the human senses. Research works on crossmodal associations between acoustic and color features (Johansson et al., 2020), faces and voices (Nagrani et al., 2018), taste and touch (Christensen, 1980), taste and smell (Stevenson and Boakes, 2004), emotion and color, colors and taste, and colors and smell (Gilbert et al., 2016) have demonstrated crossmodal interactions.

Interactions between vocal and visual features in speech perception are demonstrated by the McGurk effect (McGurk and MacDonald, 1976; Green and Norrix, 1997; Abelin, 2007, 2008; Jiang and Bernstein, 2011). Furthermore, gesture congruence is found in speech segments, setting productions, and communicative gestures. For instance, the/i/vowel sound, the lip spreading voice quality setting, and the smiling gesture indicative of positive emotional valence share the same facial movement.

In considering sound symbolism, reference to the dialogue "*Cratylus*" by Plato is inevitable (Vieira, 2014) for discussions of the natural and arbitrary links between sound and meaning. The two protagonists in the dialogue Cratylus defend different points of view on the nature of the expression and content of verbal signs. The protagonist Cratylus argues that the links between form and meaning are natural (the *physei* theoretical view of language). The other protagonist Hermogenes, however, advocates that these links are arbitrary (the *thesei* theoretical view of language).

Over the centuries the debate continued, and in the 20th century, it reappeared in a new guise through Jakobson (1977). Jakobson considers the linguistic sign as motivated in opposition to the arbitrary linguistic sign defended by Saussure [1916] (2012). The *physei-thesei* controversy raised such an interest that it became the subject of a great number of reviews and theoretical propositions in the literature. The discussion evolved in the direction of discussing the specific role of the two types of sound-meaning associations co-existing in language expression (Ahlner and Zlatev, 2010; Monaghan et al., 2014; Dingemanse et al., 2015) and, more recently, to hybridization between arbitrariness and motivation (Nobile, 2019).

Seemingly, the *physei-thesei* is an endless topic of discussion because it can be discussed from a myriad of dynamic factors such as pre-verbal and verbal communication, ritualization, gesturality, systematicity, multimodality, digital orality, expressivity, cognitive functioning, neurophysiological correlations, magical power, and esthetic values.

Fónagy (2001) raises some relevant issues concerning the arbitrary and motivated views of language expressions. He defends that signs are conventional, whether they are motivated (iconic) or not. Furthermore, he states that conventional is not to be confused with arbitrary and that different degrees of motivation (iconicity) must be considered. Arbitrariness and motivation (iconicity) seem to be placed on extremes of a gradient scale.

According to Fónagy's view, iconicity is defined as a natural tie between verbal representation and the reality represented, sound symbolism as the iconicity of sound shape, and synesthesia as the transfer mode of perception to others. These are the definitions we adopt in this study.

Sound-symbolic uses of speech segments and voice dynamic features have been explored theoretically in numerous works in linguistic literature. The following research works are highlighted here for their pioneering contributions (Köhler, 1929; Sapir, 1929; Léon, 1933), their proposals on the relevance of the communicative basis of sound (Jakobson and Waugh, 1979; Fónagy, 1983, 2001; Tsur, 1992; Poyatos, 1993; Hinton et al., 1994), or the contemporaneity of their interpretation of the sound-meaning interactions (Nobile, 2019; Körner and Rummer, 2022).

Experimental research (Köhler, 1929; Newman, 1933; Peterfalvi, 1965; Woodworth, 1991; Abelin, 1999; Blasi et al., 2016; Anikin et al., 2021; just to mention a few) on sound symbolism has yielded surprising results on how sound and sense are interrelated. However, the sound-symbolic impressions caused by the coloring of voice quality settings have not been considered in such experiments.

The relevance of the role of voice quality in signaling affect has been demonstrated in an experiment with synthesized stimuli (Yanushevskaya et al., 2013) that compared voice quality and fundamental frequency effects in expressing affective stages and concluded that F0 alone had a weaker effect than voice quality alone.

In fact, as voice quality is a long-term articulatory or phonatory configuration that "colors" all speech segments (Laver, 1980). The "coloring" effect is more perceptually evident on the key speech segments (Mackenzie Beck, 2005) that is, the segments which are more susceptible to the influence of determined voice quality settings. In this way, for instance, oral speech sounds are more susceptible to nasal vocal quality settings than nasal speech sounds, front unrounded vowels to lip rounding voice quality settings and back rounded vowels to lip spreading voice quality settings (Mackenzie Beck, 2005, 2007).

Voice quality settings are defined in this study according to the phonetic model of voice quality description (Laver, 1980), and their description follows the Voice Profile Analysis System (VPA) developed by Laver and Mackenzie-Beck (2007). Facial movements are described with the Facial Action System (FACS) developed by Ekman et al. (2002).

Similarities between the two systems, VPA and FACS, are described by Madureira and Fontes (2019). Both systems were created in the late 1970s and revised in the early 2000s. They are both perceptually oriented, theoretically based, and componential, and their components can be combined into profiles. The analytical unit in FACS is the Action Unity (AU) and that in VPA is the setting. Both systems posit a neutral element as a reference—the neutral face in FACS and the neutral voice quality setting in VPA. The intensity of the facial movement is analyzed in terms of five varying intensity levels (from A to E in FACS), and the settings of voice quality in terms of six discrete degrees (from the weakest 1 to the strongest six in VPA).

Interactions between sound and meaning and among vocal, visual, and meaning features are examined under a theoretical view of language iconicity. Some founding tenets of sound symbolism and synesthesia perception, the role of vocal and visual features in meaning expression, and three perceptual experiments on voice quality expressiveness are considered in this study.

2. The paralinguistic voice and face of speech

Facial and vocal gestures are highly communicative (Xu et al., 2013), and they are fully integrated with vocal gestures in the expression of emotions of attitudes and modality. The 7th cranial nerve, which innervates the facial muscles, is also connected to the brain motor areas responsible for speech production (Walker, 1990).

On the speech acoustic signal produced by vocal gestures, the richness and variability of the physical properties of the acoustic features afford listeners means for attributing paralinguistic meanings. Acoustic cues impress listeners with their inherent physical features and are interpreted as indices of psychological, sociological, and biological factors and/or as representations of mental objects.

For the investigation of acoustic parameters in analyzing paralinguistic meanings, the ExpressionEvaluator script (Barbosa, 2009) and the Prosody Descriptor Extractor for PRAAT (Barbosa, 2021) allow the measurement of numerous acoustic parameters shedding light on how listeners perceive such meaning effects.

A key point regarding the attribution of paralinguistic meanings by listeners is that acoustic factors map articulatory factors, and both are considered indispensable (Kawahara, 2021) in explaining sound-meaning correspondences and their patterning in languages.

On the actions concerning facial gesturality, the muscular movements of the superior and inferior parts of the face and their combinations play an important role in listener attribution of paralinguistic meanings (Swerts and Krahmer, 2006; Scherer et al., 2021). Some of the movements of the upper face are involuntary and last from 40 to 200 ms. They are called microexpressions (Ekman and Friesen, 1976).

The prevalence of vocal or facial expression in emotion perception has been debated and research findings indicate that such perception depends on the kind of stimuli used, the kind of emotion analyzed, the level of emotional intensity considered, and the interference of cultural aspects focused in the experiments.

On issues related to emotion research, a comprehensive theoretical and empirical review is found in Scherer et al. (2011) and a proposal for future research agenda on emotion studies in Scherer (2022). For accomplishing the proposed goal in the emotion research agenda, Scherer argues that central concepts, components, mechanisms, and operationalization procedures of the emotional phenomena must be identified and clarified.

In Scherer et al. (2011), emotional expression and impression are viewed as determined by psychobiological and sociocultural factors. Emotional expression is considered from a multimodal perspective which comprises vocal, facial, and gestural patterns of expression. Emotional expression is considered from a multimodal perspective, which comprises vocal, facial, and gestural patterns of expression. The authors propose that emotions are encoded as signals in the face, the voice, and the body and decoded by the perceivers who rely on the multimodal expressed cues.

