JASA ARTICLE

Adjustment of cue weighting in speech by speakers and listeners: Evidence from amplitude and duration modifications of Mandarin Chinese tone

Hui Zhang,[1,a)] Seth Wiener,[2,b)] and Lori L. Holt[3,c)]

[1]*Speech-Language-Hearing Center, School of Foreign Languages, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China*

[2]*Department of Modern Languages, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, USA*

[3]*Department of Psychology and Neuroscience Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, USA*

ABSTRACT:

Speech contrasts are signaled by multiple acoustic dimensions, but these dimensions are not equally diagnostic. Moreover, the relative diagnosticity, or weight, of acoustic dimensions in speech can shift in different communicative contexts for both speech perception and speech production. However, the literature remains unclear on whether, and if so how, talkers adjust speech to emphasize different acoustic dimensions in the context of changing communicative demands. Here, we examine the interplay of flexible cue weights in speech production and perception across amplitude and duration, secondary non-spectral acoustic dimensions for phonated Mandarin Chinese lexical tone, across natural speech and whispering, which eliminates fundamental frequency contour, the primary acoustic dimension. Phonated and whispered Mandarin productions from native talkers revealed enhancement of both duration and amplitude cues in whispered, compared to phonated speech. When nonspeech amplitude-modulated noises modeled these patterns of enhancement, identification of the noises as Mandarin lexical tone categories was more accurate than identification of noises modeling phonated speech amplitude and duration cues. Thus, speakers exaggerate secondary cues in whispered speech and listeners make use of this information. Yet, enhancement is not symmetric among the four Mandarin lexical tones, indicating possible constraints on the realization of this enhancement. © *2022 Acoustical Society of America*. https://doi.org/10.1121/10.0009378

(Received 1 October 2021; revised 7 January 2022; accepted 10 January 2022; published online 14 February 2022)

[Editor: Ewa Jacewicz] Pages: 992–1005

I. INTRODUCTION

Speech categories are realized across multiple acoustic dimensions, some of which are more diagnostic, or more perceptually weighted, than other, secondary dimensions in perception (Holt and Lotto, 2006; Idemaru and Holt, 2020; Lehet and Holt, 2016; Llanos *et al.*, 2013). The dominant role of the primary dimensions, however, does not eliminate the usefulness of the secondary dimensions, which are especially informative for phonetic category membership when the primary dimension is insufficient or unavailable (Heeren and Lorenzi, 2014; Higashikawa and Minifie, 1999; Holt and Lotto, 2006; Llanos *et al.*, 2013). In fact, it has been repeatedly reported that perceptual weights shift to emphasize secondary dimensions when primary cues are not available (e.g., Abramson, 1972; Best *et al.*, 1981; Holt and Lotto, 2006; Lin, 1988; Stevens and Klatt, 1974; Whalen and Xu, 1992). As regularities among acoustic dimensions shift in the input, listeners rapidly adapt such that the perceptual weights of acoustic dimensions in signaling speech categories shift dramatically (e.g., Idemaru and Holt, 2011; Zhang and Holt, 2018). This flexible adjustment means that the perceptual mapping of acoustic input dimensions to speech categories is dynamically adjusted in response to the listening context and its unique regularities.

There is evidence that the weighting of acoustic dimensions available in talkers' speech productions also shifts as a function of the communicative context (Lehet and Holt, 2016; Pardo, 2013; Sancier, 1997; Tilsen, 2009). Speech production appears to dynamically shift such that some acoustic dimensions are more strongly realized in speech production than others under changing communicative demands. For example, native English speakers down-weight fundamental frequency (F0) in distinguishing voiced consonants from voiceless consonants in both perception and production after short-term exposure to an "accent" of English for which the correlation of F0 and the primary cue voice onset time (VOT) reverses (Lehet and Holt, 2016).

Despite evidence for dynamic adjustment of the weight of acoustic dimensions in both speech perception and speech production, there are important gaps in our understanding.

[a)]Also at: Linguistics and Translation, City University of Hong Kong, Tat Chee Avenue, Hong Kong 999077, Hong Kong, ORCID: 0000-0002-9448-3511.

[b)]ORCID: 0000-0002-7383-3682.

[c)]Electronic mail: loriholt@cmu.edu

In particular, the literature is unclear on whether, and if so how, talkers adjust speech output to emphasize different acoustic dimensions in the context of different communicative demands. Here, we examine this issue in the context of whispering, which offers an ecologically significant example; whispering eliminates F0, leaving speakers to realize F0-dependent contrasts with secondary cues (like formant frequency and bandwidths, spectral center of gravity, amplitude, and duration) (Heeren and van Heuven, 2011; Heeren and van Heuven, 2009; Higashikawa and Minifie, 1999; Konno *et al*., 2006) in speech contrasts dependent upon F0, such as lexical tones in tonal languages, and prosodic distinctions between statements and questions.

Prior investigations have examined whether talkers might emphasize or enhance these secondary acoustic dimensions when the primary F0 dimension is unavailable in whispered speech. But, results are mixed. For example, in the context of whispered compared to phonated speech that realizes different pitch height targets, Heeren (2015a) found that whispering Dutch speakers produced larger differences in vowel spectral cues, such as formant frequency, center of gravity, and spectral balance (as compared to their phonated Dutch speech). In contrast, Heeren and Lorenzi (2014) found that when multiple prosodic events were in close proximity, whispering French speakers did not enhance formant frequency, but rather enhanced spectral tilt to convey statements as opposed to questions. In addition to the spectral correlates of vowels, Żygis *et al*. (2017) found that the spectral slopes of sibilants were enhanced in whispered speech for intonation contrasts in Polish, though the spectral slopes of fricatives /s/ and /f/ were not enhanced to realize pitch height targets in whispered Dutch (Heeren, 2015b). In the perceptual domain, Heeren and Lorenzi (2014) demonstrated that Dutch listeners were sensitive to spectral tilt for identifying intonation in whispers. As another perceptual demonstration, Higashikawa and Minifie (1999) reported that simultaneous changes in first and second formant frequencies affected listeners' judgment of the pitch height of synthesized /a/ vowels. In all, the evidence is mixed. Nonetheless, whether and how speakers' weighting of various acoustic dimensions shifts under different communicative demands and how listeners' perceptual cue weighting strategies adjust to accommodate these changes are important questions to resolve.

The present study builds from prior examinations of whispering in the context of lexical tone in Mandarin Chinese. Mandarin F0 contours (lexical tones) serve as distinctive features analogous to vowels and consonants for discriminating lexical meanings (Chao, 1965; Gandour, 1983) with four tones with distinctive F0 contours characterizing standard Beijing Mandarin: Tone 1 (high-level), Tone 2 (low-rising), Tone 3 (low-dipping), and Tone 4 (high-falling) (Ho, 1976; Howie, 1976). Under circumstances in which F0 information is unavailable for categorization, as in whispering, both spectral and non-spectral cues are likely to contribute to Mandarin tone identification. Kong and Zeng (2006) attributed the correct identification of whispered Mandarin tones to the role of a spectral envelope cue, although other effective spectral and non-spectral cues were also present in the stimuli, including amplitude envelope (Fu and Zeng, 2000; Kong and Zeng, 2006; Whalen and Xu, 1992) and duration (Blicher *et al*., 1990; Liu and Samuel, 2004).

Here, as a preliminary study of the interaction of speakers and listeners in enhancing secondary cues of Mandarin tones when F0 is not available, we focus on the role of non-spectral cues, namely amplitude and duration. If amplitude or duration dimensions are exaggerated in whispered, compared to phonated speech, we should observe an interaction of lexical tone and phonation type for amplitude contours and durations in Mandarin tone productions. If the enhancement of amplitude and duration is effective in signaling tone category to native listeners, we should observe a more accurate tone categorization of sound signals containing only the secondary amplitude and duration cues modeled by whispered, compared to phonated speech.

Two prior studies inform our approach. Jiao and Xu (2019) invited 12 native Mandarin speakers to read lexical tones as questions or statements carried by five isolated vowels in whispered and phonated contexts (/a/, /e/, /i/, /o/, /u/). Overall, vowel duration was longer in whispered compared to phonated speech, but the *relative* durations were the same; Tone 3 was the longest, and the other three tones were comparable. Similarly, although overall vowel amplitude was dramatically weakened in whispered speech, the *relative* amplitude and contour shapes across tones were similar for phonated and whispered vowels: Tone 3 exhibited a bimodal amplitude pattern and Tone 4 demonstrated a greater drop in the latter part of the syllable than Tone 1 and Tone 2. Taken together, the authors concluded that neither duration nor amplitude (mean, contour) was exaggerated as secondary cues to whispered Mandarin lexical tone.

In a subsequent perceptual experiment, Jiao and Xu (2019) asked 22 native Mandarin listeners to identify the tones of one native female's phonated speech, her whispered speech, and amplitude-modulated white noises synthesized to mimic the duration and amplitude contours of her whispered versus phonated speech. The identification accuracy of whispered tones was dramatically lower than that of phonated tones. Tone identification accuracy of the noises was comparable for noises synthesized with whispered and phonated characteristics, suggesting that the duration and amplitude cues of whispered tones did not have more perceptual value than those cues from phonated tones and echoing results of the production experiment. In a word, Jiao and Xu (2019) provided evidence that neither amplitude nor duration was enhanced in whispered speech since these two cues were not more greatly differentiated in whispered speech than phonated speech, nor did they have more perceptual value.

