JASA ARTICLE

The contrast between clear and plain speaking style for Mandarin tones

Paul Tupper,[1,a)] Keith W. Leung,[2] Yue Wang,[2] Allard Jongman,[3] and Joan A. Sereno[3]
[1]*Department of Mathematics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada*
[2]*Department of Linguistics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada*
[3]*Department of Linguistics, University of Kansas, Lawrence, Kansas 66045, USA*

ABSTRACT:
We examine the acoustic characteristics of clear and plain conversational productions of Mandarin tones. Twenty-one native Mandarin speakers were asked to produce a selection of Mandarin words in both plain and clear speaking styles. Several tokens were gathered for each of the four tones giving a total of 2045 productions. Six critical tonal cues were computed for each production: fundamental frequency (*F*0) mean, slope, and second derivative, duration, mean intensity, and a binary variable coding whether the production involved creaky voice. A linear mixed-effects regression model was used to explore how these cues changed with respect to the clear versus plain distinction for each tone, with speaking style as the fixed effect and speaker being a random effect. The strongest effects detected were that duration and mean intensity increased in clear speech across speakers and tones. Tones 2 and 3 increased in mean *F*0 and Tone 4 increased its slope. An additional finding was that, for contour tones, speakers accomplished the increase in duration by stretching out the tone contours in time while largely not changing the *F*0 range. These results are discussed in terms of signal-based (affecting all tones) and code-based (enhancing contrast between tones) change. © *2021 Acoustical Society of America*.
https://doi.org/10.1121/10.0009142
(Received 26 March 2021; revised 23 November 2021; accepted 3 December 2021; published online 27 December 2021)
[Editor: Ewa Jacewicz] Pages: 4464–4473

I. INTRODUCTION

A. Background

We modify our speech styles to suit different communicative contexts. In adverse listening environments (e.g., background noise) or to accommodate specific target audiences (e.g., hearing-impaired or non-native listeners), we use a clarified speaking style, known as "clear speech," relative to plain, conversational style, in order to enhance intelligibility (Summers *et al.*, 1988).

Such clear speech involves a more enunciated speaking manner, e.g., with increased duration, fundamental frequency (*F*0), and intensity, as well as more exaggerated and more dynamic temporal and spectral changes resulting from hyperarticulation (Leung *et al.*, 2016; Maniwa *et al.*, 2009). These characteristics are shared by numerous audience- and environment-appropriate styles, such as infant-directed speech (IDS), foreigner-directed speech, pet-directed speech, and Lombard speech (Grieser and Kuhl, 1988; Han *et al.*, 2019; Smiljanić and Bradlow, 2009; Uther *et al.*, 2007; Xu *et al.*, 2013). In this paper, we present our findings on the clear-speech changes in the acoustic characteristics of Mandarin tone productions in terms of signal-based and code-based modifications.

B. Signal-based and code-based modifications

Articulatory and acoustic modifications in these speech styles may serve different communicative functions. For instance, an overall increase in duration, *F*0, or intensity has been found to be associated with heightened attentional and affective components with the aim of gaining the audience's attention or conveying positive affect (Kuhl *et al.*, 1997; Uther *et al.*, 2007; Xu *et al.*, 2013); on the other hand, hyperarticulation, which enhances contrastivitiy of speech sound categories, such as expansion of vowel space, serves linguistic and didactic functions to aid speech intelligibility (Burnham *et al.*, 2010; Ferguson and Kewley-Port, 2002, 2007; Kuhl *et al.*, 1997; Xu *et al.*, 2013).

Thus, clear speech can be seen to involve two levels of modifications (Bradlow and Bent, 2002; Redmon *et al.*, 2020; Zhao and Jurafsky, 2009). The first is *signal-based*, changes that apply to the entire speech signal itself and are not dependent on properties of the language, which essentially serves attentional and affective functions. These may include increased overall duration, *F*0, and intensity, resulting in enhancement of the saliency of the speech signal rather than distinctions of specific speech sounds. The second is *code-based*, changes that enhance linguistic contrasts, and are therefore phoneme-specific. An example would be increased *F*2 for front vowels and decreased *F*2 for back vowels (Leung *et al.*, 2016). As such, code-based modifications are critical to distinguishing one word from another

[a)]Electronic mail: pft3@sfu.ca, ORCID: 0000-0002-4340-4481.

and thus enhance speech intelligibility (Baese-Berk and Goldrick, 2009; Wedel *et al.*, 2018). It should be noted that changes in the same acoustic cues may entail either signal-based or code-based modifications, or both, depending on language or speech context. For example, increased duration across vowels is generally considered a signal-based change. However, greater clear-speech lengthening of tense compared to lax vowels in English would result in enhanced contrasts between tense and lax vowels and is thus considered a code-based modification (Leung *et al.*, 2016). Likewise, an increase in *F*1 may be either a signal-based change across vowels in Lombard speech or a code-based change specifically for low vowels (Junqua, 1996; Tang *et al.*, 2017).

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). Such modification may be challenging in cases where cues that are modified in signal-based changes such as *F*0 also serve code-based functions, as in the case of lexical tone. As such, lexical tone provides a unique platform for testing the clear-speech principles with respect to the extent to which signal- and code-based cues are adopted in clear-speech modifications. While substantial research focuses on clear-speech perception at the segmental level, little attention has been paid to lexical tone. Previous research (Wong *et al.*, 2017; Yang, 2015) including our own study (Tupper *et al.*, 2020) has shown that both general acoustic cues (*F*0, duration, and intensity) that are universal across tones and critical tonal cues (mean *F*0, *F*0 slope, *F*0 second derivative) that are specific to individual tones are relevant. While little research has examined clear-speech tone, studies on tone production in IDS, Lombard speech, and emphatic speech may lend some references.

First, findings have shown signal-based tone-universal modifications in hyperarticulated tones. For example, Cantonese and Mandarin tones produced in noise (Lombard speech) exhibit increased *F*0 compared to tones produced in quiet across all tones (Tang *et al.*, 2017; Zhao and Jurafsky, 2009). Likewise, all hyperarticulated tones in Cantonese or Mandarin infant-directed relative to adult-directed speech appear to be indexed by higher *F*0 and longer duration (Grieser and Kuhl, 1988; Liu *et al.*, 2007; Xu Rattanasone *et al.*, 2013). Moreover, in teaching (language instruction setting) compared to natural style, Mandarin tone production exhibited an overall expanded *F*0 range (Papoušek and Hwang, 1991), and longer average duration (Han *et al.*, 2019). Consistently, research on sentential-level emphasis in Mandarin tone productions (Chen and Gussenhoven, 2008) reveals that in a focused position compared to non-focused position, tone-bearing syllables show a systematic increase in duration, suggesting signal-based changes serving prosodic functions. In contrast, *F*0 variation, which is critical in tonal category distinctions, is shown to be restricted in conveying such prosodic information.