Scherer (2022) analyzes the tenets of the basic, appraisal, and constructivist theories and focus on their complementary and shared features. Based on these commonalities, the author suggests steps to be taken toward the proposal of a convergent theoretical framework of emotional expression. The author considers the emotional process, regardless of their theoretical orientation, to arise in response to a perceived event, which is evaluated and creates a physiological reaction. Physiological reactions are then expressed in vocal, facial, and bodily actions and can be categorized and labeled.

Abelin (2007, 2008) investigated the perception of emotional expressions in unimodal and bimodal conditions using audio only, visual only, audiovisual, and conflicting audio and visual stimuli (McGurgh stimuli). Results indicated that emotional expressions were better perceived in bimodal conditions. In unimodal conditions, the auditive expression performed better, and under the McGurk condition, it was the visual expression. Depending also on the kind of emotion expressed, the vocal or the visual channel was found to perform better. The latter is consistent with Fontes (2014) findings.

In an empirical study using neuroimaging and an affective priming task to study facial and vocal expressions of anger, happiness, and neutrality, Zhang et al. (2018) found that facial expressions played a more important role than vocal prosody in eliciting emotion perception. These findings can be interpreted as related to the kinds of emotions analyzed, as highlighted by Abelin (2007, 2008).

Based on some research implications of the above-reported studies, multimodality and mappings among acoustic, articulatory, and perceptual data are key factors to be considered in experiments on speech expression paralinguistic meanings.

3. Experiments

We designed and performed three experiments to examine the impressionistic effects of voice quality settings. These experiments focused on the associations among vocal, facial, and meaning expressions. These associations are interpreted in relation to the frequency, effort, and sirenic biological codes and phonetic metaphors.

The speech samples used in the three experiments reported in this study were extracted from the audio files accompanying the book "*The Phonetic Model of Voice Quality*" by Laver (1980), and all the voice quality settings were produced by John Laver himself. The utterance is "Learning to speak well is an important and fruitful task." The description of the voice quality settings follows Mackenzie Beck (2007).

The choice of the speech samples was motivated by the fact that the speaker is the same, so physiological features are controlled by the speaker's phonetic expertise and ability to produce several voice quality settings. All the settings were produced with a moderate degree.

The choice of the settings in each experiment was guided by the perceptual effects related to the Frequency Code (Ohala, 1984), the Effort Code (Gussenhoven, 2002, 2004), the Sirenic Code (Gussenhoven, 2016), and the phonetic metaphor (Fónagy, 1983, 2001).

The Frequency Code was proposed on ethological grounds, following Morton (1977) observations on vocal communication signals of birds and mammals. It relates animal size to F0 acoustic characteristics. From a sound-symbolic perspective, a small vocal tract and thin vocal folds produce a high F0 and a small size signal, non-threatening attitudes, fragility, submissiveness, and related meanings, whereas a low F_0 conveys large size, threatening attitudes, power, assertiveness, and related meanings.

The Effort Code pertains to the articulatory effort and precision displayed in the pitch range and articulatory movements. The greater the articulatory effort, the greater the tendency toward articulatory precision and wider pitch ranges. The paralinguistic meanings conveyed by speech productions characterized by higher effort are emphasis, greater significance, insistence, and surprise (Chen et al., 2002). The opposite might be expected from the reduced effort.

The Sirenic Code (Gussenhoven, 2016) refers to the use of breathy, whispery phonation to signal femininity, attractiveness, charm, and related paralinguistic meanings. In female participants, this kind of phonation, which is produced with airflow escaping through the glottis, produces lower harmonics-to-noise ratios. In an analysis of dubbing voices, Crochiquia et al. (2020) found that one of the male characters whose voice was characterized by a whispery voice quality setting was judged pleasant and his personality was considered charming.

Phonetic metaphors (Fónagy, 1983, 2001) are defined as verbal mirror images of the mental movement inherent in the phonetic gesture. Phonetic gestures are metaphorical because articulatory actions can be interpreted as expressions of meaning.

The perception test in each experiment was administered to a different group of subjects, that is, each subject judged the stimuli of just one of the perceptual tests. In all three tests, the number of female judges was greater than the number of male subjects. As such, the results may be more representative of women's perceptions.

3.1. Experiment 1

3.1.1. Aim

This experiment aims to examine the impressionistic effect of eight voice quality settings. Paralinguistically, meaning associations related to the Frequency Code and phonetic metaphors are expected.

3.1.2. Participants

The test was performed by 44 participants, 8 men and 36 women, aged from 18 to 75 years, with a mean age of 38 years. They were undergraduates and graduates from several fields (Linguistic, Languages, Speech Therapy, Psychology, and Communication). None of them had hearing problems.

3.1.3. Stimuli

The stimuli were eight speech samples characterized by the following voice quality settings as referred to VPA (Laver and Mackenzie-Beck, 2007): Raised Larynx, Lowered Larynx, Lip Rounding, Backed Tongue Body, Lip Spreading, Nasal, Denasal, and Falsetto.

Perceptually, Backed Tongue Body, Lip Rounding, Denasal, Nasal, and Lowered Larynx are characterized by lower pitch mean and range than Falsetto, Raised Larynx, and Lip Spreading.

The perceptual effects of these voice quality settings are related to the acoustic resonance characteristics determined by the size and shape of the resonating cavity: diminished in the cases of Raised Larynx and Lip Spreading, enlarged in the cases of Lowered Larynx and Backed Tongue Body, and involving coupled resonating cavities for Denasal and Nasal (Vanger et al., 1998). In the case of Falsetto, which is produced with stretched vocal folds, the resultant phonation is high-pitched.

In the selected stimuli, the F0 maximum value and difference between F0 maximum and F0 minimum values were for the Backed Tongue Body (103 Hz; 36 Hz), for Lip Rounding (115 Hz; 47 Hz), for Denasal (152 Hz; 70 Hz), for Nasal (166 Hz; 84 Hz), for Lowered Larynx (138 Hz; 58 Hz), for Falsetto (328 Hz; 124 Hz), for Raised Larynx (230 Hz; 143 Hz), and for Lip Spreading (180 Hz, 114 Hz).

3.1.4. Perceptual test design and application procedures

The perceptual test was designed with the SurveyMonkey online survey software and a link to be sent to participants was generated. The speech samples were followed by a sliding bar, containing four pairs of polar semantic descriptors (opposing darkness/clarity, muffledness/distinctiveness, smallness/bigness, and strength/weakness) displayed on a semantic differential scale.

The participants were asked to listen to the stimuli and register their auditory impression by placing the mouse pointer on some part of the continuous scale coded between 0 and 100. Scores lower than 50 show a tendency toward the polar descriptor on the left and scores higher than 50 to the polar descriptor on the right. A score of 50 corresponds to the middle (neutral) point.

The participants' answers were collected and transferred to an Excel sheet. Mean values of the perceptual scores for the voice quality settings were calculated. To evaluate inter-rater reliability, Cronbach's Alpha test was applied. The Pareto Chart histogram was used to identify the probability distribution function of the data.

3.1.5. Results

An acceptable value of Cronbach's Alpha was obtained (0.78). In Table 1, vocal quality settings and paralinguistic features are related; higher scores indicate the choice of the rightmost descriptor in the pair and lower scores indicate those placed on the left. In the description of the vocal quality main characteristics, descriptors whose scores fell within the range between 49 and 51 were considered neutral. Thus, the lower, or the higher the score is, the more representative it is of the characteristic with which it is related.

The probability distribution function of the data was determined with a Pareto Chart histogram to check the tendency of the distribution of 80% of the answers toward the descriptor on the left or to the one on the right of the continuous scale. In this way, it was possible to identify the voice quality settings whose judgment scores in relation to the semantic descriptors were 80% placed in the range between 0 and 40 (left side of the scale) and in the range between 60 and 100 (right side of the scale). They were as follows:

TABLE 1 Perceptual scores of associations between voice quality settings and paralinguistic meanings.