It is worth noting that the speech analyzed by Jiao and Xu (2019) was read and not directed at a listener. This may be important inasmuch as the acoustics of read speech can differ dramatically from the speech in communicative contexts (Brown-Schmidt, 2005; Buxó-Lugo *et al*., 2018; Laan,

1992; Samuel and Troicki, 1998; Schober and Clark, 1989; Wang *et al.*, 2010; Buz *et al.*, 2016; Nakamura *et al.*, 2008). Potentially, putative exaggeration of secondary cues may be reduced when there is no target listener. Further, the task in Jiao and Xu (2019) asked speakers to realize intonation on lexical tones, which may have altered speakers' emphasis on tone identity (see Ouyang and Kaiser, 2015). Finally, it is possible that idiosyncratic patterns of the single female talker used in the perception experiment are not representative of the general patterns in Mandarin whispered speech. Thus, there remain open questions about whether the conclusion that Mandarin speakers do not exaggerate secondary cues to lexical tone holds more generally.

In fact, this caution is underscored by an important perceptual study lending evidence for enhancement of duration in whispered Mandarin lexical tone (Liu and Samuel, 2004). Liu and Samuel reported that listeners better identified native whispered speech than "machine" whispered speech created by replacing the voicing of phonated speech with white noise to remove F0 while maintaining other acoustic cues of phonated speech. This would suggest the perceptual value in the whispered speech that is unavailable in the phonated speech stripped of F0. Results indicate a role for enhanced duration in whispered speech, with listeners allocating more weight to duration in native whispered lexical tone recognition. A durational analysis of Liu and Samuel's auditory stimuli, however, showed that phonated and whispered speech did not differ in any individual tone category. Liu and Samuel concluded that the improved performance for the naturally whispered speech over the stimuli that were signal-processed to remove F0 could be attributed to the relative duration changes within the stimulus set.

In all, the literature leaves quite a bit of uncertainty about the extent to which speakers may enhance secondary cues when primary cues are unavailable and, if so, whether listeners are able to capitalize on this enhancement. In the present study, we build from the work of Liu and Samuel (2004) and Jiao and Xu (2019) to examine whether speakers adjust the weight of amplitude and duration under active communicative demands and whether listeners are sensitive to these changes.

Our study extends the two previous studies in several ways. First, we record a wide range of consonant–vowel syllables to simulate the speech behavior in a communicative situation in which a speaker is encouraged to relay speech to a real listener. Second, we base our results on the average across 30 different Mandarin speakers' phonated and whispered speech to minimize idiosyncratic characteristics of individual speakers and extend previous examinations. Third, we disentangle amplitude and duration to clarify their respective enhancement in production and the respective roles of enhancement in perception. The combination of acoustic analysis and perceptual performance informs whether and how speakers change the weighting of various acoustic cues when the listening context changes and how listeners' perceptual cue weighting strategy adjusts to accommodate that change.

II. EXPERIMENT 1: SPEECH PRODUCTION

A. Method

1. Participants

Thirty speakers (15 males and 15 females; mean age = 29.8 yr) were recorded. All speakers were born and grew up in Chinese provinces with dialects of Mandarin, including Shandong, Henan, Shanxi, Hebei, Sichuan, Hubei, and east of Inner Mongolia (Chao, 1943; Norman, 1988). All participants reported speaking Beijing Mandarin as a major language in daily life and occasional dialects in family communication. No participant reported any speaking problems. All participants also reported speaking English as a foreign language, with proficiency ranging from poor to excellent. At the time of recording, three participants were located in the U.S. and the rest in China. None of the participants had lived in the U.S. for more than two years.

2. Materials

We selected 19 of the 20 monosyllables used in Liu and Samuel (2004). One syllable /p^{h} ai/ was excluded because its Tone 3 is rarely used in speech. The resultant 76 combinations of syllables and tones (19 syllables × four tones) are all commonly used Mandarin monosyllabic morphemes/words as judged by a native speaker and Cai and Brysbaert (2010). (See supplementary material for a list of the stimuli with a representative Chinese character, Pinyin romanization, and IPA.)[1] The list also provides the token frequency of each syllable per million words according to SUBTLEX_CH (Cai and Brysbaert, 2010), and the number of common homophones with a frequency higher than 0.5 per million according to the Linguistic Institute of Chinese Social Academy (1982) and Cai and Brysbaert (2010).

3. Procedure

In response to the COVID-19 pandemic, our recordings were mainly made using a novel remote recording procedure. Twenty-five participants recorded speech on their home computers using Praat (Boersma and Weenink, 2018), and two participants recorded speech using a built-in iPhone recorder. An additional three participants' recordings were made using Praat (Boersma and Weenink, 2018) in the laboratory in a sound-attenuated booth before the stay-at-home order. Although the lab recording has the benefit of producing high-quality sounds, it lacks a real-life context that simulates the conditions of everyday communication. The remote recording traded the control of sound quality for a real-life listening situation. Specifically, the speakers were situated in a room where there was some level of background noise and spoke in a real virtual meeting. Although we recognize this is a departure from many acoustic phonetics experiments conducted under tightly controlled circumstances, we believe that the production data in this recording situation could more realistically represent everyday speech.[2]

All recordings were made at a sampling rate of 44 100 Hz with 16-bit resolution. Participants were asked to familiarize themselves with the pronunciation of the list of 76 words represented by common characters, pinyin romanization, and tone notation in three separate columns in an Excel file before recording. The order of the words was randomized and re-shuffled for each participant. For the lab recording, the reading list was presented on a computer screen in front of the participants, and the participants read these utterances into a Shure SM58 microphone (Niles, IL). For the remote recording, both the presentation of the reading list and the recording were done on the participants' personal laptops in a quiet room. Specifically, participants saw the reading list in an Excel file and read each target into the built-in microphone of their own laptops.

The native Mandarin experimenter (the first author and a native Mandarin speaker) acted as a listener giving feedback about what tone the speaker said. This approach was designed to motivate participants to produce listener-oriented speech instead of read speech.[3] Upon hearing the participant's utterance, the experimenter said aloud the perceived tone. If the experimenter was correct, the participant continued to the next word. If the experimenter was wrong, the participant said, "wrong answer" and repeated the word. Participants took optional breaks throughout the recordings. The two phonation modes were recorded in two separate blocks with a fixed order: phonated words first and whispered second.

During the break between the phonated mode and whispered mode, the participants were briefed on how to whisper and the situations where whispers could occur to help them whisper in a natural way, though all participants indicated to have whispered before.[4] The participants were then asked to practice whispering. The participants would proceed to the recording once the experimenter deemed them ready. The whole procedure for each participant lasted roughly 30 min. The entire recording session including instructions involved only spoken Mandarin.

Due to the experimenter's misidentification, nine phonated words and 802 whispered words (Tone 1: 479; Tone 2: 239; Tone 3: 34; Tone 4: 50) were repeated. This indicated that it was often the case that speakers did not succeed in producing understandable whispered Tone 1 and Tone 2 the first time around. For the repeated words, two native speakers of Mandarin Chinese (including the experimenter) selected one repetition that best realized its tone identity (mostly the second or third repetition) for acoustic analysis.[5] The acoustic analysis and the following perceptual experiment were, therefore, mostly based on the first repetitions for Tone 3 and Tone 4, whereas mostly based on the second or third repetitions for Tone 1 and Tone 2.[6] We return to this methodological point in the discussion section.

The recordings were saved as wav files (see supplementary material for the recordings)[1] for each participant in each phonation mode (whispered, phonated) for a total of 60 files (30 participants × 2 phonation modes) constituted by 4 560 words (30 participants × 2 phonation modes × 19 syllables × 4 tones). None of the speakers had trouble whispering, so no speaker was discarded. Forty-three words were eliminated due to recording errors, resulting in 4 517 total words. The sound files were annotated manually at the syllable level using TextGrids in Praat version 6.0.37 (Boersma and Weenink, 2018), according to the visual inspection of the spectrogram and waveform, or by auditory judgment if necessary. The word/syllable onset was marked as the onset of the initial consonants. For stops, the onset was marked by the sudden sharp spike in the spectrogram.[7] For fricatives, the onset was marked by the start of random energy distributed over a wide range of frequencies in the spectrogram and the onset of increasingly higher amplitude. For approximants, the onset was marked by the first periodic cycle in the waveform and the onset of formants in the spectrogram. The word/syllable offset was marked as the F1 offset taken at the nearest zero crossing. The offset of F1 was often accompanied by an abrupt decrease in the amplitude of the waveform.

4. Acoustic analysis

For each word, the duration and amplitude contour were measured using Praat (Boersma and Weenink, 2018). Duration was obtained by the Get selection length method in the editor window (see supplementary material for all Praat scripts).[1] Amplitude contours were obtained with 100 time-normalized amplitude points automatically extracted from each whole word/syllable in Praat (Boersma and Weenink, 2018) (pitch floor: 50 Hz, time step strategy: automatic, subtract mean pressure).