On the other hand, research also demonstrates code-based tone hyperarticulation. IDS studies have generally revealed expanded *F*0 range (Grieser and Kuhl, 1988; Liu *et al.*, 2007) and expanded tone space (defined by the area formed by *F*0 onset/offset plots of different tones, Xu Rattanasone *et al.*, 2013), reflecting increased tonal category contrasts. Likewise, tone-specific expansion of *F*0 range was identified for emphatic relative to non-emphatic sentences, with the (Mandarin) falling tone expanding more than the rising and level tones(Chen and Gussenhoven, 2008), showing that the direction and nature of modifications are aligned with the intrinsic features of these tones. As well, changes in *F*0 contours were found to be adapted to neighboring tones to maximize distinctions between tonal categories (Chen and Gussenhoven, 2008). Code-based modifications have also been identified when tonal hyperarticulation interacts with other linguistic domains. In particular, Xu and Burnham (2010) show that *F*0 modifications in hyperarticulated Cantonese tones and intonation appear to be modulated independently such that enhanced category distinctions among tones are not affected by exaggerated intonation. Further, modifications of Mandarin tones in IDS have been found to interact with prosodic focus, showing expanded tone space for tones occurring in the utterance-final but not utterance-medial positions (Tang *et al.*, 2017). Additionally, the extent of *F*0 modifications for individual tones tends to differ as a function of lexical frequency, with low-pitched tones involving heightened and more dispersed *F*0 in low-frequency words (Zhao and Jurafsky, 2009).

Apart from *F*0 and duration, non-modal, creaky phonation also potentially displays both signal- and code-based modifications. Previous research (Kuang, 2017) demonstrated that non-modal phonation covaried with *F*0 and was *not* a tone-specific feature for Mandarin. In addition, creaky phonation can be found in all Mandarin tones (Huang *et al.*, 2018). Therefore, any modifications with creaky phonation can be regarded as signal-based. However, creakiness is sometimes produced with certain tone categories only. For instance, Kuang (2017) only found creaky phonation in Mandarin Tone 3 (dipping) and 4 (falling), which involved a low phonetic pitch target. The two tones were produced with more creakiness when *F*0 range was lowered. Since tone hyperarticulation involves an expanded *F*0 range (Papoušek and Hwang, 1991), it should potentially lead to an increase in creaky phonation for tones with low pitch targets only (i.e., Tone 3 and 4). Consequently, it increases the contrast between tone categories with high and low pitch targets and indicates a code-based modification. In addition, for tone productions at the sentential level, the *F*0 lowering in emphasis position led to increased creakiness for low tone productions (Chen and Gussenhoven, 2008). The amount of creaky voice increased in focus position compared to non-focus position, along with an increase in *F*0, and occurred for Tone 4 only (Huang *et al.*, 2018). As a result, creaky phonation modification either enhances high-low pitch contrast (Chen and Gussenhoven, 2008) or the

contrast between Tone 4 and other tone categories (Huang *et al.*, 2018) and therefore is regarded as code-based.

Although these studies have found both signal- and code-based evidence of hyperarticulation in tone production, no research has systematically (and separately) examined clear-speech features of individual tones. Presumably, if clear speech is intended to make the signal more salient, we would expect signal-based modifications of the same acoustic features across all tones. Based on similar findings from IDS studies, signal-based modifications may involve overall increase in *F*0, intensity, and duration across tones. Alternatively, if clear speech is to enhance tone category distinctions, we would expect code-based modifications where features that distinguish tones would be modified so as to increase contrast. Specifically, we expect code-based enhancement to be aligned with individual tone characteristics; for example, with high tone being higher and low tone being lower in clear compared to plain speech.

C. The present study

In the present study, "clear speech" is defined in the following contexts:

As reviewed, a variety of clear speech style modifications (e.g., IDS, foreign-directed speech, pet-directed speech, Lombard speech) share similar characteristics, and studies on these speech styles have addressed similar signal- and code-based strategies to enhance and preserve phonemic category distinctions (Moon and Lindblom, 1994; Ohala, 1995; Smiljanić, 2021; Smiljanić and Bradlow, 2009; Tupper *et al.*, 2018). Therefore, we place the current research in a broad context, relating it to previous findings across speech styles.

Clear-speech elicitation instructions in previous studies vary extensively, including "speak clearly," "hyperarticulate," "speak to a nonnative speaker," or "speak to someone who is hearing impaired" (e.g., Ferguson and Kewley-Port, 2007; Lam *et al.*, 2012; Moon and Lindblom, 1994). In this study, we adopt a previously established procedure (Burnham *et al.*, 2010; Leung *et al.*, 2016), involving clear-speech productions in response to a simulated automatic speech recognition program (see Sec. II for details). Clear speech elicited as such can be defined as "corrective" clear speech (cf. Chen and Gussenhoven, 2008) and involves similar adjustments compared to human-oriented hyperarticulated speech (Burnham *et al.*, 2010). However, given that different types of instructions may still impact the magnitude of clear-speech modifications (Lam *et al.*, 2012), the current findings are interpreted within the scope of the type of clear speech used in this study.

We take productions from 21 native Mandarin speakers of both clear and plain words containing all four tones of Mandarin. We systematically examine which acoustic features characterize clear-speech tones. We predict that both types of modification will be apparent to some degree, and explore their interplay. We expect that universal signal-based clear-speech tone attributes, including general features such as overall *F*0, duration, and intensity, will increase across all tones (e.g., proportional lengthening of tones in clear relative to plain styles). On the other hand, code-based tone modifications will involve changes that not only preserve tone-intrinsic properties but also increase tonal category distinctions (e.g., high tones getting higher while low tones are getting lower, an increase in creakiness for Tone 3 and 4).

In addition to examining how cues of the four tones are modified in the change from plain to clear speech, our study provides an opportunity to study how increased duration is implemented in a clear speech style change. We can imagine two possible ways to implement increased duration for tone contours. One, GoSlower, simply stretches out the tone contour without modifying the tone values attained. The other, DoLonger, maintains the same rates of change of pitch during enunciation but just does them for longer, thereby expanding the range of pitch values attained for contour tones. We investigate which is a better model of how duration increases are implemented, and discuss how this relates to signal-based versus code-based changes.

II. METHODS

A. Participants

The participants were 21 native Mandarin speakers (11 Female, 10 Male) who were raised in Northern China or Taiwan during the first 12 years of life (aged 18–28, mean 22.6). Although Mandarin spoken in Northern China and Taiwan differs in terms of vowel space, pitch range, and mean pitch (Shi and Deng, 2006), both groups of native Mandarin speakers were recruited because previous studies showed similar modifications for the hyperarticulation of Mandarin tones in IDS and Lombard speech (e.g., increased mean *F*0 and *F*0 range) (Liu *et al.*, 2007; Tang *et al.*, 2017). They were recruited from the undergraduate and graduate population at Simon Fraser University and indicated that standard Mandarin was their native and dominant language. They reported normal hearing and no history of speech or language disorders.