Settings	Dark/Clear	Muffled/Distinct	Big/Small	Strong/Weak	Main characteristics
Raised larynx	64.36	52.45	56.52	62.80	Clear/weak/small/distinct
Lowered larynx	44.59	58.39	41.50	36.57	Strong/big/dark/distinct
Lip rounding	36.57	57.18	62.70	47.75	Dark/ big/muffled/strong
Lip spreading	72.20	70.30	52.17	50.43	Clear/distinct/small
Falsetto	69.77	53.11	73.55	81.91	Weak/small/clear/distinct
Nasal voice	56.11	64.30	48.70	43.20	muffled/dark/ strong
Backed tongue body	30.52	37.00	38.84	37.00	Dark/strong/muffled/big
Denasal	49.09	37.95	51.55	49.41	Muffled

for the descriptor "dark", Backed Tongue Body, Lowered Larynx, and Nasal; for "clear", Falsetto, Lip Spreading and Raised Larynx; for "muffled", Backed Tongue Body and Denasal; for "distinct", Falsetto; for "big", Backed Tongue Body and Lowered Larynx; for "small", Falsetto; for "strong", Backed Tongue Body and Lowered Larynx; and for "weak", Falsetto and Raised Larynx.

3.1.6. Discussion

The results of the perceptual judgments on the voice quality settings presented in Experiment I show two opposing groups in terms of strength (weak/strong), size (small/big), and brightness (clear/dark). Images extracted from Vanger et al. (1998) and analyzed with FaceReader8. One of the groups is composed of Falsetto, Raised Larynx, and Lip Spreading voice quality settings, which were judged as clear and small. Falsetto and Raised Larynx are also considered weak. The other group is composed of Lowered Larynx and Backed Tongue Body voice quality settings, which were judged as strong, big, and dark.

Raised Larynx and Lip Spreading voice quality settings are produced with a shortened vocal tract, and shorter vocal tracts tend to increase f0 and formant frequencies. Falsetto voice quality settings are characterized by even higher F0 values than Raised Larynx and Lip spreading vocal quality settings because they are produced with stretched vocal folds, which makes the vocal folds thinner, and when they vibrate, they barely touch each other (Mackenzie Beck, 2007).

Lip Rounding and Backed Tongue Body vocal quality settings, on the other hand, are produced with an increased size of the vocal tract, the former by adding an extra cavity formed by the protrusion and rounding of the lips, and the latter by enlarging the oral cavity. Both have the acoustic consequence of lowering frequencies (Fant, 1960; Stevens, 1998).

Together with the Denasal voice quality setting, the Backed Tongue Body vocal quality setting was also judged as "muffled". The Backed Tongue Body tends to centralize front vowels and make velarization, uvularization, and pharyngealization features more marked. Tongue backing has a great effect on front vowels, which are realized as central vowels (Mackenzie Beck, 2007). A front vowel such as/i/when realized as a close central unrounded vowel sounds less distinct. Laver (1980) refers to this centralizing effect as the centering setting found in voices is perceived as muffled.

According to Kawahara (2021), the fact that Nasal voice quality was perceptually associated with strong and dark. Kawara reports that the reason why nasals may be associated with large images and paralinguistic meaning expressions of roundness, heaviness, and softness may be linked to the length of nasal cavities. Acoustically, nasal sounds are characterized by low-frequency energy and a damped auditory quality caused by resonances produced by the long resonating cavity ending in small apertures.

The perceptual judgment results in this experiment corroborate sound-symbolic relations based on the Frequency Code (Ohala, 1984, 1994) and on the phonetic metaphors (Fónagy, 1983). Ohala (1994) argues that high F0 signals smallness, a non-threatening attitude, and other related semantic meanings, whereas a low F0 conveys opposite meanings.

In our experiment, judgments of the Falsetto voice quality setting clearly demonstrate that the higher the F0, the weaker, the smaller, and the clearer the voice quality setting. This result corroborates the Frequency Code predictions on the meanings associated with high frequency. Metaphorically, on the one hand, links can be established between the diminished length of the vocal tract in Raised Larynx vocal quality setting productions and smallness and weakness, and on the other hand, between the expanded length of the vocal tract in Lowered Larynx vocal quality productions or yet of the expanded oral cavity in Backed Tongue Body vocal quality productions and bigness and strength.

3.2. Experiment 2

3.2.1. Aim

This experiment aimed to examine the impressionistic effect of seven voice quality settings. The perceptual effects of these voice quality effects are related to the presence or absence of turbulent airflow and irregularity and the presence or absence of tenseness. Paralinguistically, meaning associations related to the Effort and Sirenic Codes are expected.

3.2.3. Participants

The test was performed by 50 participants, 17 men and 33 women, aged from 21 to 70 years, and with a mean age of

TABLE 2 Perceptual scores related to associations between voice quality settings and paralinguistic features.

Settings	Softness	Regularity	Pleasantness
Modal	76	80	73
Breathy	75	70	66
Tense	72	76	66
Lax	75	70	69

TABLE 3 Perceptual scores related to associations between voice quality settings and paralinguistic features.

Settings	Harshness	Irregularity	Unpleasantness
Whispery	80	59	75
Creaky	67	60	66
Harsh	91	71	83

38 years. They were undergraduates and graduates from several fields (Linguistic, Languages, Speech Therapy, Psychology, and Communication). None of them had hearing problems.

3.2.4. Stimuli

The seven voice quality settings selected comprise phonatory settings (Modal Voice, Whispery Voice, Creaky Voice, Breathy Voice, and Harsh Voice) and Vocal Tract settings (Tense Voice and Lax voice) as referred to by VPA (Laver and Mackenzie-Beck, 2007).

3.2.5. Perceptual test design and application procedures

The perceptual test was designed in the SurveyMonkey online survey software and a link to be sent to participants was generated. The speech samples were followed by a sliding bar, containing three pairs of semantic descriptors (softness/roughness; regularity/irregularity; and pleasantness/unpleasantness) displayed in a differential scale.

The participants were asked to listen to the stimuli and register their auditory impression by placing the mouse pointer at the perceived value in the continuous scale. Their answers were collected and transferred to an Excel sheet. The mean values of the perceptual scores for the voice quality settings were calculated. To evaluate inter-rater reliability, Cronbach's Alpha test was performed.

3.2.6. Results

The value of Cronbach's Alpha was acceptable (0.72). Modal, Breathy, and Lax voice quality settings, which are not produced with larynx muscle tension or pharyngeal constriction, were considered soft, regular, and pleasant. Although the tense voice quality setting is produced with constricted glottis, the F0 is not irregular (Keating et al., 2015). This might explain the affiliation of this voice quality setting with this group. The mean perceptual scores are given in Table 2.

Whispery, Creaky, and Harsh voice quality settings were considered harsh, irregular, and unpleasant, respectively. The mean perceptual scores obtained are given in Table 3.

3.2.7. Discussion

Modal and Harsh voice qualities were placed on opposite extremes of the continuum between positive (pleasant) and negative (unpleasant). Our interpretation is that this discrepancy reflects the periodicity resulting from glottal efficiency in the Modal voice quality setting production and the aperiodicity of the Harsh voice quality setting (Mackenzie Beck, 2007), due to irregular vocal fold vibration.

Whispery voice combines fricative glottal airflow with vocal fold vibration, Creaky voice combines low-frequency pulsed phonation with vocal fold vibration, and Harsh voice quality setting combines noise with irregular vocal fold vibration. Although Breathy and Whispery voice quality settings share high levels of fricative airflow through the glottis, Breathy voice is produced with a lower amplitude of vocal fold vibration and less fricative energy (Hewlett and Beck, 2013) than Whispery voice quality and with lax phonation as opposed to tense phonation in Whispery voice quality (Schaeffler et al., 2019). Breathy voice quality is considered a signal of intimacy, whereas Whispery voice quality is interpreted as a signal of confidentiality (Laver, 1980).

