Characterizing the distinctive acoustic cues of Mandarin tones

Paul Tupper,1, a) Keith Leung,2 Yue Wang,2 Allard Jongman,3 and Joan A. Sereno3

1Department of Mathematics, Simon Fraser University, Burnaby, British Colombia V5A 1S6, Canada
2Department of Linguistics, Simon Fraser University, Burnaby, British Colombia V5A 1S6, Canada
3Department of Linguistics, University of Kansas, Lawrence, Kansas 66045, USA

ABSTRACT:
This study aims to characterize distinctive acoustic features of Mandarin tones based on a corpus of 1025 monosyllabic words produced by 21 native Mandarin speakers. For each tone, 22 acoustic cues were extracted. Besides standard F0, duration, and intensity measures, further cues were determined by fitting two mathematical functions to the pitch contours. The first function is a parabola, which gives three parameters: a mean F0, an F0 slope, and an F0 second derivative. The second is a broken-line function, which models the contour as a continuous curve consisting of two lines with a single breakpoint. Cohen’s d, sparse Principal Component Analysis, and other statistical measures are used to identify which of the cues, and which combinations of the cues, are important for distinguishing each tone from each other among all the speakers. Although the specific cues that best characterize the tone contours depend on the particular tone and the statistical measure used, this paper shows that the three cues obtained by fitting a parabola to the tone contour are broadly effective. This research suggests using these three cues as a canonical choice for defining tone characteristics.

https://doi.org/10.1121/10.0001024

(Received 21 October 2019; revised 10 February 2020; accepted 18 March 2020; published online 24 April 2020)

ACoustical Society of America

I. INTRODUCTION

Many languages, including Mandarin Chinese, employ tones to convey lexical meaning. Acoustically, lexical tones are manifested primarily as changes in fundamental frequency (F0, perceived as pitch) as well as duration and amplitude (Howie and Howie, 1976; Lehiste, 1970). However, the critical acoustic cues characterizing individual tones had not been consistently identified in previous research. A difficulty of this type of investigation is that a tone contour is how F0 varies as a function of time for the whole voiced portion of the signal. It is an example of infinite-dimensional (or functional) data in that, to perfectly describe the contour, one needs to record F0 at every instant in time. In practice, of course, this is impractical and unnecessary. It is possible to summarize the tone contours with a small number of cues, as many authors have done (e.g., Barry and Blamey, 2004; Hirst and Espesser, 1993; Wong et al., 2017; Yang, 2015). This raises the question of which set of cues are the most informative for describing the tone contours of Mandarin. The current study considers a wide selection of cues, some occurring in the literature and some new ones, and uses a large corpus of spoken words in Mandarin to assess which cues are the most successful in distinguishing different tones.

There is a long history of using cues to study the tones of Mandarin. Much of the early work was based on perceptual data, using artificially generated tones and studying listeners’ discrimination judgements. An important early investigation (Gandour, 1983) used a multi-dimensional scaling technique (INDSCAL: individual difference scaling; Carroll and Chang, 1970) on dissimilarity judgements to place a selection of tone contours in a two-dimensional space. This space can be thought of as the perceptual space of the listeners when presented with tone stimuli. The nature of multi-dimensional scaling techniques in general is that they do not provide an explicit map that can then be used to compute cues for new stimuli. But Gandour analyzed the two dimensions provided by INDSCAL and determined that they are roughly what he calls “height” and “direction.” His “height” is average or mean F0 for each tone, and his “direction” appears to correspond to slope of the F0 contour for each tone, though perhaps slope toward the end of the contour is more accurate.

Acoustic studies have also identified F0 mean and slope as critical tonal cues. For example, Wong et al. (2017) used five F0 height measurements (mean, initial, final, maximum, and minimum) as well as a direction measure (the slope of the second half of the F0 contour) to characterize Cantonese tones. The second half of the syllable was selected based on the previous findings that perceptual cues for tones are carried toward the end of the syllable (Khouw and Ciocca, 2007; Xu, 2001; Xu and Wang, 2001).

Based on Gandour (1983) and other studies, a reasonable pair of cues to describe tone contours are F0 mean and slope. Mean is straightforwardly defined as the mean F0 over the duration of the periodic portion. However, there are a few different ways of defining the “slope” or “direction”
of a tone contour. Let us assume that we normalize our tone contours so that they are all the same duration in time, say, 1 time unit. Some authors (e.g., Yang, 2015) use the difference between F0 at offset and F0 at onset, which, if the contour were a straight line, would yield its slope (since we are assuming a time duration of 1). Another option is to take the range of F0 values and divide by the duration (Flemming and Cho, 2017; Jeng et al., 2006). This always gives a positive value and, if applied to a straight line, gives the absolute value of the slope. However, this measurement fails to differentiate between a falling tone and a rising tone with the same amplitude of slope. A more sophisticated method, which relies on the entire tone contour rather than just a pair of distinguished points, is to fit a linear function to the contour in the least-squares sense (e.g., Black and Hunt, 1996; Ghosh and Narayanan, 2009; Hirst and Espesser, 1993). This is a special case of least-squares fitting. More generally, the fundamental idea of least-squares fitting is to select a model with unknown coefficients (such as height and slope, in this case) and then choose the coefficients so that the integrated squared difference between the model and the data is minimized. We will use this method to compute our slope cue, as we describe later.