B. Materials

The participants produced the monosyllable /ɤ/ with four Mandarin tones in plain and clear speech. The four tones are real Mandarin words with the meaning of "graceful" (/ɤ1/), "goose" (/ɤ2/), "nauseous" (/ɤ3/), and "hungry" (/ɤ4/). The monosyllables /i/ and /u/ were used as fillers. Only /ɤ/ items were analyzed because it is a mid-central vowel and the production involves the least tongue movement among the three vowels. Presumably, it has the least interaction with the larynx that may influence tone production.

C. Procedures

The participants produced the speech materials in a sound-attenuating booth in the Language and Brain

https://doi.org/10.1121/10.0009142

Laboratory at Simon Fraser University. The recordings were conducted digitally using Sonic Foundry Sound Forge 6.4 at a sampling rate of 48 kHz. A Shure KSM microphone was placed at a 45 degree angle, about 20 cm away from the speaker's mouth. Prompts, instructions, and feedback were displayed on a computer screen.

The elicitation sessions followed the procedures developed by Maniwa *et al.* (2009) and Leung *et al.* (2016). Participants were told that we were testing a speech recognition computer program, which was actually a simulated interactive computer software that seemingly attempted to perceive and recognize the tokens produced by a speaker, developed using MATLAB (MATLAB, 2013). Participants were instructed to speak naturally first, as if in casual conversation, when a prompt showed up on the screen. Then, the program would "guess" and indicate on the screen what they produced. The participant would then indicate whether the guess was correct by clicking a box on the screen. If the guess was considered correct, the program would move on to the next stimulus. Otherwise, the program would instruct the participant to repeat the stimulus as clearly as possible. In the acoustic analyses of such "incorrect guess" trials, the productions in response to the initial prompts served as the "plain speech," whereas the repeated productions were the "clear speech." It was possible that participants would be tempted to produce the items in a clear, enunciated speaking style whenever they saw a prompt, so as to avoid being "corrected" by the computer program. To ensure distinct productions of plain versus clear speech, participants were, therefore, reminded that they should always return to their habitual speaking style, as if in casual conversation, at the beginning of each trial in order to test the computer program's ability to recognize plain as well as clear speech. Prior to the elicitation session, participants were familiarized with this task as well as the two speaking styles in a short practice session. During the practice session, we ensured that participants were able to produce a plain-clear distinction during the task.

Each elicitation session contained 180 trials described above in total, and there were 25 trials in which the "guess" was correct. Among the remaining trials, a total of 98 /ɤ/ productions were obtained in 49 elicitation trials, i.e., 49 pairs of plain-clear speech items ([11 (/ɤ1/) + 12 (/ɤ2/) + 15 (/ɤ3/) + 11 (/ɤ4/)] X 2 styles). The prompts were presented in three blocks (15 randomly selected trials in the first block and 17 each in the other two) and speakers took a 3-min break after each block. The /i/ and /u/ words were mixed with /ɤ/ words in the recordings. There were 110 productions of /i/ (55 trials) and 102 productions of /u/ (51 trials). The order of prompts and responses was the same for each participant.

Each speaker's productions were evaluated by two phonetically trained native Mandarin evaluators in a goodness rating task. The evaluators were asked to rate the quality of each tone on a scale of 1–5, where 1 referred to poor pronunciation and 5 to excellent pronunciation. Incorrect or missing productions were given a rating of 0. The mean ratings of the two evaluators for each item were obtained. Thirteen items with a rating below 3 were excluded due to poor pronunciation or production errors (11 clear speech items). There were 2045 /ɤ/ words in total (21 speakers × 49 trials × 2 styles − 13 items).

D. Acoustic analysis

The onset and offset of a tone contour were first determined by the beginning and cessation of periodicity of the waveform. Then, a tone contour was divided into 100 equidistant intervals. *F*0 values were obtained at each of the 101 sampling time points in Praat using the autocorrelation method (pitch range: 50–450 Hz; time step = 15 ms) (Boersma and Weenink, 2017). The *F*0 values were manually checked for accuracy by phonetically trained research assistants. If there were more than ten consecutive sampling time points containing inaccurate or missing data, the *F*0 values were manually measured by taking the inverse of the duration of a single period at selected time points. These time points with manual measurements were equidistant from each other. The remaining time points were treated as containing missing data. Eventually, all portions with missing data contained fewer than ten time points. Missing data were replaced by values obtained by a linear interpolation to obtain a uniformly sampled vector of length 101 of *F*0 values for all tokens. In total, there were 420 plain and 399 clear items that had missing *F*0 data.

The fact that speakers have quite different *F*0 ranges is not problematic since we are only interested in differences within the productions of each speaker between their plain and clear styles; however, we normalized *F*0 data to make different speakers comparable. For each speaker, we applied the T-value logarithmic transform [Eq. (1)] to each *F*0 value,

$$T = 5 \times \frac{\log x - \log b}{\log a - \log b}, \tag{1}$$

where x represents the observed *F*0, a and b are the maximum *F*0 and minimum *F*0 of the speaker across all their productions, both clear and plain (Wang *et al.*, 2003). This choice of normalization requires some interpretation and has some different consequences in comparison to the alternative: normalizing only on the basis of a speaker's plain productions. In particular, the scale of normalized *F*0 values is determined by the more extreme productions, whether they are clear or plain. However, relative differences in *F*0 between clear and plain productions, which are our primary interest, are affected little by the choice of these two different normalization choices, since the effect of extreme values of *F*0 are similar between clear and plain productions.

We also normalize time values, so that all utterances could be compared on the same unit time interval; see Xu (1997, 2015) for examples of this procedure. This allows *F*0 curves of a given tone to be averaged over different utterances. Furthermore, since we retain the duration of each tone

in our subsequent analysis, no information is lost in this normalization.

Based on our previous acoustic analysis of Mandarin tones (see Tupper *et al.*, 2020 for details), we examined three *F*0 cues that efficiently characterize Mandarin tones: *F*0 mean, slope, and *F*0 second derivative which we call curve. These cues were obtained by fitting a parabola [Eq. (2)] to each tone contour and finding the best coefficients c_0, c_1, c_2 in the least squared sense (Rivlin, 1981) in the expression

$$f(t) \approx c_0 + c_1(t - 1/2) + c_2\left[(t - 1/2)^2 - 1/12\right]. \quad (2)$$

The resulting c_0, c_1 and c_2 are F0 mean, slope and curve, respectively. Importantly, these coefficients are computed using time-normalized data. In Sec. III C, we discuss why this choice appears to be most appropriate for the data. Note that of all our cues, only the values slope and curve are affected by this choice.