The results can be interpreted in reference to remarks on the opposing flowing characteristics of periodicity and disturbing characteristics of aperiodicity (Tsur, 1992; Fónagy, 2001). Opposite sound-meaning relationships can be derived from chaotic associations among the noise, irregular patterns of vibration of the vocal folds, tense phonation, laryngeal and pharyngeal constrictions, and smooth associations between sonority, a regular pattern of vibration of the vocal folds, and lax phonation.

The impressionistic meaning effects of fricative glottal airflow can be interpreted in terms of the Sirenic Code and the presence or absence of articulatory effort and the amount of energy expenditure can be interpreted in terms of the Effort Code. Under low-intensity levels, as in breathy voice quality, the airflow is smooth and pleasant, but as fricative energy, tenseness, and articulatory effort increase as in Harsh or Whispery voice qualities, it becomes unpleasant.

3.3. Experiment 3

3.3.1. Aim

This experiment aimed to investigate potential associations between facial and vocal expressions of basic emotions. Our interest is in the synesthetic interactions among Action Unities, vocal quality settings, and emotional meaning expression. Action Unities describe the movements of the facial muscles, and vocal quality settings describe the movements of the articulators and the vocal folds. The impressionistic effects of both facial and vocal expressions on emotion detection can vary depending on production features such as the upward or downward direction of the movement and the presence or absence of tenseness and constriction.

3.3.2. Participants

The test was performed by 50 participants, 13 men and 37 women, aged between 18 and 73 years, with an average age of 35 years. They were undergraduates and graduates from several fields (Linguistics, Languages, Speech Therapy, Psychology, and Communication). None of them had hearing problems.

3.3.3. Stimuli

There were two kinds of stimuli: visual and vocal. The visual stimuli were pictures of a man portraying the six basic emotions that were selected for analysis: Happiness, Sadness, Anger, Disgust, Fear, and Neutrality. Such pictures were drawn by Vanger et al. (1998). Apart from the neutral expression, which was not investigated in Ekman (2016), his findings show that there was high agreement judgment on the other five emotions considered in this experiment: Anger (91%), Fear (90%), Disgust (86%), Sadness (80%), and Happiness (76%). Neutrality was included as a facial stimulus to investigate the potential relationship of this kind of facial expression with the neutral setting of voice quality.

The vocal stimuli comprised six voice quality settings in VPA (Laver and Mackenzie-Beck, 2007), four of them phonatory configurational settings (Modal Voice, Whisper, Creak, and Harsh Voice) and two of them vocal tract configurational settings (Lowered Larynx and Lip Spreading).

The choice of voice quality settings to be tested was motivated by the potential matchings between visual and vocal characteristics, such as that between neutral face, which is not characterized by muscle contractions and neutral voice quality, described as the voice quality setting produced with regular and efficient glottal fold vibration; between coincidental lip corner movements in the face expression of Happiness and Lip Spreading, which is a voice quality setting produced with stretched lips; between the downward lip corner movements in the facial expression of Sadness and low pitch perception of the Lowered Larynx Voice quality, produced with downward movement of the larynx; between the contracted muscles in the expression of Anger and the perceptual roughness and the irregularity of voice fold vibration characterizing Harsh; between the repulsive feeling expressed by the Disgust facial expression and the unpleasant feeling caused by irregular discrete audible pulses characterizing Creak; and between the conflicting fight-or-flight response to Fear (Cannon, 2016) and the conflicting presence of audible fricative airflow of air stemming from the glottis and the absence of voice characterizing Whisper.

A point must be made in relation to Whisper and Creak. According to Kreiman and Sidtis (2011), Modal voice and Whisper are placed on opposite extremes of the voicing continuum. In Whisper, the vocal folds vibrate only slightly or not at all, and noise is generated through a partially closed glottis. In VPA (Laver and Mackenzie-Beck, 2007), the voice quality setting "Whisper" is distinct from Whispery Voice and Creak from Creaky Voice. The distinctions are based on the predominance of either voicing or noise features characterizing speech production. Wideband spectrograms of speech samples produced with these four kinds of voice quality settings are presented in Figures 1, 2. They were generated in PRAAT (Boersma and Weenik, 2022), version 6.2.18.

In Figure 1, the waveforms and the wideband spectrograms of the first clause of the sentence "Learning to speak well is an important and fruitful task" produced in Whisper and Whispery voice qualities are displayed. The absence of the voicing bar in the Whisper production contrasts with the presence of the voicing bar in the Whispery production.

In Figure 2, the waveforms and the wideband spectrograms of the first clause of the sentence "Learning to speak well is an important and fruitful task" produced in Creak and Creaky voice qualities are displayed. The prevalence of creak over voicing characterizes the Creak voice quality setting.

The pictures depicting the six basic emotions were also analyzed automatically by the software FaceReader, version 8.1, from Noldus. (2022) Technology to determine the AUs involved in the picture facial expressions, their intensities, and their association with affective states. The coding of the AUs is performed in relation to the neutral face of the person under analysis.

3.3.4. Perceptual test design and application procedures

The perceptual test was designed in the SurveyMonkey online survey software and a link to be sent to participants was generated. Photos and speech samples were followed by multiple-choice questions.

The participants were asked to look at face portrait images and choose one out of six alternatives that best described the emotion expressed by the face. Each image was followed by six speech samples of the same sentence produced with six voice quality settings (Neutral, Modal, Creak, Whisper, Harsh, and Breathy). Participants were then asked to choose the speech sample which best matched the facial emotion expression.

The participants' answers were collected and transferred to an Excel sheet. The mean values of the perceptual scores for the face emotion descriptors and voice quality settings were calculated. To evaluate inter-rater reliability, Cronbach's alpha test was applied.

The Principal Component Analysis (PCA) method (Husson et al., 2009) was used to analyze the set of variables. The PCA is a multivariate and multidimensional statistical method. It is applied in three steps: identifying a common structure among the group variables; describing the specificity of each group of variables using correlation analysis; and comparing the resulting values using the individual analyses of the variables.

3.3.5. Results

The facial automatic analysis provided for each facial expression comprised the following features: the AUs, their intensity level, the emotion detected, and their percentage of detection. Intensity levels vary from the weakest (A) to the strongest (E). "A" stands for Trace, "B" for Slight, "C" for Pronounced, "D" for Severe, and "E" for Maximum. Figure 3 presents the AUs identified with the FaceReader and Table 4 specifies all the data obtained using the automatic analysis.

The indices of facial recognition of the emotions depicted in the pictures, except for Fear, were high. The lower percentage for the recognition of Fear may be related to the low intensity (A) of the AU detected.

Table 5 presents the results of the Perceptual test applied to the judges. The features in the table describe the stimuli, the kinds

TABLE 4 Stimuli, action unities (AUs) and their intensities, emotions, and percentages of identification.

Stimuli	AUs	AU description	AU intensity	Emotion detected	% Emotion identification
Picture of a sad facial expression	1	Inner brow raiser	C	Sadness	92%
	4	Brow lowerer	B		
	15	Lip corner depressor	E		
Picture of a happy facial expression	12	Lip corner puller	D	Happiness	98%
	25	Lips part	D		
Picture of an angry facial expression	4	Brow lowerer	C	Anger	87%
	5	Upper lid raiser	B		
	7	Lid tightener	A		
	23	Lip tightener	D		
	24	Lip presser	E		
Picture of a disgusted facial expression	4	Brow lowerer	B	Disgust	99%
	6	Cheek raiser	C		
	7	Lid tightener	B		
	9	Nose wrinkler	C		
	10	Lip corner depressor	D		
	17	Chin raiser	D		
Picture of a scared facial expression	5	Upper lid raiser	A	Neutrality	65
				Fear	33
Picture of a neutral facial expression				Neutrality	99

TABLE 5 Stimuli, emotions, and their percentage of identification, associated voice quality settings, and their percentage of identification.