The problem with just using mean and slope to characterize tone contours is that they do not capture what is one of the most salient features of some tone contours: their curvature. For example, Tone 3 of Mandarin starts high, of the most salient features of some tone contours: their slope cue, as we describe later.

previous studies found that the temporal envelope cues are pertinent to tone perception (Fu and Zeng, 2000; Fu et al., 1998; Kong and Zeng, 2006; Wang et al., 2011). Temporal envelope cues include three main acoustic cues: periodicity, amplitude contour, and duration (Fu and Zeng, 2000; Kong and Zeng, 2006; Rosen, 1992). Among these cues, periodicity reflects F0 in the speech signal (Rosen, 1992) and is covered by the above-mentioned F0 acoustic cues. As for amplitude, when Mandarin listeners perceive Mandarin monosyllabic stimuli without F0 or formant information, the amplitude contour becomes a useful cue in the perception of Mandarin Tone 2, Tone 3, and Tone 4 (Fu and Zeng, 2000; Whalen and Xu, 1992). Acoustic studies consistently show that the intrinsic amplitude varies across tones, with Tone 3 having the lowest, and Tone 4 the highest overall amplitude (Chuang et al., 1972). Duration is used for perceiving whispered speech (Liu and Samuel, 2004). When F0 information is present, the modifying duration of stimuli can shift the perceptual boundary between Tone 2 and Tone 3 in native and non-native speakers (Blicher et al., 2000). Accordingly, in our study we include duration as a cue, as well as three cues related to amplitude.

In all the work we have surveyed here, only a few different cues are considered, which are either extracted from data using multidimensional scaling (Gandour, 1983), or postulated by the researchers based on inspection of the data and basic knowledge of how the Mandarin tone system works. Seldom are different cues compared to see which are corresponding to mean, slope, and second derivative, respectively. Other authors who have fit a parabola to extract cues include Chen et al. (2017); Li and Chen (2016); Shih and Lu (2015); Zhang and Meng (2016).

Other studies have found additional dynamic and local cues to be important. Smith and Burnham (2012) included 18 F0-related acoustic cues to Mandarin tones, finding mean and minimum F0 velocity as well as F0 onset to be critical cues to auditory tone perception. Similarly, Barry and Blamey (2004) used F0 offset and F0 onset to study tone differentiation in Cantonese. Another approach is to construct a model of how F0 contours arise from more basic phenomena (Prom-on et al., 2012; Prom-on et al., 2009; Xu, 2005; Xu and Wang, 2001). The motivation of these authors is mainly coarticulation in polysyllabic speech.

Another cue that has been proposed based on perception studies is turning point location (TP), defined as the temporal interval between the onset and the lowest point of the contour (Moore and Jongman, 1997; Zhao and Kuhl, 2015). Moore and Jongman parametrized tone contours with TP and AF0, which was defined as the drop in F0 from onset to the turning point. TP is likely to be useful for distinguishing Tone 2 from Tone 3. We capture turning point (and many other cues) using a broken line fit. We define a broken line to be a continuous function over a time interval that consists of two straight lines with a single breakpoint. It is described by four parameters. One of these, the breakpoint, which gives the location of the breakpoint in the interval, corresponds to the turning point in these earlier studies.

Previous studies found that the temporal envelope cues are pertinent to tone perception (Fu and Zeng, 2000; Fu et al., 1998; Kong and Zeng, 2006; Wang et al., 2011). Temporal envelope cues include three main acoustic cues: periodicity, amplitude contour, and duration (Fu and Zeng, 2000; Kong and Zeng, 2006; Rosen, 1992). Among these cues, periodicity reflects F0 in the speech signal (Rosen, 1992) and is covered by the above-mentioned F0 acoustic cues. As for amplitude, when Mandarin listeners perceive Mandarin monosyllabic stimuli without F0 or formant information, the amplitude contour becomes a useful cue in the perception of Mandarin Tone 2, Tone 3, and Tone 4 (Fu and Zeng, 2000; Whalen and Xu, 1992). Acoustic studies consistently show that the intrinsic amplitude varies across tones, with Tone 3 having the lowest, and Tone 4 the highest overall amplitude (Chuang et al., 1972). Duration is used for perceiving whispered speech (Liu and Samuel, 2004). When F0 information is present, the modifying duration of stimuli can shift the perceptual boundary between Tone 2 and Tone 3 in native and non-native speakers (Blicher et al., 2000). Accordingly, in our study we include duration as a cue, as well as three cues related to amplitude.

In all the work we have surveyed here, only a few different cues are considered, which are either extracted from data using multidimensional scaling (Gandour, 1983), or postulated by the researchers based on inspection of the data and basic knowledge of how the Mandarin tone system works. Seldom are different cues compared to see which are
the most significant in distinguishing the tones of Mandarin. Our paper takes a large number of cues that have been proposed and systematically evaluates them using a large dataset. The main contributions of our paper are (i) to consider a wide variety of cues and see how well they do at characterizing the naturally occurring variation in a large corpus of tokens, (ii) identify which cues are functionally equivalent in that they roughly determine the same information about a tone contour, and (iii) extract a small subset of cues that captures the essential variations in the tone contours to determine the optimal cues for characterizing Mandarin tones.

We first assess how each individual cue does alone in distinguishing each tone from the three others using Cohen’s d (Cohen, 1988). Many cues do very well under this metric, but selecting the top distinguishing cues according to this metric is not a good strategy because the top cues are often redundant, capturing the same information about the tones. We then use sparse Principal Component Analysis (PCA; Sjöstrand et al., 2018; Zou et al., 2006) to identify groupings of correlated cues that can be viewed as redundant. Sparse PCA has many of the advantages of PCA for identifying ways in which variables co-vary but gives sparse components containing only a few variables each, rather than linear combinations that contain all variables. Ideally, we want to select at most one cue from each of the sparse components. Next, we look at all pairs of cues together and see how good they are at distinguishing tones. Since there is no multi-dimensional analogue of Cohen’s d, we instead see how well a linear classifier does using only the given pair of cues to distinguish the tones. Our choice is to use Linear Discriminant Analysis (LDA), a quite standard method for classification (Friedman et al., 2001).

II. METHODS

A. Participants

Twenty-one speakers (11 female, 10 male) were recruited from the undergraduate and graduate student population at Simon Fraser University. Participants were native speakers of Mandarin Chinese aged 18–28 (mean: 22.6). They were raised in Northern China or Taiwan during the first 12 years of life. Although some participants had knowledge of another Chinese dialect, they reported that standard Mandarin was their native and dominant language. They reported normal hearing and no history of speech or language disorders.

B. Materials

The monosyllable /ɔ/ (or e in pinyin) with four Mandarin tones was used in this study, carrying the meaning of “graceful” (/31/; Tone 1, level tone), “goose” (/32/; Tone 2, rising tone), “nauseous” (/33/; Tone 3, dipping tone), and “hungry” (/34/; Tone 4, falling tone), respectively. The /i/ and /u/ words were mixed with /ɔ/ words in the recordings, but were included to address research objectives that are not reported in this paper. Therefore, only /ɔ/ words were analyzed. The production of each token was recorded in isolation in plain and clear speaking styles. Only the plain productions were analyzed for the present study.

C. Procedures

The participants’ speech was recorded digitally in a sound-attenuating booth in the Language and Brain Laboratory at Simon Fraser University, 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. Participants were seated at a comfortable distance from the computer screen where prompts, instructions, and feedback were displayed. Before recording began, participants were asked to read the four words aloud to become familiar with the stimuli.

The recording session began with a warm-up session. Speakers produced five repetitions of each word in each of two blocks, in response to prompts appearing on a monitor. In addition to /ɔ/, /i/ and /u/ quadruplet tone words were also recorded and served as fillers. Since /ɔ/ is a mid-central vowel and the production involves the least tongue movement among the three vowels, it presumably has the least interaction with tone production and therefore was chosen in this analysis. Using a vowel-only phonetic context enabled us to avoid any effect of consonant productions on the production of tones. The warm-up sessions served to familiarize speakers with the interface and materials, and to allow them to rehearse the target words. The productions from the warm-up sessions were not included in the current analyses. Then, the participants completed three elicitation sessions and were instructed to speak naturally. Productions of these words were in a random order.

A total of 49 /ɔ/ productions were obtained per speaker, in 49 elicitation trials described above [11 (/31/) + 12 (/32/) + 15 (/33/) + 11 (/34/)]. 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 /ɔ/ words in the recordings. There were 55 /i/ and 51 /u/ productions. The order of prompts and responses was the same for each participant. Due to recording errors, four tokens were excluded, yielding a total of 1025 tokens of /ɔ/ words (21 speakers × 49 productions – 4 errors).

Each speaker’s productions were evaluated by two phonetically trained native Mandarin evaluators in a goodness rating task.

D. Acoustic analysis

The onset and offset of a tone contour were determined by the beginning and cessation of periodicity of the waveform. For Tone 4 productions, since a substantial amount of irregular cycles, indicating creakiness, was observed at the offset, the endpoint in such productions was determined by the last identifiable cycle.

The tone contour was divided into 100 intervals of equal duration. F0 values in Hertz were then obtained at the
101 time points in Praat using the autocorrelation method with a pitch range of 50–450 Hz and a time step of 0.015 s (Boersma and Weenink, 2017). The F0 values were manually checked for accuracy by phonetically trained research assistants. Manual measurements were conducted if the inaccurate or missing data portion contained more than 10 time points. A selected number of time points was manually measured by taking the inverse of the duration of a single period. The other inaccurate F0 values were removed and treated as missing data. These time points with manual measurements were equidistant from each other, and the remaining missing portions contained fewer than 10 time points. In total, 394 tokens had missing F0 data, 269 of which were identified as creaky-voiced. Missing data were replaced by values obtained by a linear interpolation to obtain a uniformly sampled vector of length 101 F0 values for all tokens. The productions for all four tones were normalized using the T-value logarithmic transform to handle inter-speaker variations in F0 range,

$$T = 5 \times \frac{\log x - \log b}{\log a - \log b},$$  \hspace{1cm} (1)

where \(x\) represents the observed F0, \(a\) and \(b\) are the maximum F0 and minimum F0, respectively, of the speaker (Wang et al., 2003). Figure 1 shows the normalized tone contours for the four tones averaged over all participants and all tokens.

To explore acoustic features for characterizing Mandarin tones, this study included 22 cues that were extracted from each token (Table I). The first cue is total duration. The next 18 cues were extracted from the 101 F0 values of the tone contour, which we describe in more detail below.

Many of our F0 features were obtained by fitting two mathematical functions. The first was fitting a parabola to the tone contour and produced three cues: a mean F0, an F0 slope, and an F0 second derivative, which we call curve. These polynomials are orthogonal on \([0, 1]\) (Komzsik, 2017). Recall that time was scaled so that the contour is always over the time interval \([0, 1]\). Let the normalized pitch be given by \(f(t)\) for \(0 \leq t \leq 1\). We fit a parabola to the tone contour by finding the best coefficients \(c_0, c_1, c_2\) in the least-squared sense (Rivlin, 1981) in the expression

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

The coefficients here are multiplying the first three Legendre polynomials translated to the interval \([0, 1]\) (Komzsik, 2017). These polynomials are orthogonal on \([0, 1]\), so \(c_0\) is the mean of \(f(t)\), and \(c_0\) and \(c_1\) are what one would get from doing the best linear fit to \(f(t)\) on the same interval, and hence \(c_1\) deserves the term slope. Finally, \(c_2\) gives one half the second derivative of \(f\) if \(f\) were a parabola, and so we refer to it as curve. In reality, the data is discrete in time (101 evenly spaced points), and so the polynomials are also sampled at these points. The parameters \(c_0, c_1, c_2\) thus obtained become our cues 2, 3, and 4, respectively.

<table>
<thead>
<tr>
<th>Cue name</th>
<th>Definition</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 duration</td>
<td>duration of tone</td>
<td>in units ms</td>
</tr>
<tr>
<td>2 mean</td>
<td>mean value of F0</td>
<td>parabolic fit</td>
</tr>
<tr>
<td>3 slope</td>
<td>mean slope of F0</td>
<td>parabolic fit</td>
</tr>
<tr>
<td>4 curve</td>
<td>mean second derivative of F0</td>
<td>parabolic fit</td>
</tr>
<tr>
<td>5 BLstart</td>
<td>F0 of BL fit at onset</td>
<td>broken-line fit</td>
</tr>
<tr>
<td>6 BLslope1</td>
<td>slope of first line segment in BL fit</td>
<td>broken-line fit</td>
</tr>
<tr>
<td>7 BLbreakpoint</td>
<td>location of break point in BL fit</td>
<td>broken-line fit</td>
</tr>
<tr>
<td>8 BLslope2</td>
<td>slope of second line segment in BL fit</td>
<td>broken-line fit</td>
</tr>
<tr>
<td>9 onglide</td>
<td>difference between F0 at onset and breakpoint in BL fit</td>
<td>broken-line fit</td>
</tr>
<tr>
<td>10 offglide</td>
<td>difference between F0 at breakpoint and offset in BL fit</td>
<td>broken-line fit</td>
</tr>
<tr>
<td>11 overall</td>
<td>difference between F0 at onset and offset in BL fit</td>
<td>broken-line fit</td>
</tr>
<tr>
<td>12 F0range</td>
<td>range of F0 values</td>
<td></td>
</tr>
<tr>
<td>13 F0onset</td>
<td>F0 at onset</td>
<td></td>
</tr>
<tr>
<td>14 F025percent</td>
<td>F0 at 25% way through the vowel</td>
<td></td>
</tr>
<tr>
<td>15 F050percent</td>
<td>F0 at 50% way through the vowel</td>
<td></td>
</tr>
<tr>
<td>16 F075percent</td>
<td>F0 at 75% way through the vowel</td>
<td></td>
</tr>
<tr>
<td>17 F0offset</td>
<td>F0 at offset</td>
<td></td>
</tr>
<tr>
<td>18 maxLocation</td>
<td>temporal location of the maximum F0</td>
<td>in interval ([0,1])</td>
</tr>
<tr>
<td>19 minLocation</td>
<td>temporal location of the minimum F0</td>
<td>in interval ([0,1])</td>
</tr>
<tr>
<td>20 meanIntensity</td>
<td>mean intensity</td>
<td></td>
</tr>
<tr>
<td>21 maxIntensity</td>
<td>maximum intensity</td>
<td></td>
</tr>
<tr>
<td>22 locationMaxIntensity</td>
<td>temporal location of the maximum intensity</td>
<td>in interval ([0,1])</td>
</tr>
</tbody>
</table>

The broken line fit was the next mathematical model used and was intended to explicitly catch TP and \(\Delta F0\). The parameters are F0 onset \(d_0\), slope of the first part of the line \(d_1\), location of the breakpoint \(d_2\), and slope of the second part of the line \(d_3\). The four parameters are then determined by finding the best least-squares approximation to \(f\). Since the model depends nonlinearly on \(d_2\), it is not possible to have an explicit least-squares fit; we therefore cycled through all possible \(d_2\) values, did a least-squares fit of the other parameters, and then at the end selected the parameters with the least error. We note that \(d_0\) is not the true onset (which we include as cue 13), but the onset of the broken line fit. The parameters \(d_0, d_1, d_2, d_3\) become cues 5 through 8. Three further cues were then obtained using these cues: onglide (cue 9) is the difference between the onset F0 of the broken line fit and the F0 at the breakpoint, offglide (cue 10) is the difference between the F0 at the breakpoint and the F0 of the broken line fit at offset, and overall (cue 11) is the difference between broken line onset and offset.

To illustrate our two classes of fitting functions used in this study, Fig. 2 shows four tokens (one for each tone) for a single speaker. The top plot shows the raw normalized F0
Additional F0 measurements were determined through more straightforward means. Cue 12 (F0range) was the range of F0, computed as the difference between the maximum and minimum F0. Cues 13 through 17 were the F0 value measured at 0%, 25%, 50%, 75%, and 100% of the way through the contour (cf. Wang et al., 2003). The first of these we denote by F0onset, and the last we denote by F0offset. Since the location of the F0 turning point as well as the F0 value at the turning point are relevant for Tone 2 and Tone 3 (Moore and Jongman, 1997), we measured the temporal location of the maximum and minimum of the tone contours, giving us cues 18 and 19.

The last three cues (20, 21, 22) are mean and maximum intensity, and the temporal location of maximum intensity (Whalen and Xu, 1992). The mean intensity values were acquired using the mean energy method, which is computed by the mean power between the onset and offset of each tone. We selected the mean amplitude measure based on previous research (e.g., Chuang et al., 1972). In addition, we included local amplitude measures (maximum intensity and its temporal location) instead of amplitude contour because amplitude contour has been found to be highly correlated with F0 contour (Fu and Zeng, 2000; Whalen and Xu, 1992).

III. RESULTS
A. Examining individual cues

For each of 21 speakers, 4 tones, and 22 cues, we determine how good a particular cue is at distinguishing the specific tone from all the other tones.

We use Cohen’s $d$ as a measure of the significance of a particular cue (Cohen, 1988). Cohen’s $d$ is traditionally used as a measure of effect size. 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 the first group has $n_1$ data points with mean $\mu_1$ and variance $s_1^2$, and likewise for the second group. Then Cohen’s $d$ is

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

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

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$ (4)

Both the magnitude and the sign of Cohen’s $d$ is important. The magnitude indicates how large the difference is between the two groups and measures how useful the cue is for distinguishing the groups. The sign indicates which group has the larger value: a positive value of $d$ means that the first group has a larger value, and a negative value of $d$ means that the second group has a larger value.

For our situation, we let the first group consist of all cues from a particular tone, and the second group consist of all cues from the three other tones and then compute $d$. As a measure of the usefulness of the cue we take the absolute value of $d$. A positive value of $d$ shows that the tone is characterized by a greater value of the cue than for the other tones, whereas a negative value shows that it has a lesser value. We compute a $d$ for each cue, for each tone, and for each subject.

Figure 3 displays $d$ for the four tones. For each tone, the top plot shows a heat map of $d$ for each cue and subject. The

FIG. 2. (Color online) Top panel: Raw normalized F0 contours for one token of each tone for one speaker. Middle panel: The best least-squares fit of a parabola to each of the tokens in the top panel. Bottom panel: The best least-squares fit of a broken line to each of the tokens in the top panel.
color of each rectangle in the plot shows \( d \) for a particular cue for a particular speaker, with the color 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, as determined by MATLAB’s boxplot function, are shown as red crosses. Whiskers go to maxima and minima, excluding outliers. The box and its dividing line show quartiles (MATLAB, 2018).

We first discuss the top left of Fig. 3, showing the results for Tone 1. The top plot provides an idea of how \( d \) varies among cues and speakers. For example, it is immediately apparent that cue 2 (mean) has a large positive cue value for most speakers, and cue 12 (F0range) has a cue value that is smaller than for the other tones.

In Fig. 3, the bottom plot for Tone 1 provides more precise information about the distribution of \( d \) among all the subjects for each cue for Tone 1. As suggested by the top plot, cues 2 and 12 have a strong effect (large \(|d|\)) for all speakers.

Results for the other three tones are also shown in Fig. 3. For Tone 2, an examination of Fig. 3, top right, shows that cue 3 (slope), cue 5 (BLstart), cue 11 (overall), and a few others stand out as having a large value of \( d \) for all speakers. For tone 3, we see in Fig. 3, bottom left, that cue 15 (F050percent) has a negative \( d \) value of magnitude greater than 2 for all speakers. Cue 2 (mean) has a similar pattern as Cue 15, but is not quite as strong or consistent across speakers. For Tone 4, Fig. 3, bottom right shows that there are many cues with large \(|d|\) values. Two important ones are cue 3 (slope) and cue 11 (overall).

As a way of systematically listing the most important cues for a given tone, we list the five cues with the largest median \( d \) value in absolute value in Table II. The parabolic
fit cues 2–4 are prominent in the list, as are the raw F0 cues 13–17. The intensity cues 20–22 do not appear, nor does duration (cue 1).

### B. Identifying redundant cues

In reality, the task of identifying a tone does not involve considering a single cue in isolation. There is no single cue that is able to distinguish all four tones of Mandarin (unlike, say, how mean F0 may be able to do so in a language with level tones such as Yoruba). So it is of interest to identify a small number of cues that together are able to distinguish all four tones for all speakers. But we cannot just pick important cues independently, because some cues are highly correlated. Adding an additional cue can only improve the ability to distinguish tones if it is to some extent independent of the cues already being used. For example, in Table II both cue 2 (mean) and cue 15 (F050percent) appear twice, suggesting that they (among other cues) are useful for distinguishing tones. But both effectively measure height, so using them both is likely to be redundant. To quantify this redundancy, we compute the correlation matrix among the 22 cues over all tokens from all speakers and all tones. Figure 4 shows the correlation of all cues across all tokens. Recall that the diagonal entries of a correlation matrix are always 1. As can be seen by the off-diagonal entries that are close to 1 in value, there are several highly correlated cues. For example, cue 2 and cue 15 have a correlation of 0.89.

One way to handle and interpret correlations between variables is Principal Component Analysis (PCA). From a set of correlated variables, a set of uncorrelated variables that are linear combinations of the original variables is generated. One shortcoming of PCA is that since the principal components are weighted combinations of all the variables, this hinders interpretability. This is especially a problem given that we want to select a small number of cues for further study. One remedy for this is sparse PCA (Zou et al., 2006). See Friedman et al., 2001, Sec. 14.5.5 for an explanation of the advantages of sparse PCA over regular PCA, and an illustrative example of its use. Sparse PCA computes a small number of principal components balancing the desiderata of orthogonality of components, sparsity of the components, and the components sequentially capturing maximal unexplained variance. Here sparsity means every component is a linear combination of just a few of the cues. We computed some sparse principal components for our tokens, ignoring the tone labels. We used the routine spca from SpaSM, a MATLAB toolbox for sparse statistical modeling (Sjöstrand et al., 2018). Using trial and error, we found that asking the sparse PCA algorithm to find 8 components and five cues with non-zero weight in each component yielded interpretable, sparse, and approximately independent clusters. We used the result of the algorithm to cluster the variables into 8 groups, assigning a cue to whichever cluster in which it was most heavily weighted.

Table III shows a summary of the computed principal components in order of their importance of explaining variation in the dataset. The analysis also provides a clustering of correlated cues. For example, the dominant component contains the cues 3, 8, 10, 11, 17 as basically slope and four other cues that correlate highly with it. The next component consists of cue 2 (mean) and its correlates. The third component is cue 4 (curve) and its correlates. Then comes mean and max intensity, which naturally are highly correlated, followed by components consisting of single cues. The remaining component combines a group of variables that do not explain much variation. The last column in Table III states which cue is most “heavily loaded” in the component.

The results of the sparse PCA analysis support using a set of variables that only includes one cue from each of the three dominant clusters.

---

### Table II

For each tone, the five most important cues in distinguishing it from the other three, ranked. The +/- indicates whether the cue is typically larger (+) or smaller (-) for the distinguished tone.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Tone</th>
<th>Cues</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 (+)</td>
<td>mean 5 (-) BLstart 15 (-) F050percent 11 (-) overall</td>
</tr>
<tr>
<td>2</td>
<td>16 (+)</td>
<td>F075percent 3 (+) slope 2 (-) mean 3 (-) slope</td>
</tr>
<tr>
<td>3</td>
<td>12 (-)</td>
<td>F0range 11 (+) overall 14 (-) F025percent 10 (-) offglide</td>
</tr>
<tr>
<td>4</td>
<td>15 (+)</td>
<td>F050percent 18 (+) maxLocation 9 (-) onglide 8 (-) BLslope2</td>
</tr>
<tr>
<td>5</td>
<td>6 (+)</td>
<td>BLslope1 19 (-) minLocation 4 (+) curve 17 (-) F0offset</td>
</tr>
</tbody>
</table>

---

FIG. 4. (Color online) Correlation matrix of all 22 cues over all tokens for all speakers and tones shown using a heat map; the color of the entry at the ith row and jth column indicates the correlation of the ith and jth cue in this dataset.
TABLE III. Summary of the results from Sparse PCA analysis of 22 cues from our tokens. The first column lists the cues in each component and the second column gives a loose description of those cues. The third column gives the most heavily loaded cue in the cluster.

<table>
<thead>
<tr>
<th>Cues in cluster</th>
<th>Interpretation</th>
<th>Max loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>3, 8, 10, 11, 17</td>
<td>slope and its correlates</td>
<td>11 - overall</td>
</tr>
<tr>
<td>2, 14, 15, 16</td>
<td>mean and its correlates</td>
<td>2 - mean</td>
</tr>
<tr>
<td>4, 5, 6, 9, 13</td>
<td>curve and its correlates</td>
<td>9 - onglide</td>
</tr>
<tr>
<td>20, 21</td>
<td>mean and max intensity</td>
<td>21 - maxIntensity</td>
</tr>
<tr>
<td>7</td>
<td>location of BL breakpoint</td>
<td>7 - BLbreakpoint</td>
</tr>
<tr>
<td>22</td>
<td>location of max intensity</td>
<td>22 - locationMaxIntensity</td>
</tr>
<tr>
<td>1</td>
<td>duration</td>
<td>1 - duration</td>
</tr>
<tr>
<td>12, 18, 19</td>
<td>miscellaneous</td>
<td>18 - maxLocation</td>
</tr>
</tbody>
</table>

C. Selecting sets of cues

Because we are considering only 22 cues, it is feasible to go through all pairs of cues and estimate how good they are at distinguishing tones, and thereby select the best pair according to some standard.

To do this we need a way of measuring how good a pair of cues is. Cohen’s $d$ does this for one cue and a pair of tones, but we need to generalize this to multiple cues and four tones. One way to measure how good a set of cues is at distinguishing tones is to see which cues allow a standard classification algorithm to best distinguish tones from each other. We choose a simple popular classification algorithm, LDA. Given cues for a set of data points with class labels, LDA first finds a linear transformation of the space of stimuli into different regions according to the classes. The output is a division of the space into regions, which we will refer to as classification accuracy. If, given a set of cues, the categories are perfectly separable by lines, then the classification accuracy will be 100%. If the cues are not at all informative, then the classification accuracy will be 25% in our case, since we have 4 categories with equal numbers of tokens.

In Table IV we show the top 5 pairs of cues that have the highest classification accuracy. We see that the top pair is (slope,curve). The next four pairs are nearly as good, but interestingly, each consists of slope or another cue highly correlated with slope and curve or another cue highly correlated with curve. For example, the fifth most highly rated pair is (offglide,overall) = (10,11); the correlation of offglide with curve is 0.93, and the correlation of overall with slope is 0.97. Since we will advocate for the use of parabolic cues as canonical cues for studying Mandarin tone contours, it is interesting to compute the LDA classification accuracy for the other pairs of parabolic cues, namely the pairs of cues (mean,slope) and (mean,curve). Both of these are significantly less accurate than the pair of cues (slope,curve).

In Fig. 5 we show plots of all tones for all speakers using all three pairs of parabolic cues. There is an ellipse for each tone for each speaker. Each ellipse is centered at the mean of all cues for the given tone and speaker. The magnitude and orientation of the ellipse represents the covariance of the tokens about the mean. (The ellipse is plotted at one standard error of the mean.) We see that for the (slope, curve) pair the ellipses for distinct tones can be separated from each other by straight lines, with two exceptions. The first exception is that one speaker (S16) is using an allophone of Tone 3 that looks like Tone 4. The second exception is that some speakers’ Tone 2 overlap with other speakers’ Tone 3. This is not surprising given the confusability of Tones 2 and 3.

A natural suggestion is to determine if including a third cue can improve the distinguishability of the tones. There are 1540 trios of distinct cues, few enough that we can repeat our calculations with the pairs of cues with all trios. In Table V we show the results for the top five trios in terms of LDA classification accuracy. Remarkably, only a trio consisting of slope (3), curve (4), and mean intensity (20), is able to improve the classification accuracy over that of the best pair (3 and 4), and then only marginally. Accordingly, we take this as a strong indication that most of the important variation between tones is captured in just two cues.

IV. DISCUSSION

We now review the results of the previous section with the aim of providing a rationale for selecting the three parabolic cues as canonical cues for studying tone contours.

First we considered single cues alone and used Cohen’s $d$ to determine how good they were at distinguishing tones. This identified many effective cues, including the parabolic cues. But no single cue is sufficient to distinguish the tone contours of Mandarin, and so we need to consider sets of cues. We want cues in our set that have a large Cohen’s $d$ for some tones, but also that convey different information about the tone contour, rather than groups of all highly correlated cues. We used sparse PCA to identify groupings of cues that are redundant. We found that the three dominant groupings were (i) mean and cues correlated with it, (ii) slope and cues correlated with it, and (iii) curve and cues correlated with it.
Selecting at most one cue from each of these groupings is a good way to avoid redundancy, but the issue remains of which cues to select. Since there is not a natural analogue for Cohen’s $d$ with more than two cues, we turned to LDA. For each pair of cues, we built a LDA classifier that only used those two cues and then observed its classification accuracy. This allowed us to rank all pairs of cues according to this classification accuracy. This method picked out slope and curve as the best pair of cues. We further investigated whether adding a third cue improved LDA classification accuracy appreciably. Adding a further cue did not significantly improve classification accuracy even in the best case, and usually worsened it. The lack of significant improvement in the best case is likely due to something observed going back to Gandour (1983), that the perceptual space of Mandarin tone is two-dimensional (Chandrasekaran et al., 2010; Gauthier et al., 2007; Peng et al., 2012).

These two results might seem contradictory at first: the sparse PCA analysis yielded mean (and its correlates) as an important variable capturing as much of the variation in the tokens, but the LDA analysis found that mean was not so useful for distinguishing between tones. The reason for the difference is that sparse PCA studies the variation of the tokens ignoring the tone labels, whereas the LDA analysis only selects cues that distinguish between tone categories.

In other words, slope and curve were important for capturing variation between tones; mean was less important for this, but was important for capturing variation between tokens. Much of the between-token variation captured by mean is within tone categories, and thus not useful for classification. Despite this, we recommend using mean as a canonical cue for tone contours, because of its simplicity, because it has been used so extensively already (Gandour, 1983; Jeng et al., 2006; Yang, 2015), and because the variation it describes is important for other reasons besides distinguishing tones, such as interspeaker variation and different vocal styles (Zhao and Jurafsky, 2009).

The critical tonal cues as determined by our approach are consistent with previous findings showing the relevance of F0 height and contour cues (e.g., Smith and Burnham, 2012; Yang, 2015), particularly in terms of F0 slope (e.g., Flemming and Cho, 2017) and curve (e.g., Li and Chen, 2016). Our approach further advances tone characterization by the following new findings. First of all, this study shows that among a large number of tonal cues, a small set of three cues does a good job of characterizing Mandarin tones.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Cue 1</th>
<th>Cue 2</th>
<th>Cue 3</th>
<th>LDA classification accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>4</td>
<td>curve</td>
<td>meanIntensity 89.5%</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>curve</td>
<td>BLbreakpoint 89.2%</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
<td>curve</td>
<td>offglide 89.2%</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>4</td>
<td>curve</td>
<td>locationMaxIntensity 89.1%</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>7</td>
<td>BLbreakpoint</td>
<td>overall 89.1%</td>
</tr>
<tr>
<td>134</td>
<td>2</td>
<td>3</td>
<td>slope</td>
<td>curve 83.3%</td>
</tr>
</tbody>
</table>

FIG. 5. (Color online) Speaker means for each of the four tones plotted using each of the three pairs of parabolic cues: top (slope, curve) middle (mean, slope), bottom (mean, curve). Each ellipse shows all the productions of one tone for one speaker. The magnitude and orientation of the ellipse represents the covariance of the tokens about the mean.

TABLE V. The top five trios of cues for maximizing classification accuracy of LDA on all tokens. The result for the trio of parabolic cues is also shown.
Further reveals that cues can be clustered into groups such that those within each group are functionally equivalent, resulting in multiple possible choices of 3-cue sets that can effectively distinguish different tones. Moreover, our analysis brings the parabolic cues to the forefront. The three cues obtained from the parabolic fit (F0 mean, slope, and curve) are considered the optimal cue set since they are easy to compute, directly interpretable, and among the best for distinguishing tones.

On the practical front, there are additional reasons besides our statistical analyses to favor using the three parabolic cues as a standard for studying tone contours. First, there is value in having a small critical set of standard cues whatever they may be. If different groups use different cues then it is difficult or impossible to compare results from different papers. For example, if an experimental manipulation is found to lead to a shift in turning point of Tone 2 in one study and an increase in curve in another, it is possible that they are observing the same phenomenon, but it is difficult to know for sure without reperforming the analysis with the raw data. Another reason to use only a small number of cues is to reduce the temptation to select the cues for any dataset that give the most significant results, thus leading to false positive results.

Finally, fitting a parabola to a curve is very well understood mathematically and statistically (unlike the nonlinear regression required for the broken line fit). There are efficient algorithms for computing them available in every mathematical software system. The fact that our three cues are based on basic mathematics (the Legendre polynomials being standard tools used in many contexts) and not any specific facts about Mandarin tone contours means that they can be generalized to other languages and tone systems.

Our results also have implications for the relationship between tone perception and production. Lexical tone perceptual studies have been largely influenced by Gandour (1983), which determined that tone perception involves “height” and “direction” as the two perceptual dimensions that have been interpreted as F0 mean and slope, respectively (e.g., Chandrasekaran et al., 2010; Francis et al., 2008; Guion and Pederson, 2007; Jongman et al., 2017). Moreover, native Mandarin listeners give stronger perceptual weightings to “direction” than to “height” (Francis et al., 2008; Gandour, 1983). Our study corroborates these findings in that slope was important for capturing between-category classifications of our participants’ tone productions, whereas curve as a tone perceptual cue has not been widely explored. Perceptual cues that are related to curve include TP and ΔF0, since any change in these cues should lead to a change in the parabolic shape of the tone contour and have been shown to influence the perception of Tone 2 and Tone 3 (Moore and Jongman, 1997). A recent study by Leung and Wang (2018) also shows that curve demonstrates a stronger correlation between the production and perception of Tone 2 than slope. However, further research is needed to examine the role of curve in the perception-production link of all Mandarin tones, as well as other tone languages.

The current results also have implications for establishing the relationship between tone acoustics and articulation. Research has revealed that, during tone production, facial (e.g., head, eyebrow, lip) movements in distance, direction, speed, and timing can be spatially and temporally equated to acoustic features of tonal changes in F0 (Attina et al., 2010; Garg et al., 2019). For example, Garg et al. (2019) shows that the upward and downward head and eyebrow movements follow the rising, dipping, and falling tone trajectories for Mandarin contour tones (i.e., Tone 2, Tone 3, and Tone 4, respectively). Additionally, the time taken for the movements to reach the maximum displacement is also aligned with these trajectories. These results suggest that the spatial and temporal dynamics in the articulation of different tones may particularly be aligned with the parabolic cues identified by the current study as the critical cues in representing tonal contours, reflecting a linguistically meaningful association between spatial and acoustic events in lexical tone production. Further research may thus explore how these cross-modal (articulatory-acoustic) resources can be incorporated in tone perception.

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 Halakoei, Abner Hernandez, Michelle Le, Kelsey Philip, Michelle Ramsay, and Joanna Xie for conducting acoustic data measurements. Portions of this research were presented at the 176th Meeting of the Acoustical Society of America and 2018 Acoustics Week in Canada, Victoria, British Columbia, Canada.