Other measures included total duration, which was the temporal difference between the onset and offset of a tone contour, and mean intensity obtained using the mean energy method in Praat based on the mean power between the onset and offset of a tone contour (Boersma and Weenink, 2017). A binary creakiness variable was created by applying a value of 1 to creaky productions and a value of 0 to non-creaky productions. The productions were determined to be creaky auditorily by the second author. In addition, these creaky productions were characterized by double pulses in the wideband spectrogram and missing or discontinuous *F*0 track (cf. Yu and Lam, 2014). When manual measurement of *F*0 values was possible, these creaky productions showed low *F*0 below 70 Hz (Drugman *et al.*, 2014; Keating *et al.*, 2015; Titze, 1994) and/or a decrease in *F*0 over 15 Hz between two consecutive sampling time points. A complete list of acoustic cues is displayed in Table I.

Other cues have been studied in the context of speech-style changes, but changes in these other cues are detectable by changes in the cues described in Table I. For example, pitch range has been reported to increase in emphasized speech (Chen and Gussenhoven, 2008). However, typically the only way to increase pitch range is to modify either slope or curve; emphasized tones in Chen and Gussenhoven (2008) clearly show an increase in the magnitude of slope. Pitch range has a disadvantage compared to slope as a cue to study, in that a rising tone and a falling tone may have identical pitch ranges, even though there is clearly a salient contrast.

Some of these cues we expect to be modified in purely signal-based changes: duration, *F*0 mean, and meanIntensity may all increase across tones in a signal-based change in style from plain to clear. In a code-based change, we may expect some features to change so as to exaggerate differences between tones. For example, the mean *F*0 of Tone 1 could increase while the mean *F*0 of Tone 3 decreased in clear speech, making it easier to distinguish these tones. Similar effects would happen in a code-based change for the cues *F*0 slope and curve. Finally, since creakiness is a distinctive feature of Tones 3 and 4 for many speakers, we could expect an increase in creakiness for these tones under a code-based change in style.

TABLE I. List of acoustic cues used in the present study and their definition. *F*0 always refers to transformed *F*0 (T-values).

	Cue name	Definition
1	Duration	Duration of tone (ms)
2	Mean	Mean value of *F*0
3	Slope	Mean slope of *F*0
4	Curve	Mean second derivative of *F*0
5	meanIntensity	Mean intensity (dB)
6	Creakiness	Whether the tone is produced with creaky voice

III. RESULTS

A. Population averages

We first show the normalized *F*0 contours for plain and clear styles averaged over all speakers and productions. Figure 1 shows the averages with the duration of all productions normalized to the same length, so that the *x* axis shows a percentage of the way through the tone. This of course obscures any effect speech style has on duration. In Fig. 2 we show the same data but with each average scaled by the average duration for all productions for that given tone and style, so that the *x* axis indicates time.

A visual comparison of tone contours in these figures suggests that there are limited plain-to-clear modifications for mean, slope, and curve (Fig. 1) (Note that slope and curve were obtained using time-normalized data). In contrast, there is an increase in duration for clear productions as compared to plain productions of all tones (Fig. 2). To further examine these observations, we proceed with a

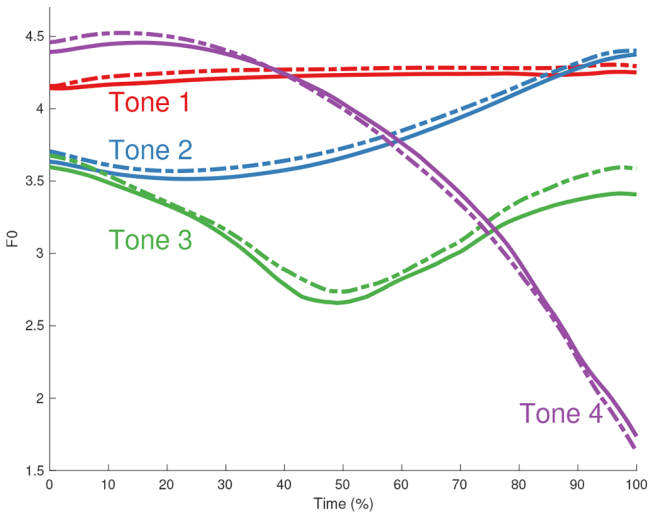

FIG. 1. (Color online) Normalized *F*0 contours averaged over all speakers and all tokens for each of the four tones. Time information was normalized so that all contours were of standard duration. The solid lines show plain productions. The dashed lines show clear productions.

https://doi.org/10.1121/10.0009142

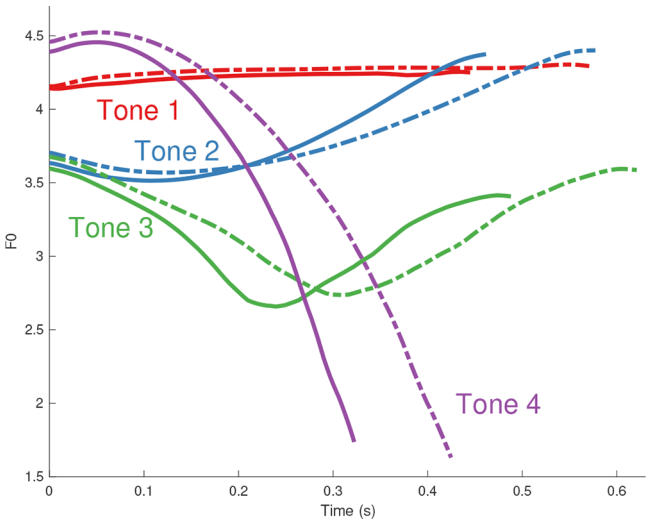

FIG. 2. (Color online) The data as in Fig. 1 except each average is scaled in time by the average duration of the productions of each tone. The solid lines show plain productions. The dashed lines show clear productions.

quantitative analysis of the effect of clearness on these four cues, as well as mean intensity and creakiness.

B. Examining individual cues

For each of 21 speakers and four tones, we determine how strongly the difference between plain and clear speaking style affects each of the six cues. We use Cohen's d as a measure of the effect size for each cue (Cohen, 1988). Given two groups of stimuli, Cohen's d is a normalized difference between the means of the stimuli. It measures how large the difference between the means is, but scaled by the variability of the stimuli. Suppose category 1 has n_1 data points with mean μ_1 and variance s_1^2, and likewise for category 2. Then Cohen's d is

$$d = \frac{\mu_1 - \mu_2}{s},$$

where s the pooled standard deviation is given by (Patten and Newhart, 2017)

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

For our study, we let category 1 consist of all the plain cues of a particular speaker and tone, and category 2 consist of all clear cues from the same speaker and tone. A positive value of d shows that clear productions have greater values of the cue than plain productions, whereas a negative value shows that they have a smaller value. We compute a d for each cue, for each tone, and for each speaker.

Figure 3 displays d for the four tones. In each panel, the top plot shows a heat map of d for each cue and speaker. The colour of each rectangle in the plot shows d for a particular cue for a particular speaker, with the colour bar at the side indicating the value. The bottom plot shows the distribution of d across the speakers for each cue using a box plot. Each cue is given its own boxplot showing the variation among speakers. Outliers are shown as red crosses, whiskers go to maxima and minima excluding outliers. The box and its dividing line show quartiles.

Clearness has a large effect on duration for all tones for most speakers. One way to measure if there was a consistent effect across speakers is to ask for which cues more than 75% of the speakers had a Cohen's d in the same direction. This can be seen by whether the quantile box is on one side of the x axis. This happens with duration for all tones, mean intensity for Tone 1 and 2, and mean $F0$ for Tone 2. Other cues were not systematically modified for particular tones by speakers, since more than 25% of speakers modified the cue in a different direction from the majority of the speakers.

We then performed a statistical analysis to determine if the clear style has any consistent effect across all speakers. For each tone, we performed a linear mixed-effects model with clearness as the independent variable and each cue (except for creakiness) as the dependent variable using the MATLAB `fitlme`. Since creakiness was a binary variable, a generalized linear mixed-effects model was carried out instead with a binomial distribution and logit linking function (MATLAB `fitglme`) (MATLAB, 2020). The random intercept and slope of clearness on speaker were included in the models with the following syntax:

$$CueName \sim clearness + (1 + clearness|speaker).$$

Table II shows the results, with + or − indicating the direction of the effect of clearness on the variable for each tone. All tones showed a significant increase in duration (Tone 1: $\beta = 0.126$, standard error (SE) = 0.033, $t(456) = 3.76$, $p < 0.001$; Tone 2: $\beta = 0.115$, SE = 0.041, $t(501) = 2.79$, $p = 0.005$; Tone 3: $\beta = 0.133$, SE = 0.050, $t(625) = 2.65$, $p = 0.008$; Tone 4: $\beta = 0.105$, SE = 0.024, $t(455) = 4.43$, $p < 0.001$) and mean intensity (Tone 1: $\beta = 1.26$, SE = 0.327, $t(456) = 3.85$, $p < 0.001$; Tone 2: $\beta = 0.968$, SE = 0.356, $t(501) = 2.72$, $p = 0.007$; Tone 3: $\beta = 1.09$, SE = 0.397, $t(625) = 2.75$, $p = 0.006$; Tone 4: $\beta = 0.594$, SE = 0.280, $t(455) = 2.12$, $p = 0.034$). The other significant effects were an increase in mean $F0$ for Tones 2 ($\beta = 0.052$, SE = 0.015, $t(501) = 3.41$, $p < 0.001$) and 3 ($\beta = 0.079$, SE = 0.024, $t(625) = 3.21$, $p = 0.001$), and a decrease in slope for Tone 4 ($\beta = -0.196$, SE = 0.098, $t(455) = -2.01$, $p = 0.045$), making it more steeply downwards for clear-speech productions.

C. Two interpretations of the clear-speech modification of duration

The largest effect of clearness, and the most consistent one across speakers, is that duration increased in the clear style versus the plain style. Increased duration is a well-known feature of clear speech. Among many possibilities, there are two very simple ways in which increased duration might be implemented for tone contours. The first, which we call *Go Slower*, involves maintaining an $F0$ contour in

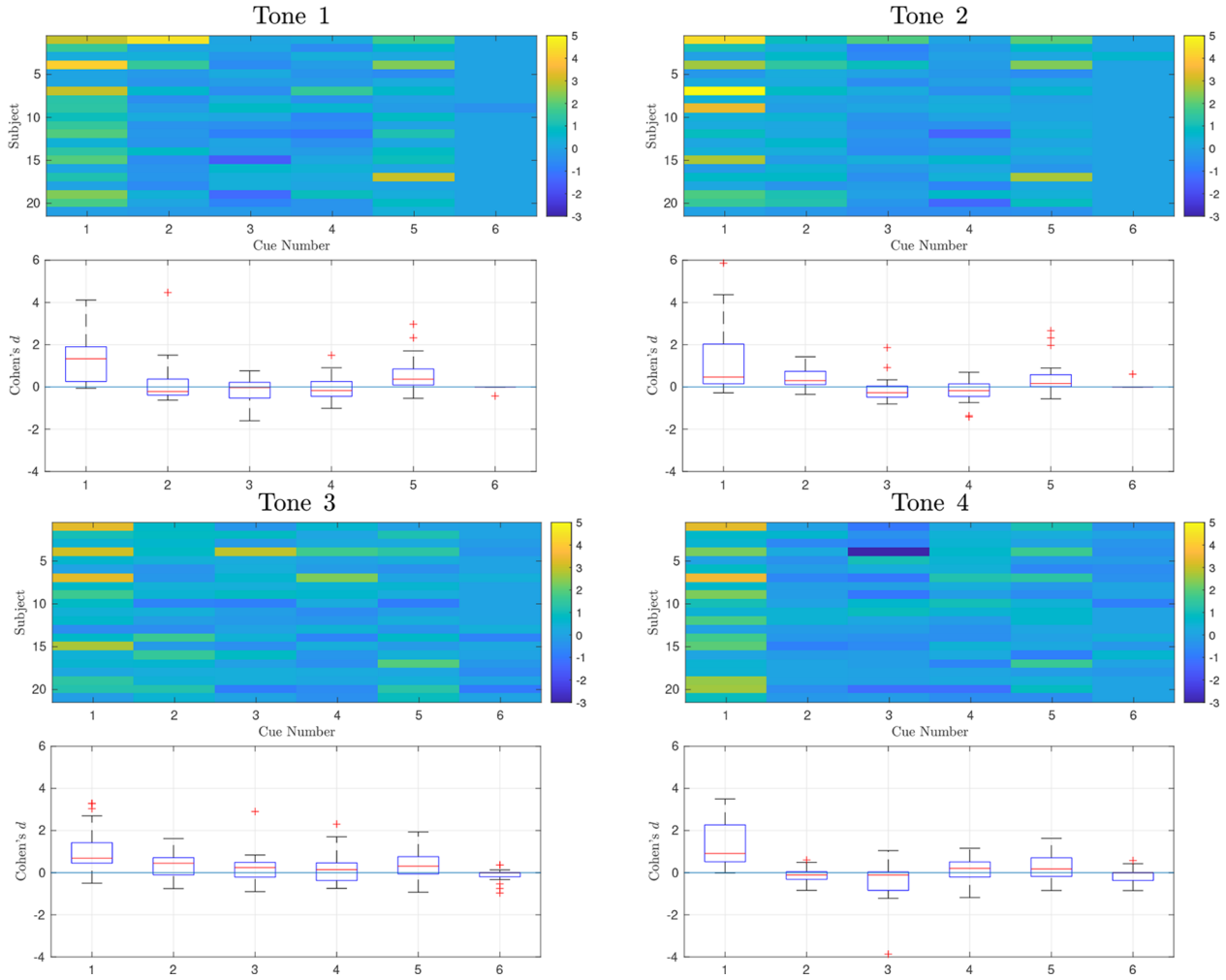

FIG. 3. (Color online) Speakers' use of the six cues in our study in the distinction between clear and plain tones. For each of the four tones, the top heat map shows d for each cue (see Table I for details) and speaker. The lower boxplot shows the distribution of d over speakers for each cue. In each bar, the red line shows the median, and the limits of the bar show the first quartile and the last quartile. The whiskers show the range of the distribution excluding outliers.

the same pitch range using a longer duration, so that the duration is increased without expanding the range of pitch. Under *Go Slower*, once time is normalized so that all contours occur on a unit interval of time, there is no difference between the plain and clear productions. The second we call *Do Longer*, which involves instead maintaining the same rate of change with respect to time throughout the tone. So for longer productions, there is an expanded pitch range, as seen for Mandarin tones emphasized in the context of corrective focus in Chen and Gussenhoven (2008). In fact, under *Do Longer*, the pitch range for each tone should increase by the same factor as the duration does. When plotted with normalized time, clear productions of contour tones should therefore have a larger range than plain. This could be interpreted as a code-based change to the $F0$ slope and curve. A glance at Figs. 1 and 2 shows that, averaged over the population, *Go Slower* appears to describe the data well. The range of $F0$ does not increase substantially with clear productions, with the possible exception of Tone 4, and this is despite substantial increases in the duration of the tone contours. We examine if this holds at the level of individual members of the population.

TABLE II. Overview of significant fixed effects of clearness on the values of each cue for each tone. + and – indicate the direction of the effect. Blank cells indicate no fixed effect.

	Cue name	Tone 1	Tone 2	Tone 3	Tone 4
1	Duration	+	+	+	+
2	Mean		+	+	
3	Slope				–
4	Curve				
5	meanIntensity	+	+	+	+
6	Creakiness				

For each speaker and each tone we compute the error of the *Go Slower* and *Do Longer* models for explaining plain versus clear styles. For *Go Slower* for each speaker and each tone we compute the average tone contour of all the plain tokens and the average tone contour of all the clear tokens, both of which are vectors of length 101. We then center

https://doi.org/10.1121/10.0009142

each vector (subtracting off the mean) so that any shifts up or down in pitch are not included, and then take the root mean squared difference between the two vectors to get the error of *Go Slower*. This should be zero if *Go Slower* is perfectly accurate. For *Do Longer*, the procedure is the same except that we multiply the plain vector by the ratio between the average duration of the clear productions and the plain productions for that speaker and tone before computing the error. If *Do Longer* is perfectly accurate, then this error should be zero, since the amount by which tones are stretched in *F*0 should match the amount by which their duration is extended. In Fig. 4, we show the errors of these two models for each speaker and tone. We see that there are many cases where both models have low errors (both when duration is not increased much and for Tone 1 which is almost flat) but that in cases where there is much difference *Go Slower* performs better than *Do Longer*. The one exception is one speaker's production of Tone 4. We conclude that *Go Slower* is a better model of how duration increases are implemented by speakers in clear speech, indicating that longer duration in clear speech did not automatically expand the pitch range.

There are other ways to implement longer duration while maintaining roughly the same *F*0 range. In our model, *Go Slower* assumes a uniform slowing through the whole production. Our calculations here are exploratory, so we did not investigate more complicated models of slowing, though we would expect strong nonuniform slowing to be apparent in larger differences between the time-normalized clear and plain productions in Fig. 1.

IV. DISCUSSION

The goal of our investigation was to see what differences Mandarin speakers implemented in tone production when changing style from plain to clear speech. In particular, we wanted to examine the extent to which speakers performed signal-based or code-based changes in their clear-speech tone productions (Bradlow and Bent, 2002; Redmon *et al.*, 2020; Zhao and Jurafsky, 2009). Compared to previous segment-based clear-speech studies, our research provides a unique testing case of clear-speech principles in that *F*0, previously identified as a signal-based acoustic cue, primarily serves code-based functions in lexical tone production.

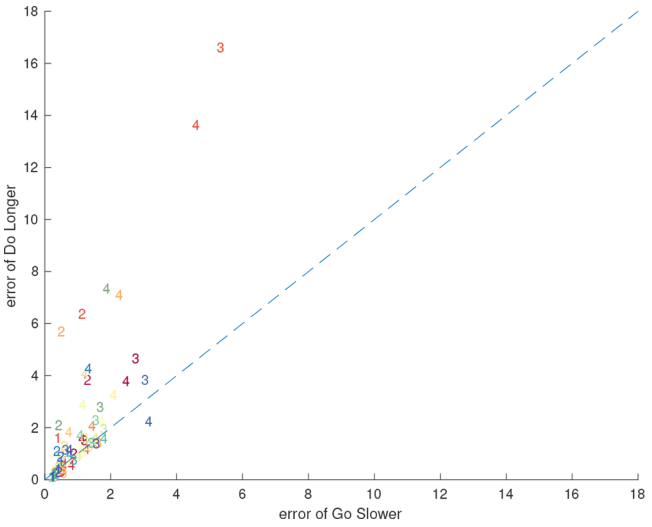

FIG. 4. (Color online) The mean squared error of the *Go Slower* and *Do Longer* models for each speaker and tone. Each speaker is represented by a different color.

We used two different ways to assess if a cue changed among speakers in a systematic way with the change of style. One was to look at which way cues changed on average, and see if more than 75% of speakers changed their productions in the same direction for a given cue and tone. We found that by this standard, duration increased for all tones, mean intensity increased for Tones 1 and 2, and mean *F*0 increased for Tone 2. We also performed a statistical analysis to see if there was a significant change in any of the cues for each of the tones using a linear mixed effects regression. Summarized in Table II, we found increase in duration for all tones, increase in mean intensity for all tones, increase in mean *F*0 for Tones 2 and 3, and a heightened slope for Tone 4.