Stimuli	Emotion	% Emotion identification	Voice quality setting	% voice quality identification
Picture of a sad facial expression and voice quality settings	Sadness	62%	Lowered larynx	58%
			Creak	30%
Picture of a happy facial expression and voice quality settings	Happiness	98%	Lip spreading	54%
			Modal voice	34%
Picture of an angry facial expression and voice quality settings	Anger	64%	Harsh voice	48%
			Creak	18%
			Lowered larynx	16%
Picture of a disgusted facial expression and voice quality setting	Disgust	92%	Creak	34%
			Harsh voice	30%
Picture of a scared facial expression and voice quality settings	Fear	90%	Whisper	78%
Picture of a neutral facial expression and voice quality settings	Neutrality	84%	Lip spreading	44%
			Modal voice	22%

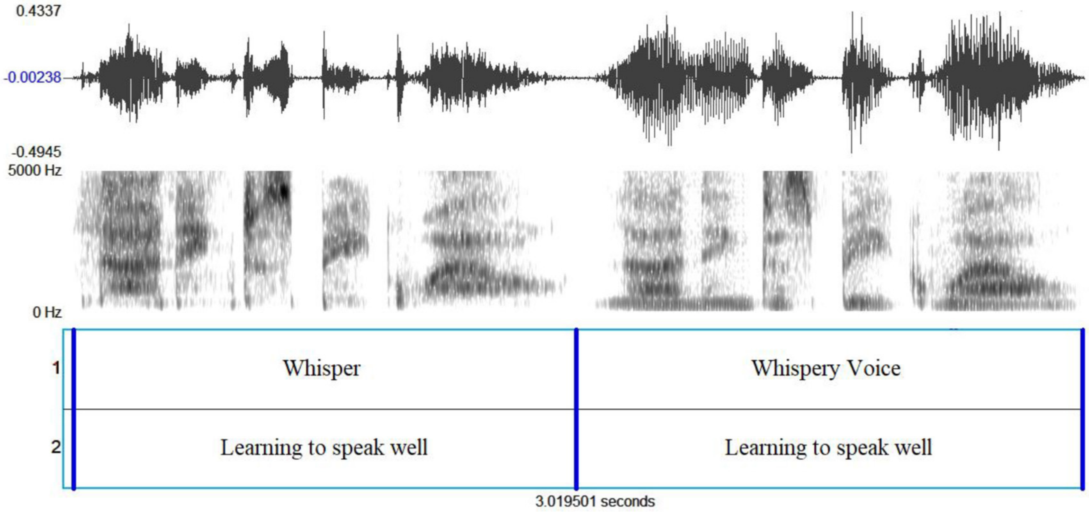

FIGURE 1
From **top** to **bottom**: the waveforms, the wideband spectrograms, and two annotation tiers with information on voice quality settings and textual transcription.

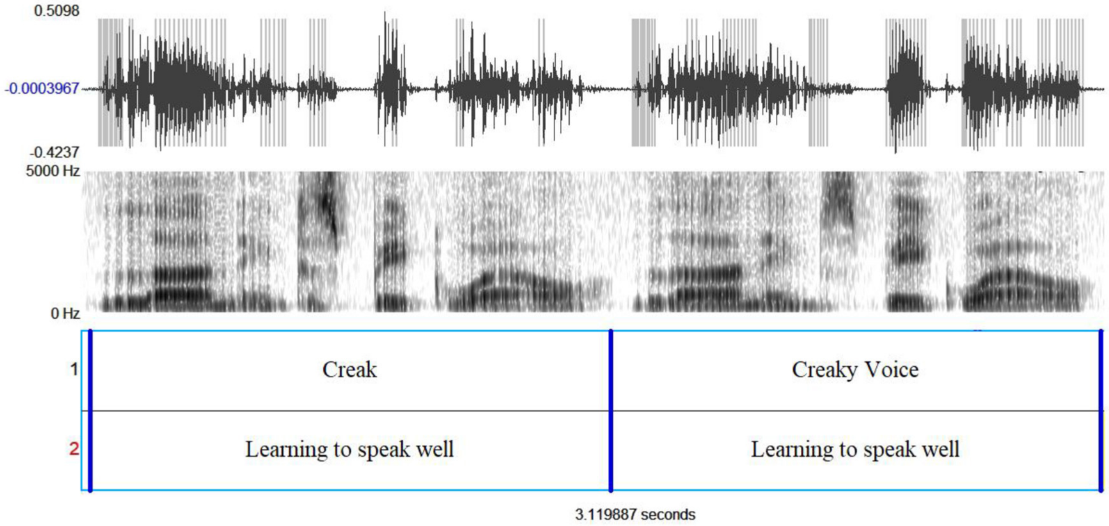

FIGURE 2
From **top** to **bottom**: the waveforms, wideband spectrograms, and two annotation tiers with information on voice quality settings and textual transcription.

and intensity levels of the emotions detected, and the kinds and percentages of identification of the voice quality settings.

The statistical method PCA was applied to analyze the emotion and the voice quality setting variables. Mean scores for both groups of variables were considered. All measures were normalized.

The PCA generated five clusters. One of the clusters grouped Whisper and Creak vocal quality settings which are characterized by voicelessness. Lip Spreading was correlated with Modal Voice, and Fear with Whisper. Creak was inversely correlated with Lip Spreading and Modal Voice.

The inertia gains in the clusters indicated that Dimensions 1 and 2 (Dim and Dim 2) could better explain the data. In Figure 4, the distribution of the variables is shown in four quadrants. On the upper left quadrant are Modal Voice, Lip Spreading, Happiness, and Neutrality; on the upper right quadrant are Lowered Larynx, Creak, and Sadness; on the left lower quadrant are Fear and Whisper; and on the right lower quadrant are Harsh, Anger, and Disgust. Significant correlations are presented in Table 6.

3.3.6. Discussion

The judges were able to identify most facial emotion expressions with high accuracy. Except for the Fear and Whisper

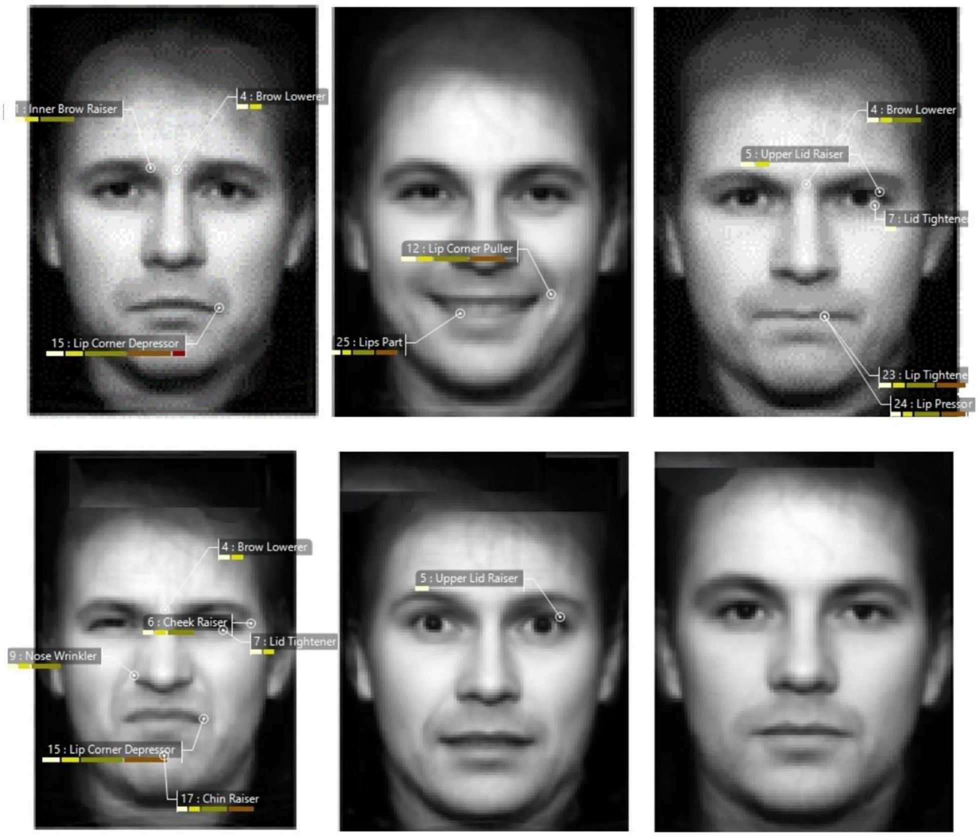

FIGURE 3
From top **left** to **right**: facial expressions of Sadness, Happiness, and Anger. From bottom left to right: facial expressions of Disgust, Fear, and Neutrality.

pairing, weaker associations of voice quality settings with facial expressions of emotions were established.