Duration and amplitude were modeled using linear mixed-effect models with the lme4 package in R (Bates *et al.*, 2015). For duration, the fixed predictors were Phonation (categorical factor, dummy coded as phonated and whispered), Tone (categorical factor, dummy coded as Tone 1, Tone 2, Tone 3, and Tone 4), and Gender (categorical factor, dummy coded as female and male) along with a two-way interaction between Phonation and Tone. The duration values were log-transformed (natural logarithm) to adjust for the skewness of the distribution. The Box–Cox procedure confirmed that the log transformation was appropriate ($\lambda = 0.15$). The maximal model included by-participant and by-syllable random intercepts, by-syllable random slopes for the three fixed predictors, and by-participant random slopes for Phonation and Tone. The optimal model was found using a backward elimination procedure (Table I). *Post hoc* comparisons among tones were made by changing the reference levels. (See supplementary material for all data and R code detailing our statistical analyses.)[1]

Since our purpose of analyzing amplitude was to show that the amplitude differs across tones and phonation types not only in terms of the average, but the trajectory over time, we modeled amplitude contours over 100 time-points using growth curve analysis (Mirman, 2014). In this way, we were able to model the slope and curvature of amplitude over time

TABLE I. Linear mixed-effects regression model[a] output on duration of four tones in whispered and phonated speech produced by two genders.

Random effects	Variance	Standard deviation	N	Observation
Participant (intercept)	0.029	0.171	30	4517
Syllable (intercept)	0.004	0.067	19	4517
Fixed effects	**Estimate**	**Standard error**	**t-value**	**p-value**
(Intercept: Tone 1, phonated, female)	−0.640	0.041	−15.694	<0.001
Whispered	0.045	0.037	1.217	0.230
Tone 2	0.023	0.026	0.860	0.395
Tone 3	0.200	0.022	9.221	<0.001
Tone 4	−0.221	0.029	−7.663	<0.001
Male	−0.133	0.041	−3.231	0.003
Tone 2: Whispered	−0.132	0.013	−10.289	<0.001
Tone 3: Whispered	0.074	0.013	5.764	<0.001
Tone 4: Whispered	−0.223	0.013	−17.398	<0.001

[a]Formula: Duration ∼ Tone + Phonation + Gender + Tone: Phonation + (1+Tone + Phonation + Gender | Syllable) + (1 + Tone + Phonation| Participant).

rather than simply the mean amplitude. Figure 1(B) shows the amplitude contours across time by Tone and Phonation. We observed that the contours of Tone 3 and possibly whispered Tone 2 exhibited four changes of directions with a dipping curve in the middle. The remaining contours showed two changes of directions with reversed U-shaped curves. The overall amplitude shapes were accordingly modeled with first-order (linear), second-order (quadratic), third-order (cubic), and fourth-order (quartic) polynomials. Of interest was the effect of Tone (categorical factor: dummy coded as Tone 1, Tone 2, Tone 3, and Tone 4) and Phonation (categorical factor: dummy coded as phonated and whispered), as well as their interaction on all time terms. We also included Gender (categorical factor: dummy coded as female and male) in the model to consider its possible effect. The model also involved random effects of speakers and syllables on all time terms. Because the model failed to converge, we removed the third- and fourth-order polynomials and the correlation between first-order and second-order polynomials from the random effects. Therefore, the random effect of speakers and syllables only included the first- and second-order polynomials. The model was specified in formula in the footnote of Table II. Female phonated Tone 1 was set as the reference level or intercept, and parameters were estimated for Tone 2, Tone 3, Tone 4, and whispered speech overall and on all time terms. We were specifically interested in Tone × Phonation interactions on all time terms as this estimate could capture larger amplitude differences among tones in whispered than in phonated speech.

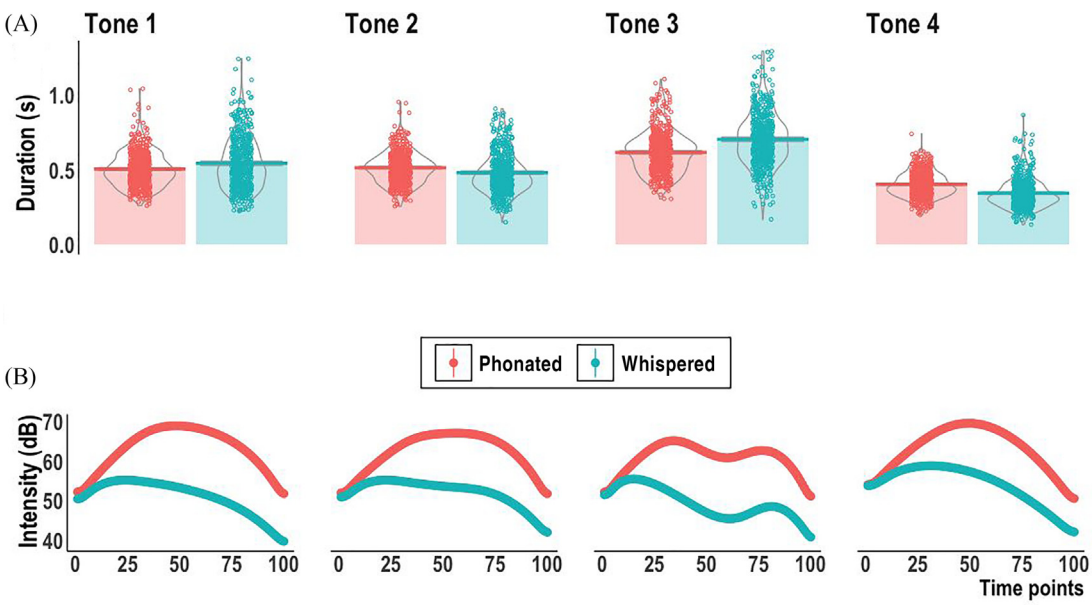

FIG. 1. (Color online) Duration and amplitude in phonated and whispered speech. (A) Duration as a function of tone and phonation with individual data points, density distributions, and means. (B) Average amplitude as a function of normalized time point, tone and phonation with 95% confidence intervals.

https://doi.org/10.1121/10.0009378

TABLE II. Growth curve analysis model[a] output on amplitude of four tones in whispered and phonated speech produced by two genders.

Random effects	Variance	Standard deviation	N	Observation
Participant (First-order polynomial)	115.130	10.730	30	451 700
Participant (Second-order polynomial)	46.250	6.801	30	451 700
Syllable (First-order polynomial)	103.950	10.196	19	451 700
Syllable (Second-order polynomial)	52.080	7.217	19	451 700
Fixed effects	**Estimate**	**Standard error**	**t-value**	**p-value**
(Intercept: Tone 1, phonated, female)	62.770	0.050	1252.703	<0.001
First-order polynomial: Tone 2: Whispered	4.952	0.900	5.505	<0.001
First-order polynomial: Tone 3: Whispered	5.588	0.901	6.204	<0.001
First-order polynomial: Tone 4: Whispered	5.673	0.901	6.299	<0.001
Second-order polynomial: Tone 2: Whispered	−3.103	0.900	−3.449	0.001
Second-order polynomial: Tone 3: Whispered	5.216	0.901	5.791	<0.001
Second-order polynomial: Tone 4: Whispered	−3.942	0.901	−4.377	<0.001
Third-order polynomial: Tone 2: Whispered	0.153	0.900	0.170	0.865
Third-order polynomial: Tone 3: Whispered	−3.696	0.901	−4.103	<0.001
Third-order polynomial: Tone 4: Whispered	3.060	0.901	3.399	0.001
Fourth-order polynomial: Tone 2: Whispered	0.286	0.900	0.318	0.750
Fourth-order polynomial: Tone 3: Whispered	−0.334	0.901	−0.371	0.711
Fourth-order polynomial: Tone 4: Whispered	0.221	0.901	0.245	0.806

[a]Formula: Amplitude ~ (First-order + Second-order + Third-order + Fourth-order)*Tone *Phonation + Gender + Gender: Phonation + (First-order |Speaker) + (First-order||Syllable) + (Second-order| Syllable) + (Second-order |Speaker).

B. Result

1. Duration

Figure 1(A) plots the duration for each phonation type by tone type. In both phonated and whispered speech, Tone 3 had the longest duration, and Tone 4 had the shortest duration. The result of the *post hoc* comparisons (see Table I and supplementary material for the result)[1] showed that Tone 3 was significantly lengthened in whispered speech [ß = 0.119, standard error (SE) = 0.037, $p = 0.002$] whereas Tone 2 and Tone 4 was significantly shortened (ß = –0.087, SE = 0.037, $p = 0.024$ for Tone 2 and ß = –0.178, SE = 0.037, $p < .001$ for Tone 4) in whispered compared to phonated speech. The duration of Tone 1 (ß = 0.045, SE = 0.037, $p = 0.230$) did not significantly differ across the two phonation modes. The magnitude of difference between phonated and whispered tones was larger for Tone 3 (0.087 s) and Tone 4 (0.060 s) than Tone 2 (0.033 s).

Figure 1(A) also shows that the effect of Tone is similar for both phonation types. Specifically, Tone 3 was longer than Tone 1 and Tone 2, which were again longer than Tone 4 in both phonated and whispered speech (*ps* <0.001). The *post hoc* comparisons among tones with a Tone × Phonation interaction indicated that the differences in duration among the four tones were larger in whispered speech than phonated speech (see supplementary material for the coefficients of interactions).[1]

2. Amplitude contour

As shown in Fig. 1(B), the amplitude contours of Tone 3 and Tone 4 resembled the canonical F0 contours of the corresponding tone categories, which were dipping and falling, respectively. The amplitude contours of Tone 1 and Tone 2, however, differed from the canonical F0 contours of the corresponding tone categories, which are level and rising. For the sake of brevity, only the Tone × Phonation interaction is shown in Table II, as this is our primary interest (see supplementary material for the outputs of *post hoc* analyses obtained by changing the reference levels).[1] The model (with female phonated Tone 1 as the reference level) revealed that the effects of Tone and Phonation were significant overall (except for Tone 1–Tone 4 difference) and on all time terms (see supplementary material for the model),[1] indicating that Tone and Phonation changed not only the average amplitude but also the slope and curvature of the amplitude contours. Releveling of the model revealed the same pattern for the other three tones. This suggests that eight unique contours were found corresponding to four tones and two phonation types, each of which differed from one another in terms of the slope and curvature. Notably, the Tone × Phonation interactions were statistically significant on first-, second-, and third-order time terms, indicating possibly larger differences in amplitude contour among tones in whispered speech than in phonated speech. Following Mirman (2014), we visualized the effects of those terms by fitting a reduced model with the specific time terms removed from the Tone × Phonation interaction and then visually compared the model fits (see R code online). The visualizations (see supplementary material for the visualization of amplitude difference among tones across time)[1] showed that the full model with Tone × Phonation interactions on first-, second-, and third-order time terms confirmed larger amplitude differences among tones in whispered speech than in phonated speech, whereas the reduced model showed identical differences among tones for these two phonation types.

C. Experiment 1: Discussion

We found that whispered Mandarin tones changed both duration and amplitude contour relative to phonated tones. In terms of duration, whispered Tone 3 became significantly longer than its phonated counterpart, and Tones 2 and 4 became significantly shorter (relative to phonated counterparts). But the magnitude of change for Tone 2 was much weaker than Tone 3 and Tone 4. In addition, tone durations differed more greatly in whispered speech than in phonated speech. Thus, duration differences across the four tones were enhanced in whispered speech. The duration measurements are consistent with Liu and Samuel (2004)'s suggestion that the relative durations of the long tone (Tone 3) and the short (Tone 4) would be made more distinct in whispered speech. Using an interactive recording procedure, we found an enhancement of duration in the speakers' production, which was not observed in Jiao and Xu (2019).

With respect to amplitude contour, all four whispered tones were produced with lower overall contours relative to their phonated counterparts, as would be expected for whispers. By considering the amplitude over time, we found that speakers essentially produced eight unique contours that each differed from one another both within and across Phonation. The Tone × Phonation interaction on first-, second-, and third-order polynomial time terms with subsequent visualization of the interaction effects indicated that speakers enhanced the difference in amplitude contours among tones in whispered speech compared with phonated speech. These novel amplitude contour results go beyond the results of Liu and Samuel (2004) and Jiao and Xu (2019) and demonstrate how phonation types can yield changes in amplitude characteristics in listener-directed speech.

Combing the data of duration and amplitude, we found that Tone 3 had the lowest and most variable amplitude contour and was the tone that was sustained for the longest time and with the most modulation. Tone 4 had the highest amplitude and was short and simple in its amplitude contour. The amplitude contours of both Tone 3 and Tone 4 simulated their respective canonical F0 contours, which were dipping and falling, respectively. Tone 1 and Tone 2, however, were in between in duration and had amplitude contours distinct from their respective canonical F0 contours, which should have been level and rising. This was possibly caused by the high demand of articulation for whispered Tone 1 and Tone 2. However, a more detailed explanation will be presented in the Discussion section. We next turn to the perceptual experiment to understand whether listeners make use of these enhancements.

III. EXPERIMENT 2: PERCEPTUAL EXPERIMENT

A. Method

1. Participants

One hundred and eighty-eight listeners (76 males and 112 females; mean age = 22.59 yr; age range: 18–39 yr) were recruited for the perceptual experiment. All listeners were born and grew up in mainland Chinese provinces where the dialects were varieties of Mandarin (Chao, 1943; Norman, 1988). All participants self-reported speaking Beijing Mandarin as a major language in their daily life and occasional dialects in family communication. No participant reported any hearing impairment. All participants reported speaking English as a foreign language, ranging from poor to proficient. The study was approved by the authors' Institut Review Board, and all participants were paid for their time.

2. Materials

The test materials for the perceptual experiment, which were inspired by the materials of Liu and Samuel (2004) and Jiao and Xu (2019), were phonated words, whispered words, and amplitude-modulated noises. Phonated words (N = 2156) and whispered words (N = 2121), which were the words recorded in Experiment 1, were used to build blocks A and B, respectively. In order to shorten the overall task length, we sampled 220 phonated words and 220 whispered words uniquely for each participant in a pseudo-randomized manner. In other words, each participant heard a roughly unique set of 220 of the possible 2156 phonated words and 220 of the possible 2121 whispered words, resulting in 55 items per tone and per phonation type ($55 \times 4 \times 2 = 440$). This meant each utterance from Experiment 1 was heard and labeled for its tone by roughly 20 different listeners. Importantly, the 55 items per tone involved at least 27 different speakers and 17 different syllables. This syllable variability more closely approximated that of a realistic speech environment.

In block C, amplitude-modulated noises were synthesized to mimic the amplitude and duration information of the phonated and whispered tones within nonspeech, white-noise stimuli through a custom Matlab script (see supplementary material for the script). The average duration [refer to Fig. 1(A)] and amplitude contour [refer to Fig. 1(B)] of each tone type (by phonation type) were used as the base to synthesize the noises. For example, we copied the amplitude contour and duration information of phonated Tone 1 to the white noise to synthesize the amplitude-modulated noise for phonated Tone 1. In this way, the tone information was reduced to only containing amplitude and duration information. The number of samples for each noise was decided by duration and sampling rate, which was 44 100 Hz. Since we only obtained 100 amplitude points for the contour analysis, we used linear interpolation for point expansion so that Matlab could have the amplitude value of each sample. The synthesis resulted in eight noises with duration and amplitude information of a particular Tone by Phonation [see the shaded stimuli along the diagonal in Fig. 2(A)]. In addition, we also synthesized stimuli incongruent between duration and amplitude by, for example, combining the duration of a phonated Tone 1 with the amplitude contour of whispered Tone 4 [see the remaining stimuli in Fig. 2(A)]. This resulted in a total of 64 items (8 *congruent* noises + 56

https://doi.org/10.1121/10.0009378

(A)

Amplitude \ Duration		Duration: Phonated (ph)			
		Tone 1	Tone 2	Tone 3	Tone 4
Amplitude: Phonated(ph)	Tone 1	ph1-ph1	ph1-ph2	ph1-ph3	ph1-ph4
	Tone 2	ph2-ph1	ph2-ph2	ph2-ph3	ph2-ph4
	Tone 3	ph3-ph1	ph3-ph2	ph3-ph3	ph3-ph4
	Tone 4	ph4-ph1	ph4-ph2	ph4-ph3	ph4-ph4
Amplitude: Whispered(wh)	Tone 1	wh1-ph1	wh1-ph2	wh1-ph3	wh1-ph4
	Tone 2	wh2-ph1	wh2-ph2	wh2-ph3	wh2-ph4
	Tone 3	wh3-ph1	wh3-ph2	wh3-ph3	wh3-ph4
	Tone 4	wh4-ph1	wh4-ph2	wh4-ph3	wh4-ph4

Amplitude \ Duration		Duration: Whispered (wh)			
		Tone 1	Tone 2	Tone 3	Tone 4
Amplitude: Phonated(ph)	Tone 1	ph1-wh1	ph1-wh2	ph1-wh3	ph1-wh4
	Tone 2	ph2-wh1	ph2-wh2	ph2-wh3	ph2-wh4
	Tone 3	ph3-wh1	ph3-wh2	ph3-wh3	ph3-wh4
	Tone 4	ph4-wh1	ph4-wh2	ph4-wh3	ph4-wh4
Amplitude: Whispered(wh)	Tone 1	wh1-wh1	wh1-wh2	wh1-wh3	wh1-wh4
	Tone 2	wh2-wh1	wh2-wh2	wh2-wh3	wh2-wh4
	Tone 3	wh3-wh1	wh3-wh2	wh3-wh3	wh3-wh4
	Tone 4	wh4-wh1	wh4-wh2	wh4-wh3	wh4-wh4

(B)

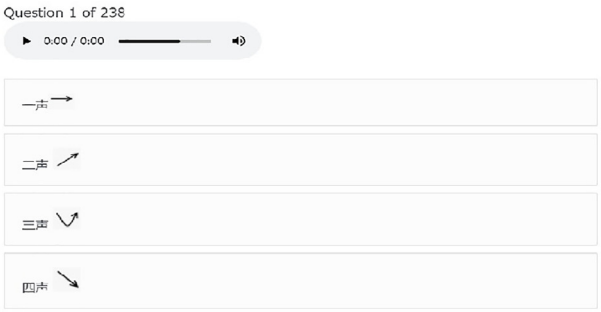

FIG. 2. (Color online) (A) Illustration of the duration and amplitude information for the 64 amplitude-modulated noises. Shaded boxes indicate congruent duration and amplitude whereas non-shaded boxes indicate incongruent duration and amplitude. The letters *ph* represent phonated and *wh* whispered. Numbers refer to tone categories. (B) The options displayed on the screen for each trial.

incongruent noises), each of which was presented four times to the participants for a total of 256 trials.

3. Procedure

Participants took part in the perceptual experiment online on the web platform Qualtrics[8] (Qualtrics, 2014). There were three blocks in the perceptual experiment, namely phonated (block A), whispered (block B), and noises (block C). The order of the blocks was counterbalanced so that there were six fully counterbalanced presentation orders. The 188 participants were randomly assigned to the six orders, with roughly 30 participants per order. Each block began with practice trials followed by 220 trials (phonated and whispered blocks, respectively) or 256 trials (noise block). Participants were instructed to set the sound volume to a comfortable level during the practice trials of each block. Then, they were told not to change sound volume within the block. In each trial, a sound was automatically played. During the sound presentation, four options displaying the labels of four tones and their tone shapes were shown on the screen [see Fig. 2(B)]. Participants were asked to select the tone that the sound best matched. The stimuli could be played only once. The perceptual experiment took about 40 min with optional breaks between blocks.

4. Data analysis

The results consist of the hit rate of tone categorization (hereafter accuracy) of the phonated words, whispered words, and amplitude-modulated noises. The accuracy of the phonated and whispered words served as a baseline showing the identification of tones with and without F0 cues, i.e., the primary cue to Mandarin lexical tone in clear, phonated speech. Accuracy was modeled using a mixed-effects logistic regression model in R. The dependent variable was accuracy (coded as 1 or 0). For the phonated and whispered words, the fixed variables included Tone (categorical factor: dummy coded as Tone 1, Tone 2, Tone 3, and Tone 4), Phonation (categorical factor: dummy coded as phonated and whispered), and Block order (categorical factor: dummy coded as ABC, ACB, BCA, BAC, CAB, and CBA). By-participant and by-item (speaker and syllable) random intercepts and random slopes were included. Since the maximal model failed to converge, we removed the interactions of Tone and Phonation from random slopes. The optimal model was found using a backward elimination procedure. The formula was specified in the footnote of Table III.

There were two stages in examining the noise identification data. The accuracy of the eight *congruent* noises was examined in the first stage. The correct answer was the tone the noises modeled. The purpose of this stage was to compare the extent to which tone identity was carried by amplitude and duration information across the two phonation types. The maximal model for noise results included the fixed variables of Tone, Phonation, and Block order, with by-participant and by-item random intercepts, by-participant random slopes for Tone and Phonation, and by-item random slope for Block order. The optimal model was found in a backward elimination procedure (formula in the footnote of Table IV). Variable coding and model fitting followed the same outlined procedure.

If the accuracy of whispered noises outperformed phonated noises, a second-stage examination disentangling the contributions of duration and amplitude enhancement was

TABLE III. Logistic mixed-effects regression model[a] output on correct whispered and phonated tones in six order versions.

Random effects	Variance	Standard deviation	N	Observation
Participant (intercept)	1.742	1.320	161	74 704
Item (intercept)	1.430	1.956	540	74 704
Fixed effects	**Estimate**	**Standard error**	**z-value**	**p-value**
(Intercept: Tone 1, phonated)	4.536	0.138	32.985	<0.001
Tone 2	−1.059	0.136	−7.814	<0.001
Tone 3	−1.006	0.160	−6.286	<0.001
Tone 4	0.724	0.169	4.277	<0.001
Whispered	−5.553	0.121	−46.044	<0.001
Tone 2: Whispered	1.143	0.114	10.031	<0.001
Tone 3: Whispered	4.063	0.119	34.042	<0.001
Tone 4: Whispered	0.486	0.137	3.538	<0.001

[a]Formula: Accuracy ∼ Tone + Phonation + Tone: Phonation + (1 + Tone + Phonation |Participant) + (1 + Tone + Phonation| Item).

conducted as a follow-up analysis. As a reminder, our production results showed that both duration and amplitude contour were enhanced in whispered speech for Tone 3 and 4. Specifically, whispered Tone 3 had a longer duration and different amplitude contour than phonated Tone 3, whereas whispered Tone 4 had a shorter duration and different amplitude contour than phonated Tone 4. To clarify the respective roles of duration and amplitude enhancement in improving the accuracy, we selected four noises of the following amplitude-duration combinations (hereafter, amplitude always precedes duration) for Tone 3 and Tone 4, respectively: Phonated amplitude (Phonamp)-Phonated duration (Phondur), Phonamp-Whispered duration (Whisdur), Whispered amplitude (Whisamp)-Phondur, and Whisamp-Whisdur. In other words, compared with the Phonamp-Phondur noises, this created three noises: amplitude-enhanced noise (Whisamp-Phondur), duration-enhanced noise (Phonamp-Whisdur), and amplitude-duration–enhanced noise (Whisamp-Whisdur). For Tone 3–based noises, we submitted listeners' responses (coded as 1 for Tone 3 and 0 for other tones) to a logistic regression model. The fixed predictors included the enhancement information of amplitude (categorical factor: enhanced and not enhanced) and duration (categorical factor: enhanced and not enhanced) and their interaction. Both by-participant random intercept and slope for the two fixed predictors were included as the random variables. For Tone 4–based noises, we submitted accuracy to a logistic regression model with the same predictors.

B. Results

1. Accuracy of phonated and whispered words

Figure 3(A) shows the identification accuracy of phonated and whispered tones. The accuracy of whispered Tone 1 decreased by 65.9% than that of phonated Tone 1 (ß = –5.553, SE = 0.121, $p < 0.001$), 60.7% for Tone 2 (ß = –4.410, SE = 0.107, $p < 0.001$), 9.2% for Tone 3 (ß = –1.491, SE = 0.110, $p < 0.001$), and 44.9% for Tone 4 (ß = –5.068, SE = 0.133, $p < 0.001$) (see Table III and supplementary material for outputs of *post hoc* analyses obtained by changing the reference levels).[1] The accuracy was 30.5% with a 95% confidence interval (CI) of 29.6%–31.5% for whispered Tone 1, and 32.3% with a 95% CI of 31.3%–33.2% for whispered Tone 2. The accuracies for whispered Tones 1 and 2 were therefore all slightly

TABLE IV. Logistic mixed-effects regression model[a] output on correct amplitude-correlated noises in whispered and phonated speech in six order versions.

Random effects	Variance	Standard deviation	N	Observation
Participant (intercept)	1.129	1.063	161	5152
Item (intercept)	0.000	0.000	8	5152
Fixed effects	**Estimate**	**Standard error**	**z-value**	**p-value**
(Intercept: Tone 1, phonated)	−1.252	0.136	−9.227	<0.001
Tone 2	0.371	0.185	2.003	0.045
Tone 3	3.086	0.247	12.518	<0.001
Tone 4	1.607	0.201	8.001	<0.001
Whispered	−0.039	0.139	−0.278	0.781
Tone 2: Whispered	−0.618	0.195	−3.164	0.002
Tone 3: Whispered	0.606	0.219	2.775	0.006
Tone 4: Whispered	1.034	0.194	5.317	<0.001

[a]Formula: Accuracy ∼ Tone + Phonation + Tone: Phonation + (1 | Item) + (1 + Tone| Participant).

https://doi.org/10.1121/10.0009378

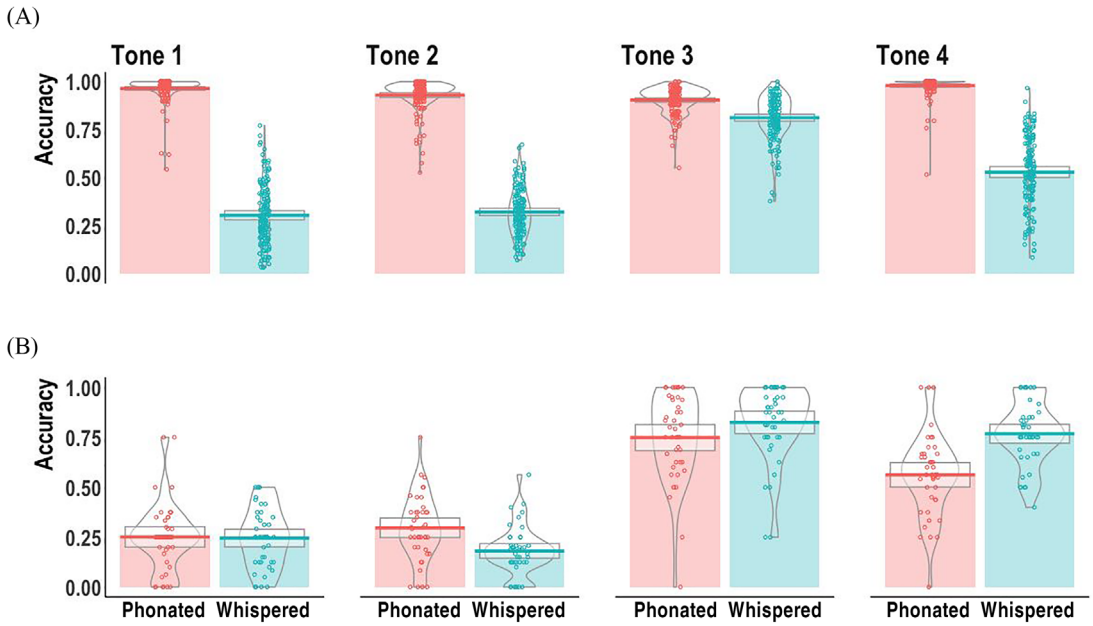

FIG. 3. (Color online) (A) Identification accuracy for phonated and whispered tones with individual data points, density distributions, and means. (B) Identification accuracy of amplitude-modulated noises as a function of tone and phonation with individual data points, density distributions, and means.

above chance level. *Post hoc* comparisons among the four tones showed that all differences reached significance ($p < 0.001$) except for Tone 2–Tone 3 ($p = 0.693$) within phonated speech (Table III) (see supplementary material for all pairwise comparisons).[1] In contrast, within whispered speech, all comparisons were significantly different ($p < 0.001$) except for Tone 1–Tone 2 ($p = 0.434$).

2. Accuracy of amplitude-correlated noises

As shown in Fig. 3(B), the mean accuracies of Tone 1–based noises (phonated: 26.6%, whispered: 25.9%) and Tone 2–based noises (phonated: 31.7%, whispered: 20.5%) were both around chance level, irrespective of phonation. Tone 3- (phonated: 76.7%, whispered: 83.1%) and Tone 4–based noises (phonated: 57.1%, whispered: 75.2%) showed higher accuracies than Tone 1 and Tone 2 (see Table IV and supplementary material for statistical comparisons),[1] suggesting that duration and amplitude of Tone 3 and Tone 4 in speakers' production were more intelligible than those of Tone 1 and Tone 2.

The comparison between noises modeled by phonated speech (hereafter, phonated noises) and whispered speech (hereafter, whispered noises) showed that whispered noises elicited lower accuracy than phonated noises for Tone 2 (ß = –0.657, SE = 0.138, $p < 0.001$). In contrast, whispered noises produced higher accuracies than phonated noises for both Tone 3 (ß = 0.568, SE = 0.169, $p = 0.001$) and Tone 4 (ß = 0.995, SE = 0.136, $p < 0.001$). There was no difference for Tone 1 (ß = –0.039, SE = 0.139, $p = 0.781$). These results indicated that the enhancement of duration and amplitude in Tone 3 and Tone 4 we observed in the production experiment had perceptual significance to listeners.

Finally, since the tone identification of whispered noises outperformed their phonated counterparts for Tone 3 and Tone 4, we further investigated which dimension, amplitude, or duration contributed to the higher accuracy. Results showed that the enhancement information of amplitude had a main effect. Noises modeling whispered amplitude (enhanced) produced higher accuracy than those modeling phonated amplitude (not enhanced) (ß = 0.569, SE = 0.216, $p = 0.009$). There was no main effect of duration enhancement (ß = –0.160, SE = 0.191, $p = 0.402$) or its interaction with amplitude enhancement (ß = 0.139, SE = 0.240, $p = 0.561$). To summarize, the higher accuracy of identifying whispered noises than phonated noises for Tone 3 was primarily attributed to amplitude enhancement.

For Tone 4–based noises, amplitude enhancement had a main effect, with accuracy of noises modeling whispered amplitude (enhanced) higher than those modeling phonated amplitude (not enhanced) (ß = 0.876, SE = 0.147, $p < 0.001$). Once again, duration enhancement did not have a main effect (ß = 0.083, SE = 0.124, $p = 0.501$), nor did it interact with amplitude enhancement (ß = 0.069, SE = 0.191, $p = 0.717$). Thus, the higher accuracy of identifying whispered noises than phonated noises for Tone 4 was primarily attributed to amplitude.

C. Experiment 2: Discussion

Our perception study yielded several interesting findings. First, we found that phonated tones were easier to categorize than whispered tones. This was expected as whispered tones lack the primary cue, F0. Since the participants adjusted their volume between blocks, our results represented the accuracy of identifying phonated and whispered words with the premise that the intensity of the words

was above the auditory threshold, which was sometimes not the case for whispered words in real speech situations. In fact, in real-life communication, listeners also try to increase the volume of whispered speech by approaching speakers' mouth or using hands to contain the sound waves, though the effect might be limited by space or background noise. Second, the overall patterns of accuracy among tones varied across phonated and whispered speech. Phonated Tone 4 was most accurately categorized followed by Tone 1, Tone 2, with Tone 3 least accurately categorized. In contrast, whispered Tone 3 was most accurately categorized, followed by Tone 4, Tone 2, with Tone 1 least accurately categorized. The asymmetry of tone intelligibility among the four whispered tones is consistent with the finding that secondary cues have more perceptual value for Tone 3 and Tone 4 than Tone 1 and Tone 2 (Fu and Zeng, 2000; Liu and Samuel, 2004; Whalen and Xu, 1992). Third, regarding noise identification, the accuracy of Tone 1- and Tone 2–based noises was at chance level for both types of noises, suggesting that amplitude and duration information in the speakers' productions are not sufficient to signal Tone 1 or Tone 2. Tone 3- and Tone 4–based noises were overall quite accurate and slightly more accurate given whispered noises compared to phonated noises. Our final analysis revealed that amplitude helped improve Tone 3 and Tone 4 identity. This indicates that although duration was an effective cue for identification, its enhancement in whispered speech in the production experiment did not make the tone identity more intelligible in the present study.

The combination of the acoustic and perceptual data revealed that speakers enhanced the features of the amplitude and duration in whispered speech for Tone 3 and Tone 4, and the enhancement had perceptual significance for listeners. In contrast, the enhancement made by speakers for whispered Tone 1 and Tone 2 did not have perceptual significance for listeners.

IV. GENERAL DISCUSSION

In the present study, we explored whether speakers enhance secondary amplitude and duration cues for Mandarin tones when F0 is not available in whispered speech. We also examined whether the potential enhancement has perceptual significance to listeners. Novel to our study, we elicited listener-directed speech by encouraging speakers to produce tones intelligible to listeners, using a wide range of phonological contexts and a large number of speakers and listeners.

For duration, the Tone × Phonation interactions showed greater differentiation among tones in whispered speech than phonated speech. The comparison between the phonated and whispered speech offered more details of enhancement: Tone 3 became significantly longer in whispered speech, and Tones 2 and 4 were shorter in whispered speech. This is inconsistent with the results of Jiao and Xu (2019), who found that the magnitude of difference among whispered tones was comparable to phonated tones. The reason for the incongruent findings is likely to lie in the different types of speech elicited in the recording process. In their study, participants read speech rather than produce listener-directed speech. The participants, therefore, had the freedom to read or spontaneously produce speech. In the present study, however, the speakers were invited to make their speech understood by the listeners, i.e., to keep the listeners' needs in mind. They even repeated productions if the experimenter did not correctly perceive the tone. Note that most whispered tones were first repetitions for Tone 3 and Tone 4, whereas second/third repetitions for Tone 1 and Tone 2. Thus, the greater enhancement of Tone 3 and Tone 4 in the present study may reflect the higher communicative demand to clarify a tone identity, as also suggested by Jiao and Xu (2019). The marginal durational adjustments in speakers' second/third repetitions for Tone 2 indicates the speakers' attempt to make their productions clearer after hearing the experimenter's feedback. The combination of Jiao and Xu (2019), Liu and Samuel (2004), and the present study suggests that speakers have the potential to enhance multiple acoustic dimensions contributing to lexical tone when communication demands it, but they may not do this regularly in all whispered situations, perhaps especially those that do not demand communication directed to a real listener.

The observation of the duration pattern among the four tones replicates Liu and Samuel (2004) but differs from Jiao and Xu (2019), who found Tone 3 was longer than the other three tones, but the other three tones were comparable. This difference might be caused by the superimposition of tone and intonation on the utterances in Jiao and Xu (2019), who asked the speakers to produce a tone as a statement or a question. In other words, each utterance carried both tone and intonation functions. Intonation has been reported to affect both the F0 register and duration of Mandarin words (Wang and Xu, 2011; Wang *et al.*, 2017; Wang *et al.*, 2018; Yang and Yang, 2012). Therefore, the intonation the utterances carried may have altered the speakers' emphasis on the tones' category, which was the only thing being emphasized and focused on by speakers in Liu and Samuel (2004) and the present study. Since we did not include intonation as a variable, the possible influence of intonation needs to be clarified in future studies (cf. Ouyang and Kaiser, 2015).

In addition to duration, speakers also adjusted their amplitude contours given both tone type and phonation type. Moreover, speakers enhanced the difference in amplitude contours among tones in whispered speech than phonated speech. Visual inspection revealed that the amplitude contour of Tone 3 and Tone 4 resembled the canonical F0 contours of the corresponding tone categories for both speech modes, which were dipping and falling, respectively. In contrast, the amplitude contours of Tone 1 and Tone 2 did not imitate the F0 contours, which should have been level and rising. Instead, the amplitude contours of Tone 1 and Tone 2 were falling, despite adjustment of slope and curvature across speech modes.

In the perceptual experiment investigating whether the enhancement made by the speakers had perceptual

significance to listeners, we observed mixed results. Specifically, the accuracy of identifying whispered noises was higher than that of their phonated counterparts for Tone 3 and Tone 4. The accuracy for Tone 1 and Tone 2, however, did not show improvement in whispered noises compared to phonated noises; for Tone 1 and Tone 2, all four noises were identified at or near chance. This suggested that the greater enhancement of Tone 3 and Tone 4 in whispered speech had perceptual significance, demonstrating the interaction between the speaker and listener. Indeed, this is what we found in the comparison between the phonated and whispered results: listeners showed significant differences in Tone 3 and Tone 4 categorization accuracy given the phonation type.

In contrast with whispered Tone 3 and Tone 4, the enhancements in whispered Tone 1 and Tone 2 were not enough to make the tone features stand out. Unexpectedly, the enhancement of whispered Tone 2 reduced the identification to around chance level. Our perceptual data echoed the production data in that it was harder to produce an intelligible whispered Tone 1 and Tone 2 than Tone 3 and Tone 4 without the primary F0 cue. Note that most whispered tones were first repetitions for Tone 3 and Tone 4, whereas second/third repetitions for Tone 1 and Tone 2. Our results, therefore, suggest that the listener-directed speech contains sufficient information for whispered Tone 3 and Tone 4, but not for whispered Tone 1 or Tone 2, even with listener's feedback as guidance.

On the one hand, speakers did not consistently enhance the duration features of Tone 1, as we found in the production experiment. This is in line with the previous literature showing that Tone 1 and Tone 2 do not exhibit a consistent and marked pattern in duration (Chao, 1965; Fu and Zeng, 2000; Howie, 1976; Tseng, 1990). When needed, speakers can lengthen Tone 1 in whispered speech, and listeners can use duration as a perceptual cue (e.g., Liu and Samuel, 2004). However, our results from 30 speakers suggest that in listener-directed speech, only Tone 3 and Tone 4 secondary cues become enhanced.

On the other hand, the enhancement of amplitude contour in whispered tones was not effective in improving the tone identity. The high resemblance of amplitude and F0 contours was thought to contribute to the improved identification accuracy (Fu and Zeng, 2000; Ho, 1976; Whalen and Xu, 1992). In the present study, however, the amplitude contours of Tone 1 and Tone 2 did not resemble their corresponding F0 contours. It is possible that enhancing amplitude in whispered speech, particularly for Tone 1 and Tone 2, is somehow constrained by the mouth closing since airflow is further modulated by the production of syllables, that is, through the opening/closing of the mouth. Based on the association with F0 contour, the amplitude contours of Tone 1 and Tone 2 should have been high level and rising. However, it was hard for speakers to keep a constantly strong airflow throughout the syllable for Tone 1 or strengthen airflow across time from an already mid-high level for Tone 2. The production experiment showed a flat, falling slope after the peak in the amplitude of whispered Tone 1 and Tone 2. The adjustments, however, did not improve perceptual accuracy. Our tentative interpretation is that speakers enhanced the amplitude contours of whispered tones, but this type of information does not help the listener decide on tone identity, or the mouth opening/closing hinders the speakers in their efforts to use amplitude to imitate F0 trajectories.

Finally, data on the perceptual accuracy of amplitude-enhanced and duration-enhanced noises showed that only the enhancement of amplitude (and its combination with duration) improved the accuracy, though duration was an effective cue for Tone 3 and Tone 4 when F0 was absent. This is consistent with the findings that listeners relied more on amplitude than duration in tone identification (Fu and Zeng, 2000; Whalen and Xu, 1992). We note that our noise-modulated results do not necessarily mean that duration is not an effect or useful secondary cue in natural speech. Clearly, our speakers and listeners made use of duration to varying degrees of effectiveness. The fact that Tone 3 and Tone 4 showed consistent durational differences (and exaggerations) indicates that duration plays a vital role for these two tones in line with Liu and Samuel (2004).

Overall, our results revealed a greater enhancement of amplitude and duration when the primary cue F0 was unavailable, which is consistent with what Liu and Samuel (2004) suggested based on their perceptual experiment. However, we also observed that the greater enhancement for the four tones was asymmetric. This again extends Liu and Samuel (2004)'s results and indicates that the characteristics of each tone category affect how well secondary cues can be enhanced. Taken together, the current study sheds light on several issues of the production and perception of Mandarin tones. First, speakers differentiate amplitude and duration more when F0 is unavailable than when available. The adjustment suggested the speakers' awareness of the association between tone categories and these two dimensions and their efforts to find alternatives to compensate for the loss of information to satisfy the communicative purpose. Second, the more accurate tone categorization of the enhanced secondary cues indicated the compensatory behavior affected both speakers and listeners. They collaborated in preserving tone identity when the primary cue F0 was not available. However, as we previously noted, listeners were not sensitive to the enhancement of duration made by speakers in the present study. Third, the effort in enhancements was at the same time constrained by the saliency of the feature a dimension has (i.e., whether a tone category is marked by long or short in duration) and/or the physiological limits of speech production (i.e., whether it is hard to realize). Consequently, speakers will not show consistent enhancement for tones without distinctive features, such as the duration for Tone 1. Despite greater enhancement, speakers are still likely to undershoot because they do not have a powerful control of airflow during exhalation. It is worth mentioning here that our recording procedure produced more repetitions for whispered Tone 1 and whispered Tone 2 than

whispered Tone 3 and Tone 4 due to the experimenter's misidentification of the first productions. These second (and third) repetitions served as the tokens for the acoustic analysis and perceptual study. It is possible that the initially misunderstood utterances involved enhancements, such as an exaggerated duration for Tone 1. Our findings suggest that speakers most likely undershot realizing whispered Tone 1 and Tone 2, even when they repeated productions according to listener's feedback.

To summarize, we found in the production experiment that speakers may attempt to dynamically enhance the duration and amplitude contours of Mandarin tones when they whisper. This enhancement has perceptual significance, demonstrating speaker–listener coordination. However, not all enhancement has perceptual significance. As a result, comprehension of Tone 1 and Tone 2 suffered more in whispers than comprehension of Tone 3 and Tone 4. Several limitations are calling for future studies. First, since our experiments were mostly done remotely via the participants' computers and phones, our results represent the situations where speakers and listeners are communicating online via electronic devices. Generalization to other communicative situations entails the designing of the respective recording setup. Moreover, because the present study only focused on non-spectral cues, spectral cues need investigation in future studies as their role for Mandarin tone identification and the interaction between speakers and listeners on their enhancement in whispered speech are theoretically important but remain unclear.

ACKNOWLEDGMENTS

This research was supported in part by the National Institutes of Health (Grant No. 1R03HD099382-01). We are grateful to Dr. Ewa Jacewicz, Dr. Arthur Samuel, and one anonymous reviewer for their valuable comments, engagement, and time committed to this paper.

[1]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0009378 for a list of the stimuli with a representative Chinese character, Pinyin romanization, and IPA; the wav recordings; the result of all *post hoc* comparisons; coefficients of interactions; all Praat scripts; the visual comparison of lab-recording and remote recording; and outputs of *post hoc* analyses obtained by changing the reference levels; for all data and R code detailing our statistical analyses.

[2]To make sure that the recordings were of sufficient quality for acoustic measurements, we attempted to strictly control the recording environment and recording procedure. First, we increased the number of participants to 30, which exceeded our initial power analysis (N = 1) and ensured that our sample was robust. Before recording, we instructed the participants to sit in a quiet room with all the windows closed and send a practice recording to the experimenter. The experimenter checked that the file was auditorily intelligible and that there was no obvious background noise, such as air-conditioning, wind, ambient noise, and so on. Then. the participants were asked to position their mouths approximately 20 cm from the microphone and keep still during the recording. In addition, we visually compared the lab recordings (participant: N = 3) to the remote recordings (participant: N = 27) in terms of amplitude contour shape and the first formant (F1). The measurement of F1 was the basis for marking the syllable onset and offset and thus, was the premise for duration analysis. (See supplementary material for the visual comparison of lab recording and remote recording.)[1] The comparison confirmed very similar amplitude contour shapes and F1 values between lab recording and remote recording. This suggests that remote recording did not result in false tracking of the measurements that we focused on, namely amplitude or F1.

[3]There was a minor difference in the way feedback was given in the lab recording versus remote recording. Specifically, the experimenter sat outside the sound booth with her back facing the participant and used gestures to indicate the tone for the lab recording, whereas the experimenter talked with the participant via audio conference (WeChat audio conference) and responded orally with the tone for the remote recording. To avoid the head movement, the participants were asked to hold the device for audio conference to the left/right of their mouth and keep still during the recording.

[4]Specifically, the experimenter first demonstrated the whisper and explained the mechanism by saying that the vocal fold would vibrate in phonated speech but would not in whispered speech and that the speaker could feel the difference on the front of the neck. She then mentioned two situations where whispering often occurs: within a library and the bedroom where a baby was sleeping.

[5]Specifically, the two native speakers of Mandarin Chinese listened to the second and third repetitions and wrote down which repetition was the most intelligible one. When there was disagreement, they made a final decision through discussion.

[6]There are at most three repetitions for each utterance since there are only four possible answers. If the third repetition was still misidentified, the answer could be inferred without a fourth repetition.

[7]Since all stops were at the onset of each word, the silence before the stop could not be measured. Therefore, we systematically underestimated stop duration for all the recorded utterances.

[8]We note that data collection of auditory experiments using web platforms allows experimenters to not only collect a large amount of data in a short period of time, but also collect data from a larger sample beyond the typical sample only available to lab-based researchers, such as young adult college students. Because participants may get distracted during web-based studies, we made the following efforts to ensure the reliability of our data: First, we increased the number of participants to 188 for greater statistical power. Second, before the experiment, we asked the participants to sit in a quiet room and turn off all other sounds from other sources, such as music and message/system alerts. Third, the experiment was done by appointment and the participants were informed that the experimenter was simultaneously supervising their progress and response data. Fourth, the use of headphones was required (and confirmed) during the experiment. Finally, data inspection was conducted following the experiment to remove unreliable data. Consequently, 27 subjects were removed for not finishing the experiment or for scoring less than 50% in the phonated trials, which should have been very easy for native listeners.

Abramson, A. S. (**1972**). "Tonal experiments with whispered Thai," Papers linguistics phonetics to memory Pierre Delattre **54**, 31–44.

Bates, D., Maechler, M., Bolker, B., Walker, S., Haubo Bojesen Christensen, R., *et al.*, (**2015**). "lme4: Linear mixed-effects models using eigen and s4. r package version 1.1–7. 2014," arXiv:1406.5823 (2015)

Best, C. T., Morrongiello, B., and Robson, R. (**1981**). "Perceptual equivalence of acoustic cues in speech and nonspeech perception," Percept. Psychophys. **29**(3), 191–211.

Blicher, D. L., Diehl, R. L., and Cohen, L. B. (**1990**). "Effects of syllable duration on the perception of the mandarin tone 2/tone 3 distinction: Evidence of auditory enhancement," J. Phon. **18**(1), 37–49.

Boersma, P., and Weenink, D. (**2018**). "Praat: Doing phonetics by computer (version 6.0.37) [computer program]," http://www.praat.org (Last viewed March 14, 2018).

Brown-Schmidt, S. (**2005**). "Language processing in conversation," Ph.D. dissertation, University of Rochester, Rochester, NY.

Buxó-Lugo, A., Toscano, J. C., and Watson, D. G. (**2018**). "Effects of participant engagement on prosodic prominence," Discourse Process. **55**(3), 305–323.

Buz, E., Tanenhaus, M. K., and Jaeger, T. F. (**2016**). "Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations," J. Mem. Lang. **89**, 68–86.

Cai, Q., and Brysbaert, M. (**2010**). "Subtlex-ch: Chinese word and character frequencies based on film subtitles," PloS One **5**(6), e10729.

Chao, Y. R. (**1943**). "Languages and dialects in china," Geogr. J **102**(2), 63–66.

Chao, Y. R. (**1965**). *A Grammar of Spoken Chinese* (Univ of California Press, Berkeley).
Fu, Q.-J., and Zeng, F.-G. (**2000**). "Identification of temporal envelope cues in chinese tone recognition," Speech, Lang. Hear. **5**(1), 45–57.
Gandour, J. (**1983**). "Tone perception in far eastern languages," J. Phon. **11**(2), 149–175.
Heeren, W. F. (**2015a**). "Vocalic correlates of pitch in whispered versus normal speech," J. Acoust. Soc. Am. **138**(6), 3800–3810.
Heeren, W. F. (**2015b**). "Coding pitch differences in voiceless fricatives: Whispered relative to normal speech," J. Acoust. Soc. Am. **138**(6), 3427–3438.
Heeren, W. F., and Lorenzi, C. (**2014**). "Perception of prosody in normal and whispered french," J. Acoust. Soc. Am. **135**(4), 2026–2040.
Heeren, W. F. L., and Van Heuven, V. J. (**2009**). "Perception and production of boundary tones in whispered dutch," in *Proceedings of Interspeech, September 6-10, 2009, Brighton, UK*, pp. 2411–2414.
Heeren, W. F. L., and Van Heuven, V. J. (**2011**). "Acoustics of whispered boundary tones: Effects of vowel type and tonal crowding," in *Proceedings of the 17th International Congress of Phonetic Sciences*, edited by W.-S. Lee and E. Zee and (City University of Hong Kong, Hong Kong), pp. 851–854.
Higashikawa, M., and Minifie, F. D. (**1999**). "Acoustical-perceptual correlates of' whisper pitch' in synthetically generated vowels," J. Speech Lang. Hear. Res. **42**(3), 583–591.
Ho, A. T. (**1976**). "The acoustic variation of mandarin tones," Phonetica **33**(5), 353–367.
Holt, L. L., and Lotto, A. J. (**2006**). "Cue weighting in auditory categorization: Implications for first and second language acquisition," J. Acoust. Soc. Am. **119**(5), 3059–3071.
Howie, J. M. (**1976**). *Acoustical Studies of Mandarin Vowels and Tones*, vol. **18** (Cambridge University Press, London).
Idemaru, K., and Holt, L. L. (**2011**). "Word recognition reflects dimension-based statistical learning," J. Exp. Psychol.-Hum. Percept. Perform. **37**(6), 1939–1956.
Idemaru, K., and Holt, L. L. (**2020**). "Generalization of dimension-based statistical learning," Atten. Percept. Psychophys. **82**, 1744–1762.
Jiao, L., and Xu, Y. (**2019**). "Whispered mandarin has no production-enhanced cues for tone and intonation," Lingua **218**, 24–37.
Kong, Y.-Y., and Zeng, F.-G. (**2006**). "Temporal and spectral cues in mandarin tone recognition," J. Acoust. Soc. Am. **120**(5), 2830–2840.
Konno, H., Kanemitsu, H., Toyama, J., and Shimbo, M. (**2006**). "Spectral properties of Japanese whispered vowels referred to pitch," J. Acoust. Soc. Am. **120**(5), 3378–3378.
Laan, G. P. (**1992**). "Perceptual differences between spontaneous and read aloud speech," in *Proc. of the Institute of Phonetic Sciences Amsterdam*, Vol. 16, pp. 65–79.
Lehet, M., and Holt, L. L. (**2016**). "Adaptation to accent is proportionate to the prevalence of accented speech," J. Acoust. Soc. Am. **139**(4), 2164–2164.
Lin, M. (**1988**). "Putonghua shengdiao de shengxue texing he zhijue zhengzhao [the acoustic characteristics and perceptual cues of tones in standard chinese]," J. Chin. Linguist. **204**(3), 182–193.
Linguistic Institute of Chinese Social Academy. (**1982**). *Modern Chinese Dictionary* (Commercial Press, Beijing).
Liu, S., and Samuel, A. G. (**2004**). "Perception of mandarin lexical tones when f0 information is neutralized," Lang. Speech **47**(2), 109–138.
Llanos, F., Dmitrieva, O., Shultz, A., and Francis, A. L. (**2013**). "Auditory enhancement and second language experience in spanish and english weighting of secondary voicing cues," J. Acoust. Soc. Am. **134**(3), 2213–2224.
Mirman, D. (**2014**). "Growth curve analysis: A hands-on tutorial on using multilevel regression to analyze time course data," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 36, pp. 51–52.
Nakamura, M., Iwano, K., and Furui, S. (**2008**). "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance," Comput. Speech Lang. **22**(2), 171–184.
Norman, J. (**1988**). *Chinese* (Cambridge University Press, London).
Ouyang Iris, C., and Kaiser, E. (**2015**). "Prosody and information structure in a tone language: An investigation of Mandarin Chinese," Lang. Cogn. Neurosci. **30**(1-2), 57–72.
Pardo, J. (**2013**). "Measuring phonetic convergence in speech production," Front. Psychol. **4**, 559.
Qualtrics, L. (**2014**). "Qualtrics (version july, 2020) [software]," Provo, Utah (Last viewed: 7/27/20).
Samuel, A. G., and Troicki, M. (**1998**). "Articulation quality is inversely related to redundancy when children or adults have verbal control," J. Mem. Lang. **39**(2), 175–194.
Sancier, M. L., and Fowler, C. A. (**1997**). "Gestural drift in a bilingual speaker of brazilian portuguese and english," J. Phon. **25**(4), 421–436.
Schober, M. F., and Clark, H. H. (**1989**). "Understanding by addressees and overhearers," Cogn. Psychol. **21**(2), 211–232.
Stevens, K. N., and Klatt, D. H. (**1974**). "Role of formant transitions in the voiced-voiceless distinction for stops," J. Acoust. Soc. Am. **55**(3), 653–659.
Tilsen, S. (**2009**). "Subphonemic and cross-phonemic priming in vowel shadowing: Evidence for the involvement of exemplars in production," J. Phon. **37**(3), 276–296.
Tseng, C-y. (**1990**). "An acoustic phonetic study on tones in Mandarin Chinese," Ph.D. dissertation, Brown University, Providence, RI.
Wang, B., and Xu, Y. (**2011**). "Differential prosodic encoding of topic and focus in sentence-initial position in mandarin chinese," J. Phon. **39**(4), 595–611.
Wang, B., Xu, Y., and Ding, Q. (**2017**). "Interactive prosodic marking of focus, boundary and newness in mandarin," Phonetica **75**(1), 24–56.
Wang, Y.-T., Green, J. R., Nip, I. S., Kent, R. D., and Kent, J. F. (**2010**). "Breath group analysis for reading and spontaneous speech in healthy adults," Folia Phoniatr. Logop. **62**(6), 297–302.
Wang, B., Kügler, F., and Genzel, S. (**2018**). "Downstep effect and the interaction with focus and prosodic boundary in Mandarin Chinese," in *Proceedings of the 6th International Symposium on Tonal Aspects of Languages (TAL 2018)*, pp. 22–26.
Whalen, D. H., and Xu, Y. (**1992**). "Information for mandarin tones in the amplitude contour and in brief segments," Phonetica **49**(1), 25–47.
Yang, X., and Yang, Y. (**2012**). "Effects of topic structure and syntax on boundary pitch variations in standard chinese," in *Speech Prosody 2012*, Shanghai, China, pp. 543–546.
Zhang, X., and Holt, L. L. (**2018**). "Simultaneous tracking of coevolving distributional regularities in speech," J. Exp. Psychol.-Hum. Percept. Perform. **44**(11), 1760.
Zhang, H., Holt, L. L., and Wiener, S. (**2022**). Dynamic adjustment of cue weighting in speech. osf.io/mw8c7 (Last viewed: 1/7/22).
Żygis, M., Pape, D., Koenig, L. L., Jaskuła, M., and Jesus, L. M. (**2017**). "Segmental cues to intonation of statements and polar questions in whispered, semi-whispered and normal speech modes," J. Phon. **63**, 53–74.