Overall, there is strong evidence for signal-based changes in production in the change of style from plain to clear and limited evidence for code-based change. Duration and intensity are generally not contrastive features in Mandarin tones so those changes cannot be code-based, and they changed in the same direction for all tones, so it is unlikely that any between-category tone contrast was induced. Duration and intensity increases in clear speech coincided with what was found for IDS in Liu *et al.* (2007) and Tang *et al.* (2017), for Lombard speech in Tang *et al.* (2017), and for foreigner-directed speech in the language instruction setting in Han *et al.* (2019).

Signal-based change where mean *F*0 increases across all tones has been observed in IDS and Lombard speech (Grieser and Kuhl, 1988; Liu *et al.*, 2007; Tang *et al.*, 2017; Xu Rattanasone *et al.*, 2013; Zhao and Jurafsky, 2009), but not in foreigner-directed speech (Han *et al.*, 2019). The current results are consistent with the patterns observed for foreigner-directed speech in that we did not observe a statistically significant increase in mean *F*0 across tones. These differences are presumably because, unlike IDS, foreigner-directed speech and the current corrective type of clear speech do not necessarily need to engage overall *F*0 changes to gain attention or convey positive affect (Burnham *et al.*, 2010). Thus, signal-based changes in *F*0 can be restricted (Chen and Gussenhoven, 2008) to prioritize its primary code-based function in tonal category distinctions.

Indeed, some IDS studies have observed code-based changes in mean *F*0 (Liu *et al.*, 2007; Xu and Burnham, 2010). Since mean *F*0 is a critical cue distinguishing Mandarin tones, it was certainly conceivable that contrast could have been enhanced by, for example, lowering Tone 3 and raising Tone 1. However, in the current study, speakers may be reluctant to further increase mean *F*0 for Tone 1, since the tone contour is inherently high. The lack of

increase in *F*0 for Tone 1 may be taken as evidence of a code-based constraint in clear-speech modifications due to the intrinsic characteristics of T1. The inherently high-pitch nature of Tone 1 may make it more resistant to further increase in *F*0. This is consistent with segmental clear-speech findings. For example, the high-front tense vowel /i/ has demonstrated limited capacity for further articulatory excursions in clear speech because of its intrinsically extreme articulation (Granlund *et al.*, 2012; Leung *et al.*, 2016). The code-based modification resulting in enhancement of category distinction lies in the steeper slope and increased *F*0 range for Tone 4 in clear speech. This change also demonstrates that clear-speech modifications are aligned with the dynamic nature of Tone 4, as was also observed in IDS studies (Liu *et al.*, 2007). The finding of an increase in *F*0 range for Tone 4 only is consistent with the result from Chen and Gussenhoven (2008) that different tones revealed different limits on *F*0 range expansion, with Tone 4 expanding much more than the other tones. Together, these patterns suggest that clear speech production is modulated by code-based enhancements and constraints, in that individual tones are in tune with their inherent attributes to enhance as well as preserve category distinctions.

Consistent with these observations, *GoSlower* was observed to be the predominant strategy for manipulating duration in the form of clear speech of this study, meaning that duration was increased without a corresponding increase in the range of pitches attained, consistent with a signal-based change.

Regarding non-modal phonation, this study did not find clear-speech modification for creakiness across four Mandarin tones. It is not surprising since creaky phonation is expected to covary with *F*0 (Kuang, 2017). The increase in creakiness was found to be accompanied by a lowering of *F*0 (Chen and Gussenhoven, 2008; Kuang, 2017) or an increase in *F*0 for Tone 4 specifically (Huang *et al.*, 2018). These modifications were not found in the current study (Table II).

Taken together, placing the current results of clear-speech tone modification in the context of other goal-oriented speaking styles such as IDS, Lombard speech, and foreigner-directed speech, it appears that lengthened duration and increased intensity are universal signal-based strategies to enhance signal saliency in hyperarticulated tones. The use of *F*0 exhibits more complex patterns, involving either signal-based overall increases when serving prosodic functions to convey attentional and affective components, or code-based modifications of individual tones to enhance category distinctiveness (as revealed by the stretching of tone space or steepening of tone contour slope). Such dual functions that *F*0 has to serve dictate that the use of *F*0 in clear-speech modifications may be restricted as *F*0 variations are primarily reserved for use as tone-intrinsic cues for category distinctions. Our results consistently suggest that speakers primarily rely on signal-based duration and intensity changes in clear-speech modifications, rather than on code-based *F*0 changes that enhance the contrast between tones.

ACKNOWLEDGMENTS

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. 435-2012-1641). We thank SFU Language and Brain Lab members Zoe Beukers, Anisa Dhanji, Hasti Halakoeei, Abner Hernandez, Michelle Le, Kelsey Philip, Michelle Ramsay, and Joanna Xie for conducting acoustic measurements.

Baese-Berk, M., and Goldrick, M. (**2009**). "Mechanisms of interaction in speech production," Lang. Cogn. Process. **24**(4), 527–554.
Boersma, P., and Weenink, D. (**2017**). "Praat, a system for doing phonetics by computer (version 6.0.28) [computer program]," http://www.fon.hum.uva.nl/praat/ (Last viewed March 23, 2017).
Bradlow, A. R., and Bent, T. (**2002**). "The clear speech effect for non-native listeners," J. Acoust. Soc. Am. **112**(1), 272–284.
Burnham, D., Joeffry, S., and Rice, L. (**2010**). "D-o-e-s-Not-C-o-m-p-u-t-e: Vowel hyperarticulation in speech to an auditory-visual avatar," in *Proceedings of the 9th Auditory-Visual Speech Processing*, Norwich, UK (September 10-13), p. P18.
Chen, Y., and Gussenhoven, C. (**2008**). "Emphasis and tonal implementation in Standard Chinese," J. Phon. **36**(4), 724–746.
Cohen, J. (**1988**). *Statistical Power Analysis for the Social Sciences* (Erlbaum, Hillsdale, NJ).
Drugman, T., Kane, J., and Gobl, C. (**2014**). "Data-driven detection and analysis of the patterns of creaky voice," Comput. Speech Lang. **28**(5), 1233–1253.
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**(1), 259–271.
Ferguson, S. H., and Kewley-Port, D. (**2007**). "Talker differences in clear and conversational speech: Acoustic characteristics of vowels," J. Speech Lang. Hear. Res. **50**(5), 1241–1255.
Granlund, S., Hazan, V., and Baker, R. (**2012**). "An acoustic-phonetic comparison of the clear speaking styles of finnish-english late bilinguals," J. Phon. **40**(3), 509–520.
Grieser, D. L., and Kuhl, P. K. (**1988**). "Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese," Dev. Psychol. **24**(1), 14–20.
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**(4), 263–286.
Huang, Y., Athanasopoulou, A., and Vogel, I. (**2018**). "The effect of focus on creaky phonation in Mandarin Chinese tones," Univ. Pennsylvania Work. Papers Ling. **24**(1), 12, see https://repository.upenn.edu/pwpl/vol24/iss1/12/.
Junqua, J.-C. (**1996**). "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," Speech Commun. **20**(1), 13–22.
Keating, P., Garellek, M., and Kreiman, J. (**2015**). "Acoustic properties of different kinds of creaky voice," in *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK (August 10-14), pp. 0821.1–0821.5.
Kuang, J. (**2017**). "Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice," J. Acoust. Soc. Am. **142**(3), 1693–1706.
Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., and Lacerda, F. (**1997**). "Cross-language analysis of phonetic units in language addressed to infants," Science **277**(5326), 684–686.
Lam, J., Tjaden, K., and Wilding, G. (**2012**). "Acoustics of clear speech: Effect of instruction," J. Speech Lang. Hear. Res. **55**(6), 1807–1821.
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**(1), 45–58.
Liu, H.-M., Tsao, F.-M., and Kuhl, P. K. (**2007**). "Acoustic analysis of lexical tone in Mandarin infant-directed speech," Dev. Psychol. **43**(4), 912–917.

Maniwa, K., Jongman, A., and Wade, T. (**2009**). "Acoustic characteristics of clearly spoken English fricatives," J. Acoust. Soc. Am. **125**(6), 3962–3973.
MATLAB (**2013**). *Version 8.1* (*R2013a*) (The MathWorks Inc., Natick, MA).
MATLAB (**2020**). *Version 9.9.0.1467703* (*R2020b*) (The MathWorks Inc., Natick, MA).
Moon, S.-J., and Lindblom, B. (**1994**). "Interaction between duration, context, and speaking style in English stressed vowels," J. Acoust. Soc. Am. **96**(1), 40–55.
Ohala, J. J. (**1995**). "Clear speech does not exaggerate phonemic contrast," in *Fourth European Conference on Speech Communication and Technology*, Mandarin, Spain (September 18–21).
Papoušek, M., and Hwang, S.-F. C. (**1991**). "Tone and intonation in Mandarin babytalk to presyllabic infants: Comparison with registers of adult conversation and foreign language instruction," Appl. Psycholinguist. **12**(4), 481–504.
Patten, M. L., and Newhart, M. (**2017**). *Understanding Research Methods: An Overview of the Essentials* (Routledge, New York).
Redmon, C., Leung, K., Wang, Y., McMurray, B., Jongman, A., and Sereno, J. A. (**2020**). "Cross-linguistic perception of clearly spoken English tense and lax vowels based on auditory, visual, and auditory-visual information," J. Phon. **81**, 100980.
Rivlin, T. J. (**1981**). *An Introduction to the Approximation of Functions* (Courier Corporation, Chelmsford, MA).
Shi, F., and Deng, D. (**2006**). "The comparison on phonetic system between putonghua and Taiwan mandarin," in *Mountain Lofty, River Long: Festschrift in Honor of Professor Pang-Hsin Ting on His Seventieth Birthday* (Academia Sinica, New York), pp. 371–393 (in Chinese).
Smiljanić, R. (**2021**). "Clear speech perception: Linguistic and cognitive benefits," in *The Handbook of Speech Perception*, edited by J. S. Pardo, L. C. Nygaard, R. E. Remez, and D. B. Pisoni (Wiley-Blackwell, Hoboken, NJ).
Smiljanić, R., and Bradlow, A. R. (**2009**). "Speaking and hearing clearly: Talker and listener factors in speaking style changes," Lang. Ling. Compass **3**(1), 236–264.
Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (**1988**). "Effects of noise on speech production: Acoustic and perceptual analyses," J. Acoust. Soc. Am. **84**(3), 917–928.
Tang, P., Xu Rattanasone, N., Yuen, I., and Demuth, K. (**2017**). "Phonetic enhancement of Mandarin vowels and tones: Infant-directed speech and Lombard speech," J. Acoust. Soc. Am. **142**(2), 493–503.
Titze, I. R. (**1994**). "Vocal registers," in *Principles of Voice Production* (Prentice-Hall, Inc., Englewood Cliffs, NJ), pp. 252–259.
Tupper, P., Leung, K., Wang, Y., Jongman, A., and Sereno, J. A. (**2020**). "Characterizing the distinctive acoustic cues of Mandarin tones," J. Acoust. Soc. Am. **147**(4), 2570–2580.
Tupper, P. F., Jian, J., Leung, K. K. W., and Wang, Y. (**2018**). "Game theoretic models of clear versus plain speech," in *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, Madison, WI (July 25–28).
Uther, M., Knoll, M., and Burnham, D. (**2007**). "Do you speak E-NG-L-I-SH? a comparison of foreigner- and infant-directed speech," Speech Commun. **49**(1), 2–7.
Wang, Y., Jongman, A., and Sereno, J. A. (**2003**). "Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training," J. Acoust. Soc. Am. **113**(2), 1033–1043.
Wedel, A., Nelson, N., and Sharp, R. (**2018**). "The phonetic specificity of contrastive hyperarticulation in natural speech," J. Mem. Lang. **100**, 61–88.
Wong, P., Fu, W. M., and Cheung, E. Y. (**2017**). "Cantonese-speaking children do not acquire tone perception before tone production—A perceptual and acoustic study of three-year-olds' monosyllabic tones," Front. Psychol. **8**, 1450.
Xu, N., and Burnham, D. (**2010**). "Tone hyperarticulation and intonation in Cantonese infant directed speech," in *Speech Prosody 2010-Fifth International Conference*, Chicago, IL (May 11–14).
Xu, N., Burnham, D., Kitamura, C., and Vollmer-Conna, U. (**2013**). "Vowel hyperarticulation in parrot-, dog- and infant-directed speech," Anthrozoös **26**(3), 373–380.
Xu, Y. (**1997**). "Contextual tonal variations in Mandarin," J. Phon. **25**(1), 61–83.
Xu, Y. (**2015**). *Speech Prosody: Theories, Models and Analysis* (Cambridge Scholars, Cambridge, UK).
Xu Rattanasone, N., Burnham, D., and Reilly, R. G. (**2013**). "Tone and vowel enhancement in Cantonese infant-directed speech at 3, 6, 9, and 12 months of age," J. Phon. **41**(5), 332–343.
Yang, B. (**2015**). *Perception and Production of Mandarin Tones by Native Speakers and L2 Learners* (Springer, New York).
Yu, K. M., and Lam, H. W. (**2014**). "The role of creaky voice in cantonese tonal perception," J. Acoust. Soc. Am. **136**(3), 1320–1333.
Zhao, Y., and Jurafsky, D. (**2009**). "The effect of lexical frequency and lombard reflex on tone hyperarticulation," J. Phon. **37**(2), 231–247.