Despite the lower percentage of matchings between voice quality settings and facial emotion expressions, emotions with negative valence were associated with voice quality settings characterized by irregularity and fricative noise (Harsh and Whisper) or low pitch (Lowered Larynx), while emotions with neutral or positive valence with Lip Spreading.

Analysis of the associations between voice quality settings and facial expressions of emotions showed certain matchings and mismatchings between those two aspects. Lip Spreading and Modal Voice settings were chosen for the facial expression of Happiness; Harsh and Whisper were not chosen for the facial expression of Sadness; Whisper, Lowered Larynx, Creak, and Harsh were not chosen for Happiness; Whisper, Lip Spreading, and Modal Voice were not chosen for Anger; Lip Spreading and Modal were not chosen for Disgust; and Whisper and Harsh Voice were not chosen for Neutrality.

The matches and mismatches among face, emotion, and voice quality settings, as interpreted in accordance with the Emotion Wheel (Scherer, 2005), show primitive emotion-related valence and control features. Happiness is described as exhibiting positive Valence and high Control; Anger and Disgust, negative Valence and high Control; and Sadness and Fear, negative Valence and low Control. Taking these emotional primitives into account, associations between Happiness and Lip Spreading; among Anger, Disgust, and Harsh Voice; between Sadness and Lower Larynx Voice; and between Fear and Whisper are highlighted.

The associations of voicing and higher pitch (Lip Spreading) with Happy facial emotion expressions, of voice irregularity (Harsh Voice) with Anger and Disgust facial emotion expressions, and of voicing and lower pitch with Sadness emotion expressions, and the absence of voicing in Whisper can be considered in terms of the Frequency, Production, and Sirenic codes.

In the recognition of face and vocal expressions of emotions, the direction of the gestures, the regularity and irregularity of patterns, and the presence or absence of constriction were found to be influential factors in these associations.

Concerning vocal expression, the acoustic cues related to the presence of voice source only, voice and noise sources, and noise source only in speech production and their corresponding acoustic outputs and perceptual features related to the mean pitch influenced the listener's judgments.

Concerning facial expression, the presence or absence of muscle contraction and tenseness and the direction of the muscle movement were the influential factors in the attribution of emotions.

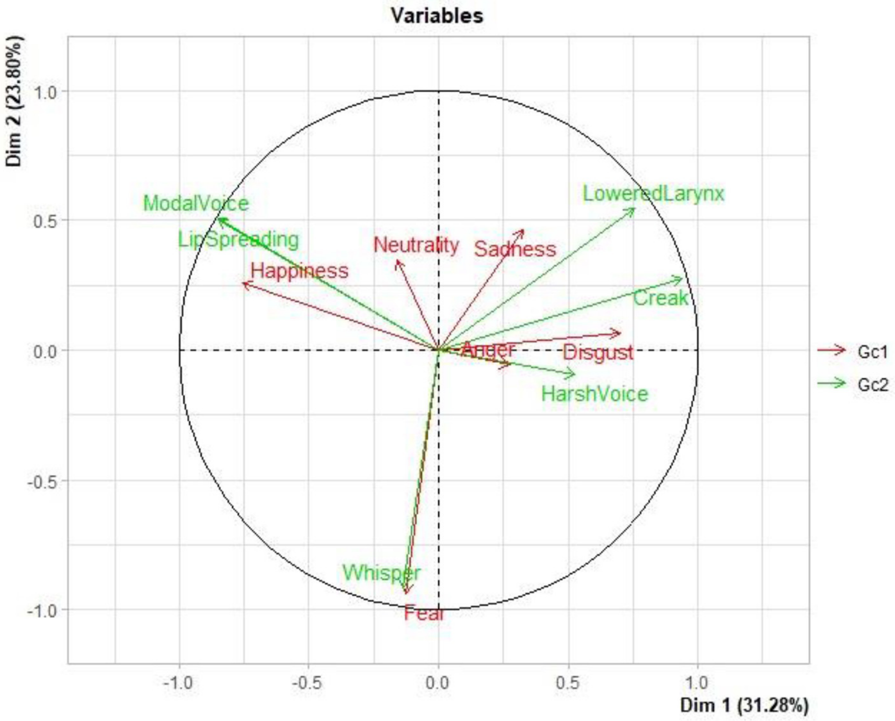

FIGURE 4
Distribution of the variables in dimensions 1 and 2 (Dim 1 and Dim 2).

TABLE 6 Voice quality settings, emotions, correlation, and probability values.

Dim.1	Correlation	P value
Creak	0.9393	0.0054
Lip spreading	−0.8487	0.0326
Modal voice	−0.8519	0.0313
Dim.2	Correlation	P value
Whisper	−0.9183	0.0097
Fear	−0.937	0.0058

Synesthetically, the smiling gesture was heard (Lip Spreading) and seen (Happy face). Iconically, the downward gestures presented in Sad facial expressions and low-pitched voice qualities (Lowered Larynx and Creak) were associated with Sadness. Metaphorically, the perceptually loud rough voice produced by the Harsh voice quality setting and the sour aggressive face displaying contracted and pressed muscles represent the facial-vocal expression of Anger. Also, from a metaphorical point of view, the conflicting absence of voice and the presence of fricative airflow in the Whisper production were associated with the facial expression of fear, echoing the saying “voiceless out of fear.”

4. Conclusion

The experiments in this study considered the expressive role of voice quality settings under a sound-symbolic, synesthetic, and metaphorical perspective, focusing on the auditory impressions these settings might have on listeners’ attributions of meaning effects and associations between vocal and visual features related to emotional expression.

The first experiment examined the impressionistic effects of eight voice quality settings characterized by pitch differences. The opposing auditory impressions of higher vs. lower pitched voice quality settings echoed the premises of the Frequency Code whereas the opposing auditory impressions of small vs. big size reflected metaphorical judgments of acoustic outputs of the articulatory gestures configuring the length of the resonating cavities in voice quality setting productions.

The second experiment examined the impressionistic effect of seven voice quality settings characterized by productions with the presence or absence of turbulent airflow, irregularity, and tenseness, and the results showed a strong iconic effect between sound and meaning. Voice quality settings characterized by highly turbulent airflow, irregularity, and constricted muscles were judged negatively whereas voice quality settings characterized by slightly turbulent flow, regularity, and relaxed muscles were judged positively.

The third experiment investigated associations between facial expressions of basic emotions and voice quality characteristics. Visual cues were more reliable in identifying emotions than auditory cues. Modal and Lip Spreading voice qualities were associated with high control positive valence emotions. Lowered Larynx was associated with low control and negative valence emotions, Creak and Harsh with high control negative valence emotions, and Whisper with low control negative valence.

The results of the present study encourage further research on the expressive uses of vocal quality settings and the interactions between voice and face expressions through the integration of more refined procedures, a larger, and more gender-balanced number of speakers and judges, and a wider variety of semantic descriptors to investigate the physical, psychological, sociological, and cultural aspects of the inherent multimodality scope of the use of gesture for communication.

Our investigation of the impressionistic effects of voice qualities, and associations between vocal and visual characteristics shifts the focus of sound-symbolic correspondences from speech segments to settings and gestures. In this way, it explores the four dimensions of expressivity implicit in the three Frequency, Sirenic, and Effort biological codes and in a metaphorical framework.

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

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. Written informed consent was obtained from the individual(s) for the publication of any identifiable images or data included in this article.

Author contributions

This paper contributes to the discussion of sound and meaning links by considering paralinguistic meanings associated with phonetic characteristics of voice quality settings. It also considers synesthetic associations between vocal and facial gestuality in expressing emotions. Three experiments are conducted and their results corroborate the communicative relevance of impressionistic and expressive uses of voice and face to express meanings. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the Pontifical Catholic University of São Paulo PIPEq funding Award No.: 21672.

Acknowledgments

The authors thank Dr. Plinio Barbosa and the reviewers for their valuable suggestions. SM also acknowledges a grant from PIPEQ-PUCSP No. 21672.

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/fcomm.2023.1114447/full#supplementary-material

References

Abelin, Å. (2007). "Emotional McGurk effect in Swedish," in *Proceedings of Fonetik TMH-QPSR* (Stockholm) 73–76.

Abelin, Å. (2008). "Seeing glee but hearing fear? Emotional Mc Gurk effect in Swedish," in *Proceedings of Speech prosody 2008* (Campinas) 713–716.

Abelin, A. (1999). *Studies in Sound Symbolism.* Doctoral dissertation. Gothenburg: Göteborg University.

Ahlner, F., and Zlatev, J. (2010). Cross-modal iconicity: a cognitive semiotic approach to sound symbolism. *Sign. Syst. Stud.* 38. 298–348. doi: 10.12697/SSS.2010.38.1-4.11

Anikin, A., Pisanski, K., Massenet, M., and Reby, D. (2021). Harsh is large: nonlinear vocal phenomena lower voice pitch and exaggerate body size. *Proc. Biol. Sci.* 288, 20210872. doi: 10.1098/rspb.2021.0872

Barbosa, P. (2021). "*Prosody Descriptor Extractor*" [Praat script]. Available online at: https://github.com/pabarbosa/prosodyscripts/tree/master/ProsodyDescriptorExtractor (accessed June 18, 2023).

Barbosa, P. A. (2009). "Detecting changes in speech expressiveness in participants of a radio program," in *Tenth Annual Conference of the International Speech Communication Association* (Brighton) 2155–2158. doi: 10.21437/Interspeech.2009-615

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., and Christiansen, M. H. (2016). Sound-meaning association biases evidenced across thousands of languages. *Proc. Nat. Acad. Sci. USA.* 113, 10818–10823. doi: 10.1073/pnas.1605782113

Boersma, P., and Weenik, D. (2022). "*Praat: doing phonetics by computer*" (Version 6.2.18), Available online at: https://www.fon.hum.uva.nl/praat/ [Computer program], online: http://www.praat.org (accessed June 18, 2023).

Cannon, W. B. (2016). *Bodily Changes in Pain, Hunger, Fear and Rage: An Account of Recent Researches into the Function of Emotional Excitement. Martino Fine Books.* New York, London: D. Appleton and Company.

Chen, A. J., Gussenhoven, C., and Rietveld, T. (2002). "Language-specific uses of the Effort Code," in *Proceedings of the Speech Prosody,* eds. B. Bel, and I. Marlien (Aix-en-Provence: Université de Provence) 215–218.

Christensen, C. M. (1980). Effects of taste quality and intensity on oral perception of viscosity. *Percept. Psychophys.* 28, 315–320. doi: 10.3758/BF03204390

Crochiquia, A., Eriksson, A., Fontes, M. A., and Madureira, S. (2020). Um estudo fonético das vozes de personagens do filme Zootopia na dublagem em português brasileiro: o papel dos estereótipos. *DELTA.* 36, 311. doi: 10.1590/1678-460x2020360311

Dingemanse, M., Blasi, D., Lupyan, G., Christiansen, M., and Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends Cogn. Sci.* 19. 603–615. doi: 10.1016/j.tics.2015.07.013

Ekman, P. (2016). What Scientists Who Study Emotion Agree About. *Perspect. Psychol. Sci.* 11, 31–34. doi: 10.1177/1745691615596992

Ekman, P., and Friesen, W. V. (1976). Measuring facial movement. *Environ. Psychol. Nonverbal Behav.* 1, 56–75. doi: 10.1007/BF01115465

Ekman, P., Friesen, W. V., and Hager, J. C. (2002). *Facial Action Coding System (2nd ed.).* Salt Lake City, UT: Research Nexus eBook.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton and Co.

Fónagy, I. (1983). *La vive voix: Essais de psycho-phonétique*. Paris: Payot.

Fónagy, I. (2001). *Languages within Language: An Evolutive Approach.* Amsterdam: John Benjamins. doi: 10.1075/fos.13

Fontes, M. A. S. (2014). *Gestualidade vocal e visual, expressão de emoções e comunicação falada. Tese de Doutorado.* Pontifícia Universidade Católica de São Paulo.

Gilbert, A. N., Fridlund, A. J., and Lucchina, L. A. (2016). The color of emotion: A metric for implicit color associations. *Food Quality Prefer.* 52, 203–210. doi: 10.1016/j.foodqual.2016.04.007

Green, K. P., and Norrix, L. W. (1997). Acoustic cues to place of articulation and the McGurk effect: the role of release bursts, aspiration, and formant transitions. *J. Speech, Lang. Hear. Res.* 40, 646–655. doi: 10.1044/jslhr.4003.646

Gussenhoven, C. (2002). "Intonation and interpretation: Phonetics and phonology," in *Proceedings of the 1st International Conference on Speech Prosody* (Aix-en-Provence), 47–57.

Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511616983

Gussenhoven, C. (2016). Foundations of intonation meaning anatomical and physiological factors. *Topics Cogn. Sci.* 8, 425–434. doi: 10.1111/tops.12197

Hewlett, N., and Beck, J. M. (2013). *An Introduction to the Science of Phonetics*. New York, NY: Routledge. doi: 10.4324/9780203053867

Hinton, L., Nichols, J., and Ohala, J. J. (1994). *Sound Symbolism.* Cambridge: Cambridge University Press.

Husson, F., Lê, S., and Pagès, J. (2009). *Exploratory Multivariate Analysis by Example Using R.* London: Chapman and Hall/CRC The R Series.

Jakobson, R. (1977). *Seis Lições sobre o Som e o Sentido.* Lisboa: Moraes Editores.

Jakobson, R., and Waugh, L. R. (1979). *The Sound Shape of Language.* Bloomington, IN: Indiana University Press and Harvester Press.

Jiang, J., and Bernstein, L. E. (2011). Psychophysics of the McGurk and other audiovisual speech integration effects. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 1193–1209. doi: 10.1037/a0023100

Johansson, N., Anikin, A., and Aseyev, N. (2020). Color sound symbolism in natural languages. *Lang. Cogn.* 12, 56–83.

Kawahara, S. (2021). Phonetic bases of sound symbolism: A review. *PsyArXiv [Preprint].* doi: 10.31234/osf.io/fzvsu

Keating, P., Garellek, M., and Kreiman, J. (2015). "Acoustic properties of different kinds of creaky voice. In the Scottish Consortium for ICPhs 2015," in *Proceedings of the 18th International Congress of Phonetic Sciences* (Glasgow: University of Glasgow). Available online at: http://www.internationalphoneticassociation.org/icphs536proceedings/ICPhS2015/Papers/ICPHS1041.pdf

Köhler, W. (1929). *Gestalt Psychology.* New York, NY: Liveright.

Körner, A., and Rummer, R. (2022). Articulation contributes to valence sound symbolism. *J. Exp. Psychol.: Gen.* 151, 1107–1114. doi: 10.1037/xge0001124

Kreiman, J., and Sidtis, D. (2011). *Foundations of Voice Studies: Interdisciplinary Approaches to Voice Production and Perception.* Boston, MA: Wiley-Blackwell.

Laver, J. (1976). "Language and nonverbedsal communication," *in Handbook of Perception, Vol. VII, Language and Speech*, eds E. C. Carterette and M. P. Friedman (New York, NY: Academic Press), 345–362.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge, MA: Cambridge University Press.

Laver, J., and Mackenzie-Beck, J. (2007). *Vocal Profile Analysis Scheme -VPAS [handout].* Edinburgh: Queen Margareth University College, Research Centre.

Léon, P. R. (1933). Précis de phonostylistique. Parole et expressivité, in the series Nathan Université. Paris: Nathan. *Canad. J. Lisguistique/ Revue Canadienne Dee linguistique.* 39, 369–371. doi: 10.107/S0008413100015590

Mackenzie Beck, J. (2005). "Perceptual analysis of voice quality: the place of vocal profile analysis," in *A Figure of Speech. A Festschrift for John Laver*, eds. W. Hardcastle and J. Mackenzie-Beck (London/Mahwah, NJ: Laurence Erlbaum Associates), 285–322.

Mackenzie Beck, J. (2007). *Vocal Profile Analysis Scheme: A User's Manual.* Queen Margaret University College-QMUC, Speech Science Research Centre, Edinburgh, United Kingdom.

Madureira, S., and Fontes, M. A. S. (2019). "The analysis of facial and speech expressivity: tools and methods," in *Subsidia: Tools and Resources for Speech Sciences, eds.* J. M. Lahoz-Bengoechea, and R. P. Ramón (Málaga: Universidade de Málaga) 1–150.

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

Monaghan, P., Shillcock, R. C., Christiansen, M. H., and Kirby, S. (2014). How arbitrary is language. *Philos. Trans. R. Soc. B. Biol. Sci.* 369, 20130299. doi: 10.1098/rstb.2013.0299

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *Am. Natur.* 111, 855–869. doi: 10.1086/283219

Nagrani, A., Albanie, S., and Zisserman, A. (2018). "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 8427–8436. doi: 10.1109/CVPR.2018.00879

Newman, S. S. (1933). Further experiments in phonetic symbolism. *Am. J. Psychol.* 45, 53–75. doi: 10.2307/1414186

Nobile, L. (2019). Introduction: Sound symbolism in the age of digital orality. A perspective on language beyond nature and culture. *Signifiances (Signifying).* 3, XXXVI–LXVIII. doi: 10.18145/signifiances.v3i1.248

Noldus. (2022). *FaceReader^{TM} 8.1: Tool for Automatic Analysis of Facial Expressions*. Wageningen, The Netherlands: Noldus Information Technology.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica* 41, 1–16. doi: 10.1159/000261706

Ohala, J. J. (1994). "The frequency codes underlies the sound symbolic use of voice pitch," in *Sound symbolism,* eds. L. Hinton, J. Nichols, and J. J. *Ohala* (Cambridge: Cambridge University Press) 325–347. doi: 10.1017/CBO9780511751806.022

Peterfalvi, J.-M. (1965). Les recherches expérimentales sur le symbolisme phonétique. *L'année Psychol.* 65, 439–474. doi: 10.3406/psy.1965.27443

Poyatos, F. (1993). *Paralanguage: A Linguistic and Interdisciplinary Approach to Interactive Speech and Sound*. Amsterdam: Journal of Benjamins Publisher. doi: 10.1075/cilt.92

Sapir, E. (1929). A study in phonetic symbolism. *J. Exper. Psychol.* 12, 225–239. doi: 10.1037/h0070931

Saussure, F., [1916] (2012). *Curso de Linguística Geral.* São Paulo: Cultrix.

Schaeffler, F., and, E., and Matthias, and Beck, J. (2019). "Towards ordinal classification of voice quality features with acoustic parameters," in *Proceedings of The Conference on Electronic Speech Signal Processing* (TU Dresden) 288–295.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Soc. Sci. Inf.* 44, 695–729. doi: 10.1177/0539018405058216

Scherer, K. R. (2022). Theory convergence in emotion science is timely and realistic. *Cogn. Emot.* 36, 154–170. doi: 10.1080/02699931.2021.1973378

Scherer, K. R., Clark-Polner, E., and Mortillaro, M. (2011)." in the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion. *Int. J. Psychol.* 46, 401–435. doi: 10.1080/00207594.2011.626049

Scherer, K. R., Dieckmann, A., Unfried, M., Ellgring, H., and Mortillaro, M. (2021). Investigating appraisal-driven facial expression and inference in emotion communication. *Emotion.* 21, 73–95. doi: 10.1037/emo0000693

Stevens, K. (1998). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Stevenson, R. J., and Boakes, R. A. (2004). "Sweet and sour smells: learned synesthesia between the senses of taste and smell," in *The Handbook of Multisensory Processes,* eds. G. A. Calvert, C. Spence, and B. E. Stein (Boston: MIT Press) 69–83. doi: 10.7551/mitpress/3422.003.0008

Sučević, J., Janković, D., and Ković, V. (2013). The sound-symbolism effect disappears: the differential role of order and timing in presenting visual and auditory stimuli. *Psychology.* 4, 11. doi: 10.4236/psych.2013.47A002

Swerts, M. G. J., and Krahmer, E. J. (2006). "The importance of different facial areas for signalling visual prominence," in *Proceedings of the International. Conference on Spoken Language Processing (Conference on Spoken Language Processing (Interspeech 2006).* doi: 10.21437/Interspeech.2006-377

Tsur, R. (1992). *What Makes Sound Patterns Expressive? The Poetic Mode of Speech Perception.* Durham, NC: Duke University Press. doi: 10.2307/j.ctv1131366

Vanger, P., Hoenlinger, R., and Haken, H. (1998). Computer aided generation of prototypical facial expressions of emotion. *Methods Psychol. Res. Online* 3, 25–38.

Vieira, C. O. (2014). *Crátilo, ou sobre a correção dos nomes.* São Paulo: Paulus.

Walker, H. K. (1990). "Cranial nerve VII: the facial nerve and taste," in *Clinical Methods: The History, Physical, and Laboratory Examinations,* 3rd edition. eds. H. K. Walker, W. D. Hall, and J. W. Hurst (Boston: Butterworths).

Westbury, C. (2005). Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain Lang.* 93, 10–19. doi: 10.1016/j.bandl.2004.07.006

Woodworth, N. L. (1991). Sound symbolism in proximal and distal forms. *Linguistics* 29, 273–299. doi: 10.1515/ling.1991.29.2.273

Xu, Y., Kelly, A., and Smillie, C. (2013). "Emotional expressions as communicative signals," in *Prosody and Iconicity,* eds. S. Hancil and D. Hirst (Amsterdam: John Benjamins) 33–60. doi: 10.1075/ill.13.02xu

Yanushevskaya, I., Gobl, C., and Chasaide, A. (2013). Voice quality in affect cueing: Does loudness matter? *Front. Psychol.* 4, 335. doi: 10.3389/fpsyg.2013.00335

Zhang, D., Zhou, Y., and Yuan, J. (2018). Speech prosodies of different emotional categories activate different brain regions in adult cortex: an fNIRS study. *Sci. Rep.* 8, 218. doi: 10.1038/s41598-017-18683-2

Frontiers in Communication

Investigates the power of communication across culture and society

A cross-disciplinary journal that advances our understanding of the global communication revolution and its relevance across social, economic and cultural spheres.

Discover the latest Research Topics

See more →

frontiers

Frontiers in Communication

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact