Speech Communication 113 (2019) 47–62

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Computer-vision analysis reveals facial movements made during Mandarin tone production align with pitch trajectories

Saurabh Garg[a,d], Ghassan Hamarneh[b], Allard Jongman[c,*], Joan A. Sereno[c], Yue Wang[d]

[a] Pacific Parkinson's Research Centre, University of British Columbia, Canada
[b] Medical Image Analysis Lab, Simon Fraser University, Canada
[c] KU Phonetics and Psycholinguistics Lab, Department of Linguistics, University of Kansas, USA
[d] Language and Brain Lab, Department of Linguistics, Simon Fraser University, Canada

ARTICLE INFO

Keywords:
Mandarin lexical tone
Visual cues
Facial movements
Features
Tone classification
Discriminative analysis
Computer vision
Image processing
Machine learning

ABSTRACT

Using computer-vision and image processing techniques, we aim to identify specific visual cues as induced by facial movements made during Mandarin tone production and examine how they are associated with each of the four Mandarin tones. Audio-video recordings of 20 native Mandarin speakers producing Mandarin words involving the vowel /3/ with each of the four tones were analyzed. Four facial points of interest were detected automatically: medial point of left eyebrow, nose tip (proxy for head movement), and midpoints of the upper and lower lips. The detected points were then automatically tracked in the subsequent video frames. Critical features such as the distance, velocity, and acceleration describing local facial movements with respect to the resting face of each speaker were extracted from the positional profiles of each tracked point. Analysis of variance and feature importance analysis based on random forest were performed to examine the significance of each feature for representing each tone and how well these features can individually and collectively characterize each tone. Results suggest alignments between articulatory movements and pitch trajectories, with downward or upward head and eyebrow movements following the dipping and rising tone trajectories respectively, lip closing movement being associated with the falling tone, and minimal movements for the level tone.

1. Introduction

This study aims to identify the visual-articulatory features of Mandarin Chinese tones. Mandarin employs tone, a prosodic entity, to convey lexical meaning. Tones are acoustically manifested by changes in fundamental frequency (F0, perceived as pitch) primarily as well as duration and amplitude, which are triggered by glottal and sub-glottal activities independent of vocal tract configurations (Howie 1976; Yip 2002). This poses the question as to whether tones are visually distinctive.

Substantial research has shown that complementary information gathered from visual cues provided by speakers' facial movements, particularly lip movements such as opening, rounding, and spreading, can strengthen the signal quality of speech segments and facilitate speech segmental perception (Kim et al., 2014b; Perkell et al., 2002; Tang et al., 2015; Traunmüller et al., 2007). On the other hand, research has not been conclusive about how prosody, including tone, may benefit from visual information, presumably because prosodic production does not rely on vocal tract configurations and may thus be less visually salient. There has been evidence that head, jaw, eyebrow, neck, and lip movements may convey visual information in tonal and/or general prosodic production and perception (Attina et al., 2010; Burnham et al., 2001; Chen et al., 2008; Cvejic et al., 2010; Kim et al., 2014a; Munhall et al., 2004; Swerts et al., 2010; Yehia et al., 2002). However, the extent to which such movements provide linguistically meaningful cues to signal tonal category distinctions or are general attention-grabbing cues is not clear. In particular, research has not agreed on which specific movements are used to characterize the visual differences of different tones, or on which methods can effectively identify and quantify these visual tonal distinctions.

The present study systematically examines how visual cues are employed in Mandarin tone production, using state-of-the-art computer-vision and image processing techniques. On pre-recorded video, we identify specific visual cues induced by facial movements in the production of each tone, measure the manner and extent of these movements using both distance- and time-based metrics, and rank their relative prominence in characterizing each tone.

* Corresponding author.
E-mail address: jongman@ku.edu (A. Jongman).

https://doi.org/10.1016/j.specom.2019.08.003
Received 23 April 2019; Received in revised form 21 July 2019; Accepted 12 August 2019
Available online 17 August 2019

1.1. Mandarin tones

Four lexical tones are used in Mandarin. They can be described by F0 contour as level, rising, and falling; and F0 register as high, mid, and low (Howie 1976). Tone 1 (high-level) has a steady, high F0 contour; Tone 2 (mid-high rising) briefly falls to mid-high and rises to a high F0 level; Tone 3 (mid-low dipping) falls to mid-low and rises to mid-high F0; and Tone 4 (high-falling) begins at a high F0 level and drops quickly to a low F0 level (Chao 1948; Howie 1976; Wang et al., 2003). Mandarin tones also vary in duration, with Tone 3 being the longest and Tone 4 the shortest (Lin 1965). Additionally, it has been observed that the F0 turning point (the point in time at which the F0 contour changes from falling to rising) for Tone 2 occurs earlier than for Tone 3 (Dreher et al., 1968). Further research reveals that the acoustic cues typically used in perception not only include static cues such as F0 height and contour direction (Gandour 1983), turning point F0 and time (Moore et al., 1997), and overall duration (Blicher et al., 1990), but also dynamic cues characterizing F0 slope and contour shape, such as the velocity and acceleration of F0 fall or rise (Krishnan et al., 2009; Prom-on et al., 2009; Prom-on et al., 2012; Xu et al., 2006). These acoustic tonal features may be articulatorily manifested as spatial and temporal changes in the distance, direction, duration and speed of movements, since pitch has been claimed to be audio-spatial in representation (Connell et al., 2013; Hannah et al., 2017).

1.2. Prosodic and tonal visual cues

Research has demonstrated that head movements occur more frequently and are larger in prosodic constituents with a larger amount of variance in F0 (Munhall et al., 2004; Yehia et al., 2002), for example, in sentences with strong focus (Kim et al., 2014a; Swerts et al., 2010), stressed syllables (Scarborough et al., 2009), and interrogative intonation (Srinivasan et al., 2003). Head motion has also been found to be associated with F0 in lexical tone production. Burnham et al. (2007) showed that head movements (e.g., nodding, tilting, rotation towards the back), as computed from the principal component analysis on kinematic sensor data, were correlated with F0 changes in Cantonese tones. Using the same approach, Attina et al. (2010) further found back and forth head movements to be correlated with F0 modulation of contour tones (Tones 2–4 in Mandarin), while head nodding was correlated with production of a high tone (Tone 1 in Mandarin). However, since the data in these studies were not quantified in terms of the direction and magnitude of movement, it is not clear if and to what extent these head movements correspond to specific changes in tone (F0) height and contour direction. One piece of evidence showing a directional association between head movement and F0 is the case of Tone 3 in Mandarin. Chen et al. (2008) reported improvement of Tone 3 identification when perceivers' attention was directed to the head dipping movement (as well as movements in the neck) in the production of this tone. Similarly, a lowered jaw position was found in the production of a low tone (Tone 3) in low vowel contexts (Shaw et al., 2014). These patterns suggest a positive correlation between head/jaw movements and changes in F0 in the production of prosodic tonal variations. It has been speculated that head and jaw lowering or raising can be triggered by the reduced or tightened vocal folds (movements of the cricothyroid muscle and ligaments) associated with low- or high-pitched tones (Moisik et al., 2014; Smith et al., 2012; Yehia et al., 2002). However, quantitative data are still needed to further determine the magnitude and trajectory of head/jaw movements in individual tone articulation.

Eyebrow movements have also been observed to be associated with prosodic articulation (Cvejic et al., 2010; Kim et al., 2014a; Munhall et al., 2004; Swerts et al., 2010; Yehia et al., 2002), although no research has focused on tone. Data from kinematic measures reveal larger vertical eyebrow displacement (from the neutral baseline position) and higher peak velocity of eyebrow movements for the focused word in a sentence (Kim et al., 2014a). Furthermore, eyebrow raising has been shown to occur more frequently and align better with accented than unaccented syllables, and with strongly than weakly accented syllables (based on video analysis by human annotators, Flecha-García 2010; Swerts et al., 2010); and to be greater in displaced distance for phrasal stress (by measurements of eyebrow displacements acquired through motion tracking, Scarborough et al., 2009). These results indicate that eyebrow movements may be coordinated with F0 for prosodic contrasts, although their specific relevance to F0 changes (in terms of height and direction) is not straightforward or invariably evident (Ishi et al., 2007; Reid et al., 2015). Although they did not specifically focus on prosody, Huron et al., (2013) did report a causal relationship between vertical eyebrow displacement and F0 height through manipulation of eyebrow movements. By instructing the speakers to raise or lower their eyebrows to different degrees during reading, the authors found higher eyebrow placement to be associated with higher vocal pitch. These results motivate the inclusion of vertical eyebrow movements in the present study as potential correlates of tone height and direction.

Lip movements typically signal segmental rather than prosodic contrasts, since the articulation of prosody does not rely on vocal tract configuration. Nonetheless, there has been evidence that lip movements may be spatially and temporally aligned with prosodic changes (Dohen et al., 2005; Dohen et al., 2006; Scarborough et al., 2009). For example, Scarborough et al. (2009) found that the largest magnitude of lip movements (in terms of lip opening displacement as well as inter-lip distance) was associated with lexical and phrasal stress. For Mandarin tone production, Attina et al. (2010) reported a general correlation between lip closing and F0 (irrespective of tones), as well as unique patterns for individual tones (Mandarin Tones 1 and 2). In particular, Tone 1 was characterized by lip raising (as well as jaw advancement), suggesting a potential link between these movements to the height or the lack of contour of this high-level tone; in contrast, Tone 2 production was mainly distinguished by lip protrusion, which presumably could be related to the rising contour. Mixdorff et al. (2005) tested the perceptual relevance of lip movements in Mandarin tone production. By showing only the lower half of the speaker's face, perceivers were forced to focus on articulatory movements of the lips and chin. Results show that tone identification improved significantly when additional visual information from the chin and lips were provided compared to the audio-only condition, suggesting facilitative effects of lip and chin movements in tone perception. However, from of all these studies, it remains unclear what kinds of lip movements characterize each individual tone, and whether and how they correspond to changes in tone height and contour.

Taken together, these results collectively suggest that specific movements of the head, eyebrows and lips are correlated with tonal articulation, and are likely coordinated with the spatial and temporal dynamics of the production of different tones. The current study thus examines visual cues to Mandarin tone production to systematically quantify the displacement distance, time, and kinematics in order to determine the magnitude, direction, and manner of these movements in individual tone articulation, as well as how they characterize each tone.

1.3. Analysis methods in prosodic and tonal studies

Research has not been consistent with respect to the methods used to acquire and analyze articulatory movement data. One traditional method involves annotated video analysis conducted by human annotators (e.g., Flecha-García (2010), in their study of pitch accent articulation). The constraint of this method is that only those movements that are observable by the annotators of a particular study are logged. As such, data are not only judged in a subjective manner, but are also limited to frequency judgment (e.g., number of occurrences of eyebrow rise) and thus preclude examination of the intensity or magnitude of the motion.

Sensor-based devices enable acquisition of more precise and quantified data. Shaw et al. (2014) used electromagnetic articulography (EMA) in their study of tone-vowel coproduction. By tracking the flesh points

(markers) attached to speakers' tongue, lips, and jaw, it is possible to measure the spatial variations such as jaw displacement and tongue-to-jaw distance with millimeter precision. A commonly used sensor-based motion capture system is the OPTOTRAK (Northern Digital Inc.) system, which involves infrared emitting markers positioned on various locations on the head (e.g., Burnham et al., 2007; Kim et al., 2014a; Attina et al., 2010). For example, Kim et al. (2014a) used OPTOTRAK to capture eyebrow and jaw movements for sentence focus and quantified the movements in terms of displacement and peak velocity. Similarly, in Scarborough et al. (2009), retro-reflectors were attached to the speaker's face for recording by the Qualisys motion capture system, allowing analysis of lip, eyebrow, and head displacement magnitude and movement velocity. However, the use of sensor-based motion tracking for data analysis has several limitations. First, since sensors are physically attached to the speaker's face, speakers may experience discomfort and consequently exhibit unnatural facial movements. Another limitation with the use of sensor-based analyses is that only the regions where sensors are placed can be examined. Additionally, although quantification of movements is possible using a sensor-based approach (Kim et al., 2014a; Scarborough et al., 2009), no such data have been reported for tones (e.g., Burnham et al., 2007; Attina et al., 2010).

To overcome the limitations of the previous video- and sensor-based approaches, our research team has developed an approach using computer-vision and image-processing techniques, which bypass the involvement of human annotators or use of sensors (Tang et al., 2015). This approach has been employed successfully in determining the facial cues for English vowel production (Tang et al., 2015). Specifically, the techniques combine the use of video capture and image analysis algorithms to respectively record and extract features that describe facial movements. This marker-free approach makes it possible for speakers to speak naturally, for any region on the speaker's face to be tracked and analyzed, and for the same speaker's face to be directly used in perception studies.

1.4. The present study

The present study focuses on several underexplored research directions for understanding the role of visual cues in Mandarin tone production. Firstly, it is unclear which specific visual cues are used in characterizing individual Mandarin tones and which cues make each tone distinct from the other tones. Secondly, research has not systematically quantified the visual cues in tone production in terms of the magnitude, direction, relative time and kinematics in association with tone height and contour changes. Thus, the extent to which such cues are linguistically meaningful has not been determined. Moreover, regarding methodology, since previous research employed a variety of data acquisition and analysis techniques and focused on different facial regions, there is a lack of consistency and comparability in the findings across studies.

The present study aims to address these questions using computer-vision and image-analysis techniques to systematically identify and examine the features extracted from motion captures of speakers' productions of Mandarin tones in single words. Accordingly, our objectives are to: (1) present a computerized framework that can systematically characterize facial movements of speakers made during tone production as captured in videos; (2) determine a set of features that describe the local facial movements associated with the production of each tone; (3) examine the representation power of each feature for each tone; and (4) explore how well these features can individually and collectively characterize each tone via traditional statistical methods and machine-learning algorithms.

With respect to methodology, the present approach extends the Tang et al. (2015) study on segments to the prosodic domain, namely tone. This involves motion tracking of additional anatomical regions (e.g., eyebrows in addition to lips) and using additional measures and (machine-learning) algorithms for analysis, based on a collection of features previously shown to be relevant for general prosodic or tonal articulation (Attina et al., 2010; Kim et al., 2014a; Krishnan et al., 2009; Scarborough et al., 2009). Specifically, a set of features based on the distance, relative time, and kinematics such as velocity and acceleration of the keypoints on the head, eyebrows, and lips were extracted from videos and subsequently analyzed. This approach may thus help us identify the unique features characterizing each tone in Mandarin.

2. Materials

2.1. Speakers and stimuli

Twenty native Mandarin speakers (12 female and 8 male) between the ages of 18–28 years old with an average age of 22.6 years were recruited. All the speakers are native speakers of Mandarin and residing in Canada for less than five years. From each of the speakers, approximately 100 pronunciations of tone quadruplet words were recorded in three sessions with two breaks. The monosyllable /3/ with four Mandarin tones was used in this study, carrying the meaning of graceful (/$\bar{3}$/; Tone 1, level tone), goose (/$\acute{3}$/; Tone 2, rising tone), nauseous (/$\check{3}$/; Tone 3, dipping tone) and hungry (/$\grave{3}$/; Tone 4, falling tone) respectively. In addition to the /3/ word we also recorded /i/ and /u/ words as fillers. Speakers were asked to read out each of these monosyllabic tone words that was presented on the screen first in a plain speech style and then in a clarified speech style. Both styles were included in the analyses to provide a broader range of within-category representation of each tone. Each word was presented individually. The average duration of the target stimuli was 580 ms (SD = 193 ms) across styles, tones and speakers.

2.2. Data acquisition

All recordings were made in a sound-attenuated booth in the Language and Brain Lab at Simon Fraser University. The stimuli for elicitation were displayed on a 15-in LCD monitor that was situated three feet in front of the speaker, positioned at eye-level to facilitate the placement of a front-view video camera, which was placed below the monitor on a desktop tripod. Each speaker was recorded individually and was instructed to sit with his/her back against a monochromatic green backdrop. High definition front-view video recordings were made with a Canon Vixia HF30 camera, recording at a frame rate of 29 fps.

3. Methods

3.1. Video analysis

3.1.1. Overview

Our fully automatic video-analysis was implemented using MATLAB and consists of the following steps: (1) video segmentation; (2) point-detection to locate the keypoints on the face viz. the medial end of the eyebrow, the nose tip and the cupid bows of the lips; (3) keypoint tracking to record the spatial coordinates of the detected points over time (over the duration of the utterance of the token); (4) feature extraction from the tracked keypoints.

3.1.2. Video segmentation

Tone utterances were segmented first using automatic tools. Segmentation of the video tokens was based on the audio signal (Garg et al., 2018). Only the portion of the video frames whose audio power (amplitude) was above a certain threshold ($\rho = 0.2\times$ maximum value) was extracted. This was done to remove any extraneous noise from the recording (such as cough sounds or keystrokes of keyboard) and to keep only the audio corresponding to the word spoken. The value of the threshold was empirically decided and was set at 20% of the maximum value found in the token. We further added a fixed amount of buffer on both

sides (10 frames, approx. 0.3 s) of the video token to compensate for imprecision errors due to segmentation. Each video token was then evaluated by two native Mandarin speakers for quality of the video (e.g., no eye blinking during production). Lastly, each token was rated by two other native Mandarin speakers for intelligibility of the audio signal. Only the tokens that were correctly perceived as the intended tone words were included. Across all tokens, 1.85% were rated as incorrectly pronounced and thus excluded from further analysis.

3.1.3. Detection of regions-of-interest and keypoints

After segmentation of the tokens, a set of keypoints were identified that would be tracked in the video during the utterance. We examined the movements at three regions-of-interest (ROI): the head, eyebrow (we randomly selected the left side), and the lips. At first, a rough bounding box on each relevant facial part was localized using the cascade filter approach of Lienhart et al. (2003). Subsequently, part-specific detectors were used to obtain better localizations of the ROIs. Details are provided below for each of the ROIs examined.

Face: Before detecting any parts of the face, the face itself needed to be detected. For this purpose, a set of Local Binary Pattern (LBP) cascade filters (Ojala et al., 2002) was used to obtain an initial set of possible candidates. Since we know at any time there is only one face in the video recording, we used different merge-thresholds in increasing order until the filter provided one output bounding box. Using the threshold, groups of co-located detections that meet the threshold value were merged to produce one bounding box around the target object. Thus, if the threshold is low, fewer detections are merged and the cascade filter will suggest several face possibilities; if the threshold is high enough, it will merge similar detections and the detector will show fewer face possibilities. In our analyses, one fixed threshold did not seem to work for all video segments and for all speakers reliably. Hence, we automatically tried different threshold values until we obtained one face detection.

Once the face was detected, the search was narrowed down to the area present within the detected face region. Limiting the search space to the face region helped reduce the large number of false positive detections that could be present in the background.

Left eyebrow: We first detected the eyes by using a set of LBP cascaded filters of two types: the first type was used to detect both eyes which were then localized with a single bounding box, while the second type of filters were used to localize the left and right eye independent with two separate bounding boxes. Detecting the entire set of eyes helped narrow down the search space further. Similar to the face, merge threshold was adjusted for the pair of eyes as well to get one estimate of the location of eye-pair. Then we applied a left-eye specific filter to get different possibilities of the left eye. The left eye bounding box was then chosen based on the distance from the top left corner of the bounding box of the detected eye-pair to the left eye bounding box. The box that had a minimum distance was selected.

Once the left eye was robustly detected, the contour of the left eyebrow (superciliary ridge line) was identified. Several sources of information relating to edges, approximate position relative to the bounding box of the detected eye, and color (e.g., eyebrows are darker in color as compared to the skin) were used to estimate the left eyebrow. The estimated eyebrow contour was then refined using an active contour model (Chan et al., 2001).

Nose/Head: We used the nose as a proxy for head movement as it is rigid and treated any movements observed for the nose as head motion as done in previous studies (Cai et al., 2012; Tu et al., 2009). The nose was again detected using a cascade filter and adjusting the merge-threshold, until there remained one bounding box. Since the lower part of the nose has a well-defined edge, as a post-processing step, an active contour model was used to find the exact boundaries of the nose.

Lips: Lips were detected using the mouth cascade detector (Castrillón et al., 2007). Once the bounding box was obtained, the exact contour of lips was detected in HSV space (as opposed to RGB color space). This was done since the differences are much more amplified in HSV space. In HSV, a threshold value was used to estimate a rough contour around the lips which was further refined using active contour modelling.

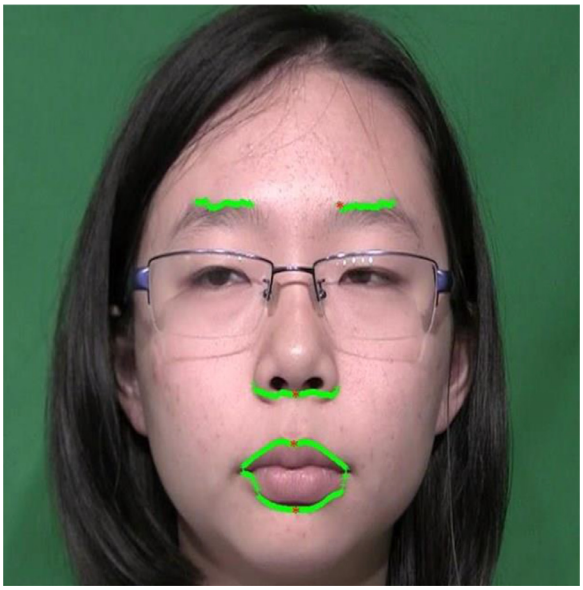


Fig. 1. Green points are those detected automatically by our keypoint-detection algorithm, four of which, shown in red, were then subsequently tracked for motion characterization of each region of interest.

For each ROI, several points of interest along the boundary of the specific part of the face were sampled. These points vary in number based on the length of the edge in the first frame as well as the distance of the speaker from the camera.[1]

3.1.4. Tracking of keypoints

Given the set of contours extracted for each ROI as described above, 1 to 2 keypoints on each contour were further extracted and motion-tracked. In particular, the eyebrow keypoint was computed as the geometric mean of the detected superciliary ridge contour of the eyebrow, while the nose keypoint was computed as the geometric mean of the detected nose contour. For the motion-tracking of the lips, one point on the upper lip (indicated by cupid's bow) and another point on the lower lip (indicated by the center point between the two oral commissures on the lower vermillion border) were detected and subsequently tracked. Examples of the extracted points on a randomly chosen video frame are shown in Fig. 1.

Once the aforementioned keypoints were identified on the first frame of each video token, they were tracked on the rest of the video frames using the Kanade-Lucas-Tomasi (KLT) feature-tracking algorithm (Lucas et al., 1981; Tomasi et al., 1991), which is a computationally efficient and robust registration technique that employs intensity gradients of each image frame to derive pixel-wise correspondences between two consecutive frames.

After performing tracking with the KLT algorithm, we obtained a set of motion trajectories for each keypoint. Any motion due to the head was subsequently removed from the eyebrow and the lips by subtracting the head displacements from the displacement of the eyebrow and lip keypoints.

3.1.5. Feature extraction

With a set of motion trajectories computed from the head, eyebrow, and the lips, we next computed a set of features selected to quantify

[1] The closer the speaker was positioned to the camera, the more points were sampled because the ROI would be larger.

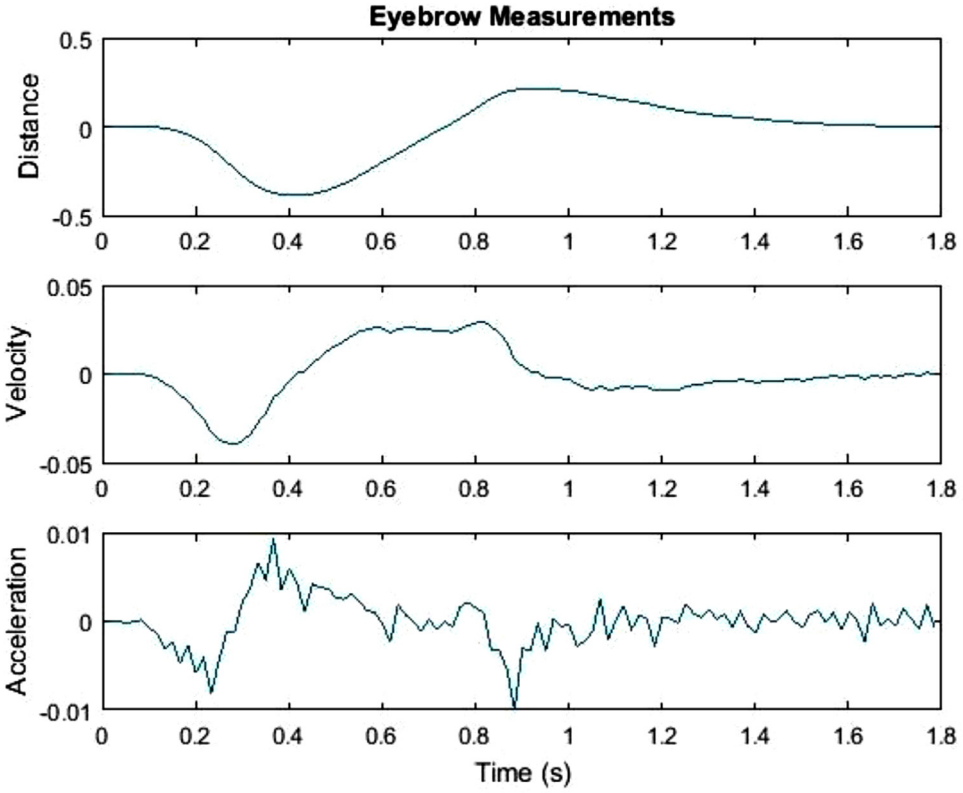


Fig. 2. The measurements shown are normalized measurements (normalized to the head size) and the units are in pixels, pixels/frame, and pixels/frame^2 for distance, velocity and acceleration, respectively.

the motion dynamics of each for the four tracked keypoints and to provide summary statistics of the local (eyebrows and lips) and rigid (head) movements.

These features can be broadly classified into three categories: (1) *distance-based*, such as the minimum and maximum total displacement of a keypoint from its initial resting position to a position marked by a target event; (2) *time-based*, such as the time it took the displacement of the keypoint of an ROI to reach maximum or minimum distance; and (3) *kinematic*, such as the velocity and acceleration of a keypoint at specific time instances marked by some target events (e.g., instance when velocity reached a maximum).

An example of how the trajectory of an eyebrow keypoint is used to compute the features is shown in Fig. 2. The plots in this figure show how distance (top), velocity (middle), and acceleration (bottom) of an eyebrow keypoint changed over time when a target word was uttered in a randomly selected video token. The top figure plots the distance traveled by an eyebrow keypoint over time with respect to its initial position. The middle and bottom figures, respectively, plot the velocity and acceleration at which the same keypoint was traveling.

A summary of all features extracted in this study is listed in Table 1. These features were chosen in an attempt to capture the different variations that could be introduced by the pronunciation of different tones.

These features were extracted only for the video frames where the speaker's voice was detected on the audio channel. Furthermore, all the feature values were normalized to account for inter-speaker differences in head size and differences in the distance of the speaker from the camera. Normalization was done by dividing the feature values by a normalization factor computed as the shortest distance between the line joining the two eyes and nose tip.

Note that the features were measured only in the vertical direction and thus any movements in the horizontal direction were not included. This is intentional as the raising and lowering of the pitch have been found to be associated only with vertical upward or downward articulatory movements (Kim et al., 2014a; Huron et al. 2013).

Fig. 3(a) shows a schematic diagram illustrating how the distance-based features and kinematic features are related. Fig. 3(b) further explains how the time-based features relate with the rest of the extracted features. In Fig. 3(b), the violet regions mark the time instances when a tracked keypoint moved downward and hence represents the lowering movements of the head and eyebrow, or closing of the lips, while the pink region represents rising movements of the head and eyebrow, or the opening of the lips. As distance-based features, we calculated the minimum and maximum distances that each of the tracked keypoints moved from its initial resting state. We also calculated their corresponding minimum and maximum velocities. Note that velocity is indicated by the slopes of the curve (i.e., computed as rate of change of the curve) and the acceleration is computed by the rate of change in velocity. We also computed the relative time at which the minimum or maximum occurred with respect to the total duration.

3.2. Two-part analyses of the extracted features

Two analyses were conducted. Part 1 of our analyses was a discriminant approach where we formulated a series of tone classification problems to allow us to directly relate which features best characterize each tone. With the most relevant features identified in Part 1, we then performed post-hoc analyses to examine the individual features on a per tone basis in Part 2.

The next section provides technical details on the discriminant analysis approach adopted for the tone classifications.

3.3. Analysis of the associations between features and tone class via a discriminant approach

Given our dataset of video tokens (each of which was represented by a set of features and a label corresponding to tone class), we used a one-versus-all (OVA) approach where we trained a random forest (RF) classifier to discriminate each tone from the other three tones using features extracted from the videos of the speaker's utterances.

Table 1
The set of features used to represent each video token. ROI is region of interest. Please refer to text for details.

ROI	Index	Category	How this feature is computed
Head	1	Distance	Maximum displacement of the head while head-raising from its starting position
Head	2	Distance	Maximum displacement of the head while head-lowering from its starting position
Head	3	Distance	Average distance head moved during the utterance
Head	4	Distance	Total distance traveled by head during the utterance
Eyebrow	5	Distance	Maximum displacement of the eyebrow keypoint from its starting position
Eyebrow	6	Distance	Maximum displacement of the eyebrow while eyebrow-lowering from its starting position
Eyebrow	7	Distance	Average distance eyebrow moved during utterance
Eyebrow	8	Distance	Total distance eyebrow moved during the utterance
Lips	9	Distance	Maximum lip-opening distance
Lips	10	Distance	Maximum lip-closing distance
Lips	11	Distance	Average distance lips moved during utterance
Lips	12	Distance	Total distance lip moved during the utterance
Head	13	Time	The relative time at which the displacement of the head while head-raising was maximum
Head	14	Time	The relative time at which the displacement of the head while head-lowering was maximum
Head	15	Time	The relative time at which the head velocity was maximum during head-raising
Head	16	Time	The relative time at which the head velocity was maximum during head-lowering
Eyebrow	17	Time	The relative time at which the displacement of the eyebrow while eyebrow-raising was maximum
Eyebrow	18	Time	The relative time at which the displacement of the eyebrow while eyebrow-lowering was maximum
Eyebrow	19	Time	The relative time at which the eyebrow keypoint reached maximum velocity during eyebrow-raising
Eyebrow	20	Time	The relative time at which the eyebrow velocity during eyebrow-lowering was maximum
Lips	21	Time	The relative time at which the amount of lip-opening reached maximum
Lips	22	Time	The relative time at which the amount of lip-closing reached maximum
Lips	23	Time	The relative time at which the lip velocity during lip-opening was maximum
Lips	24	Time	The relative time at which the lip velocity during lip-closing was maximum
Head	25	Kinematic	Maximum head velocity during head-raising
Head	26	Kinematic	Maximum head velocity during head-lowering
Head	27	Kinematic	Maximum absolute acceleration of the head
Eyebrow	28	Kinematic	Maximum eyebrow velocity during eyebrow-raising
Eyebrow	29	Kinematic	Maximum eyebrow velocity during eyebrow-lowering
Eyebrow	30	Kinematic	Maximum absolute acceleration of the eyebrow
Lips	31	Kinematic	Maximum lip velocity during lip opening
Lips	32	Kinematic	Maximum lip velocity during lip closing
Lips	33	Kinematic	Maximum absolute acceleration of the lips

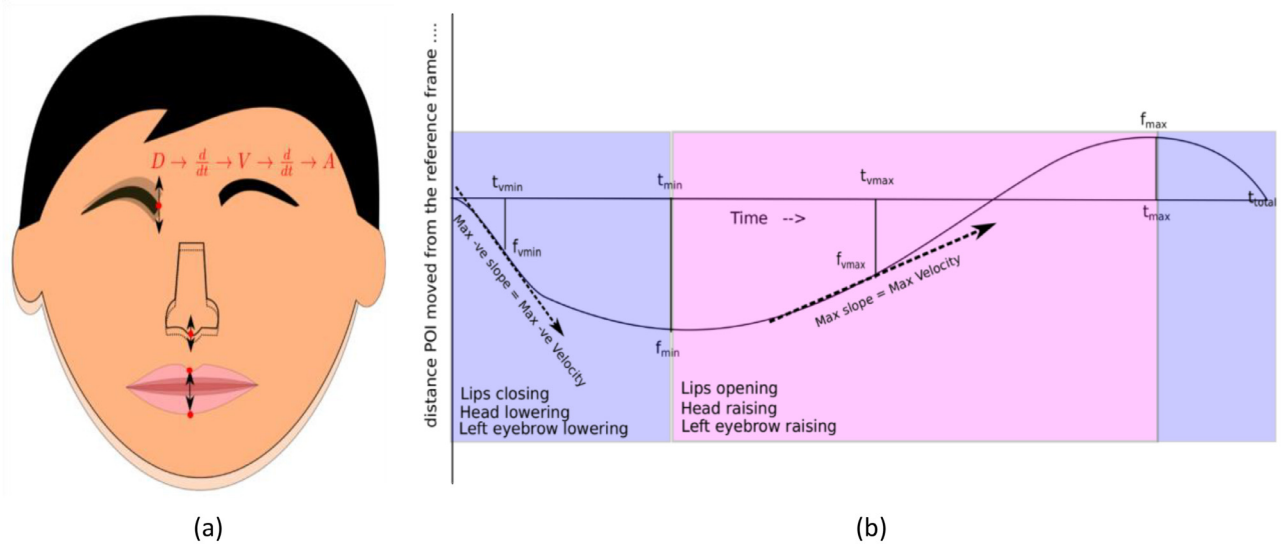


Fig. 3. (a) Schematic diagram showing the position and direction of movement of the 4 tracked points. The shadows near eyebrow/jaw/nose indicate motion. (b) A visual summary of the different features we extracted from the 4 tracked keypoints. POI is position of interest. Please see main text for details.

Our choice of random forest as the classifier was made based on the work by Paul et al. (2015). Briefly, RF operates by training and subsequently deploying an ensemble of *t* simple decision trees, each of which predicts (or outputs) a class label given an input pool of features such that the final class label is obtained by computing the mode of the class label predictions obtained from each tree. Randomness is injected by training each tree with a random subset of samples and another random subset of features.

RF has been used in many binary and multi-class classification problems such as emotion recognition (Noroozi et al., 2017). One of RF's key advantages over its counterparts is its inherent ability to provide a ranking of importance of each input feature for a given classification task. Furthermore, Paul et al. (2015) recently proposed extending analysis of feature importance derived from RF-training by evaluating the significance of each feature towards the classification accuracy. In summary, their method works by using out-of-bag samples to measure the impact of each feature on the classification accuracy after RF-training. Paul et al. (2015) showed that using only a subset of features that are significant led to substantial improvement over using the original feature set. This method is thus particularly suitable for the present study

as it enables us to examine both importance ranking and the significance of each feature to guide our understanding of which features are most representative of each tone. We next outline the procedure described in Paul et al. (2015) to compute the importance measure of each feature that effectively determines the statistical significance of each feature.

3.3.1. Statistical significance of feature

After training the RF classifier, each feature dimension was randomly permuted across the out-of-bag samples. Changes in the distribution of the class votes obtained by permuting a particular feature were then measured via a contingency table that summarizes the classification and misclassification rates (i.e. True Positive, True Negative, False Positive and False Negative) when the feature in question is permuted (or not). This procedure is repeated multiple times. A set of p-values were then obtained by running Pearson's χ^2 test of independence on these measures. After corrections for multiple comparisons, features with adjusted p-values that are below the standard confidence level ($p < 0.05$) are henceforth regarded as "significant".

Note that using the dataset with class labels of 4 tones as OVA classification (2 classes) in the same way as for a multi-class problem naturally leads to imbalanced classes. To address this, we employed bootstrapped sampling (Liu et al., 2009) so that the number of random samples r drawn for each class is the same. This step was repeated N times to eliminate bias towards any class.

In our experiments, we set $r = 500$ and set N empirically to 300 (we did not find any difference in classification performance when $N > 100$). We also employed $t = 500$ random trees for each tone classification task. Lastly, 90% of the samples in the entire dataset was used for training and 10% was used for testing in each of the sampling iterations.

4. Results

We present the results of the analyses that employed an RF classifier to distinguish each tone from the other tones using extracted features. These analyses help us understand which extracted features are most representative for each tone.

4.1. Measurements in physical units and the corresponding normalized data in pixels

The feature analysis in this study was performed on pixel data of the captured video. In order to relate the extracted measurements from pixels to physical units (i.e. mm), we thus measured the physical head sizes of two randomly chosen speakers (1 male, 1 female) to approximate the extracted measurements in physical units. A summary of the obtained feature measurements in both physical units and in pixels is listed in Table A.1. From this table, we show for example that the average head displacement during head lowering for the male speaker is 0.604 mm (1.817 pixels) while that of the female speaker is 1.328 mm (3.650 pixels), and that the maximum eyebrow velocity during eyebrow raising for the male is 0.290 mm/s (0.872 pixels/s) while that for the female is 0.516 mm/s (1.418 pixels/s).

4.2. Tone classification accuracy using the extracted features

For classification of Tone 1, average accuracy achieved by a trained random forest classifier as computed over 1000 iterations was 0.6317 ± 0.1021 when all features were used on the test set. When only using a subset of the extracted features selected using Paul's method (Paul et al., 2015), accuracy improved to 0.6497 ± 0.0958. Similarly, average classification accuracy improved from 0.5789 ± 0.0927 to 0.6063 ± 0.0929 for Tone 2 classification, from 0.6822 ± 0.1376 to 0.6892 ± 0.1302 for Tone 3 classification, and from 0.5987 ± 0.0784 to 0.6311 ± 0.0831 for Tone 4 classification.

Accuracy rates of each tone classification task were computed under 3 settings: (1) using all the features; (2) using K features that are significant as determined by Paul's approach (i.e., by applying a threshold on J_{χ^2}); (3) using the top K features as ranked by Breiman's importance measure (J_a). Using approach 2 (Paul's approach; only the features that were determined to be significant) yielded better performance for all tones as compared to using approach 1 (using all features). Also, selecting the K significant features generally led to better performance over selecting the top K features based on Breiman's importance weighting, thereby supporting the use of the analysis approach of Paul et al. (2015).

4.3. Most relevant features per tone

By computing and applying a threshold on J_{χ^2} (Paul et al., 2015), we found that 10 features were significant for Tone 1 (see Fig. 4(b)), 5 features were significant for Tone 2 (see Fig. 5(b)), 15 features were significant for Tone 3 (see Fig. 6(b)), and 8 features were significant for Tone 4 (see Fig. 8(b)). The more features are selected, the better the accuracy, ranging from 5 features (61%) to 15 features (69%).

Feature patterns observed for each tone when they were analyzed in the context of separate tone classification problems (see Section 3.3) were obtained. We first examine the importance ranking of the significant features for each tone classification task and then examine the significant features individually.

4.3.1. Tone 1

For Tone 1, we plotted the importance weight[2] of the 10 significant features for Tone 1 classification (Fig. 4(b)). From this figure, one can see that most of the significant features were those that describe head and eyebrow movements. These, in the order of importance weight, include: (i) maximum head-raising velocity; (ii) relative time when eyebrow-lowering distance is maximum; (iii) relative time when lips opening velocity is maximum; (iv) maximum lips-opening distance; (v) maximum head-lowering distance; (vi) relative time when head-lowering distance is maximum; (vii) relative time when eyebrow-lowering velocity is maximum; (viii) relative time when eyebrow-raising velocity is maximum; (ix) maximum eyebrow-lowering velocity; and (x) relative time when head-raising velocity is maximum.

We next examined the discriminatory power of each of these 10 significant features by examining their mean values pooled over each tone class. More specifically, for each tone, we compared its mean feature value with the mean of those computed from all other tones. Further, we tested the significance of observed differences in the mean values using a student's t-test. Prior to the t-test, we confirmed that the data was normally distributed per Lilliefors test. Results of the comparisons are given in Fig. 4(c). Whenever the means between one tone and the other three tones differed significantly ($p < 0.05$), we placed an asterisk above each bar.

Tone 1 is generally produced with minimal head and eyebrow movement compared to the other tones. Specifically, the maximum head-raising velocity, head-lowering distance, and eyebrow-lowering velocity exhibited by Tone 1 was smallest in value, reflecting that articulation of this tone required either smaller movements or slower velocities. Examining the plots in Fig. 4(c), one trend observed is that the times taken by the head and eyebrow keypoints to reach maximum velocity, during head raising and during eyebrow raising, respectively, were the longest for Tone 1, suggesting that the height of motion happened quite late for this tone when compared to the other tones. Overall, we could see that these significant features that best discriminate Tone 1 from the other tones are related to the velocities of keypoints and the times (of an event), and that Tone 1 generally involved lower mean values (i.e. smaller movements) for these features.

[2] Note that Paul's importance value J_{χ^2} and Breiman's importance value (J_a) are highly correlated and functionally similar as shown in Paul et al. (2015). For simplicity, we thus chose and plotted J_a as it is more familiar to the computational/statistics community than J_{χ^2}, which is also inversely proportional to common intuition of importance (i.e. lower probability value implies greater importance).

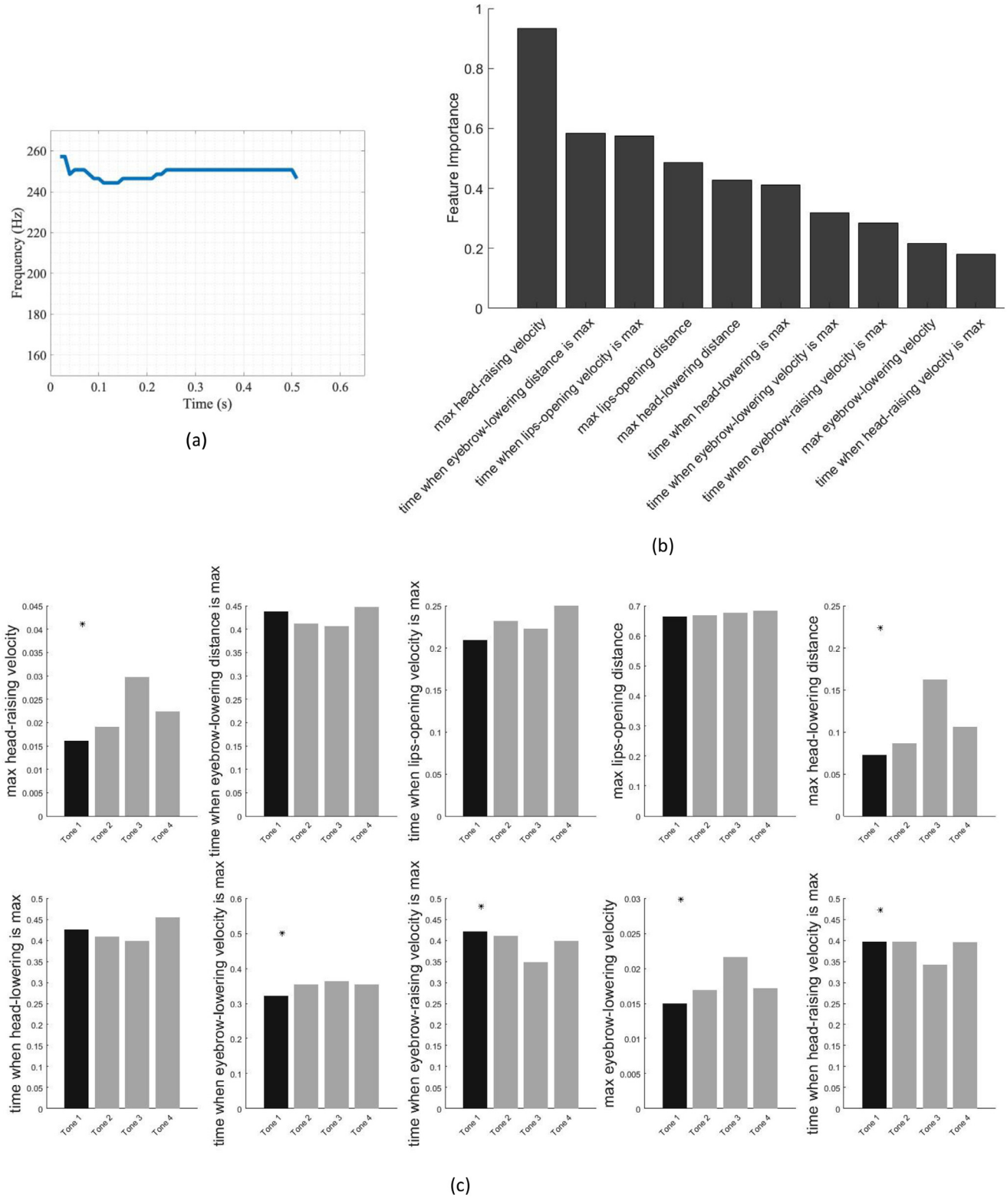


Fig. 4. (a) An example fundamental frequency contour for Tone 1. (b) Importance value of all the features determined to be significant for Tone 1 classification. (c) Comparisons of the group means of the 10 features that were significant for Tone 1 classification. 6 features were found to exhibit statistically significant differences between Tone 1 and all the others combined. A ∗ above a bar represents a p-value smaller than 0.05.

4.3.2. Tone 2

For Tone 2, only 5 of the 33 features extracted were significant. Their feature importance rankings are given in Fig. 5(b), including (i) maximum eyebrow lowering distance; (ii) maximum eyebrow raising distance; (iii) maximum head lowering distance; (iv) relative time when head raising distance was maximum; and (v) relative time when eyebrow raising velocity was maximum. None of the significant features were derived from the lip regions.

Feature comparisons of Tone 2 against the other tones were conducted using t-tests. As shown in Fig. 5(c), Tone 2 exhibits longest time

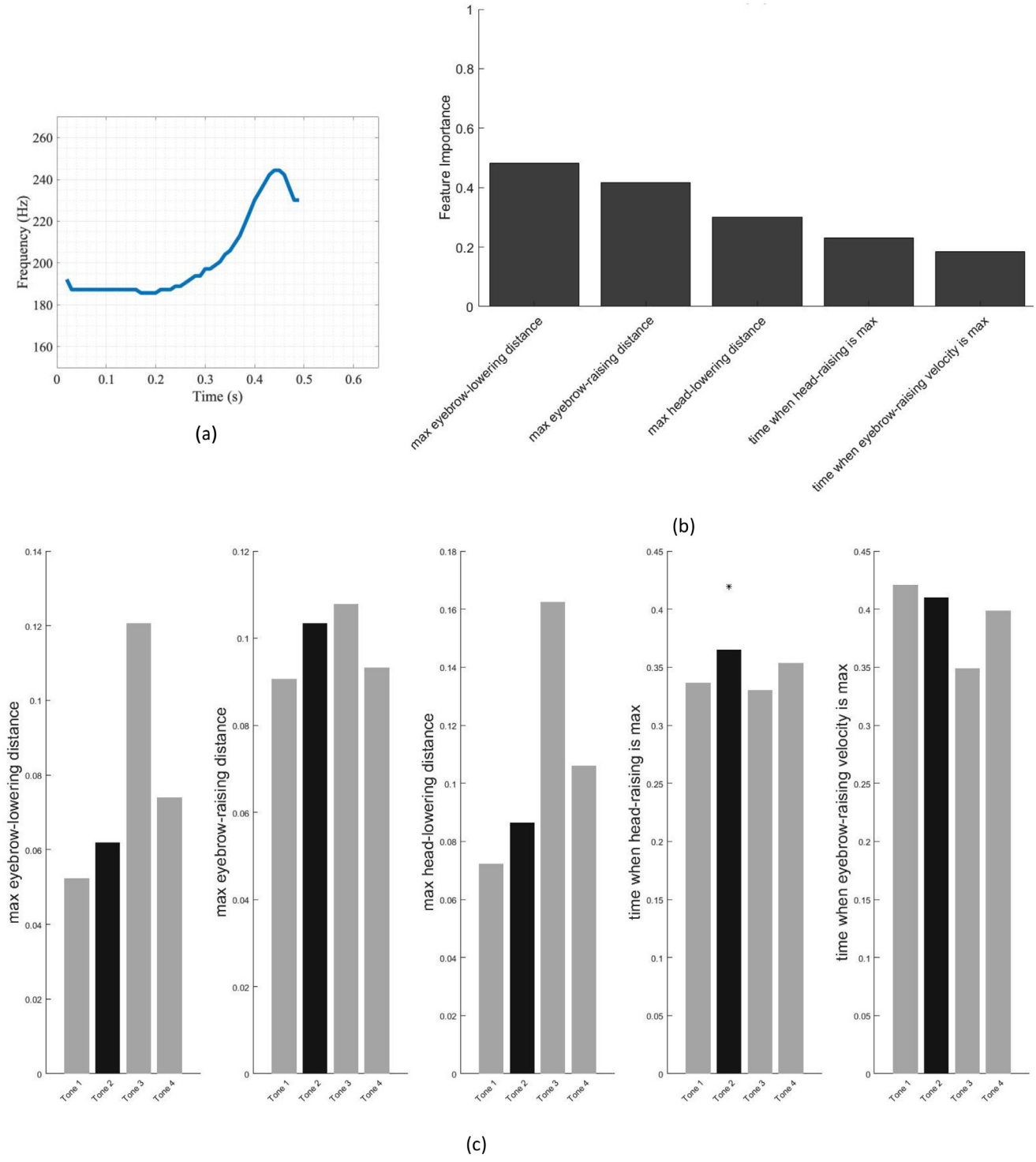


Fig. 5. (a) An example fundamental frequency contour for Tone 2. (b) Importance value of all the features determined to be significant for Tone 2 classification. (c) Comparisons of the group means of the 5 features that were significant for Tone 2 classification. One feature was found to exhibit statistically significant differences between Tone 2 and all the others combined. A $*$ above a bar represents a p-value smaller than 0.05.

to reach maximum head raising distance as compared to the other tones. This suggests that the feature maxima happened at the later part of the tone. Additionally, we observed larger eyebrow raising in the later part of the motion, albeit a t-test did not reveal statistical significance. The larger rising motion in the later part corresponds to the rising contour of Tone 2.

4.3.3. Tone 3

We next turn to Tone 3, which is the most dynamic of all four Mandarin tones. Compared to the other tones, many more features were found to be significant for the Tone 3 classification. Fig. 6(b) ranks the significant features by their measured importance as done similarly before for Tones 1 and 2. These include: (i) maximum eyebrow-

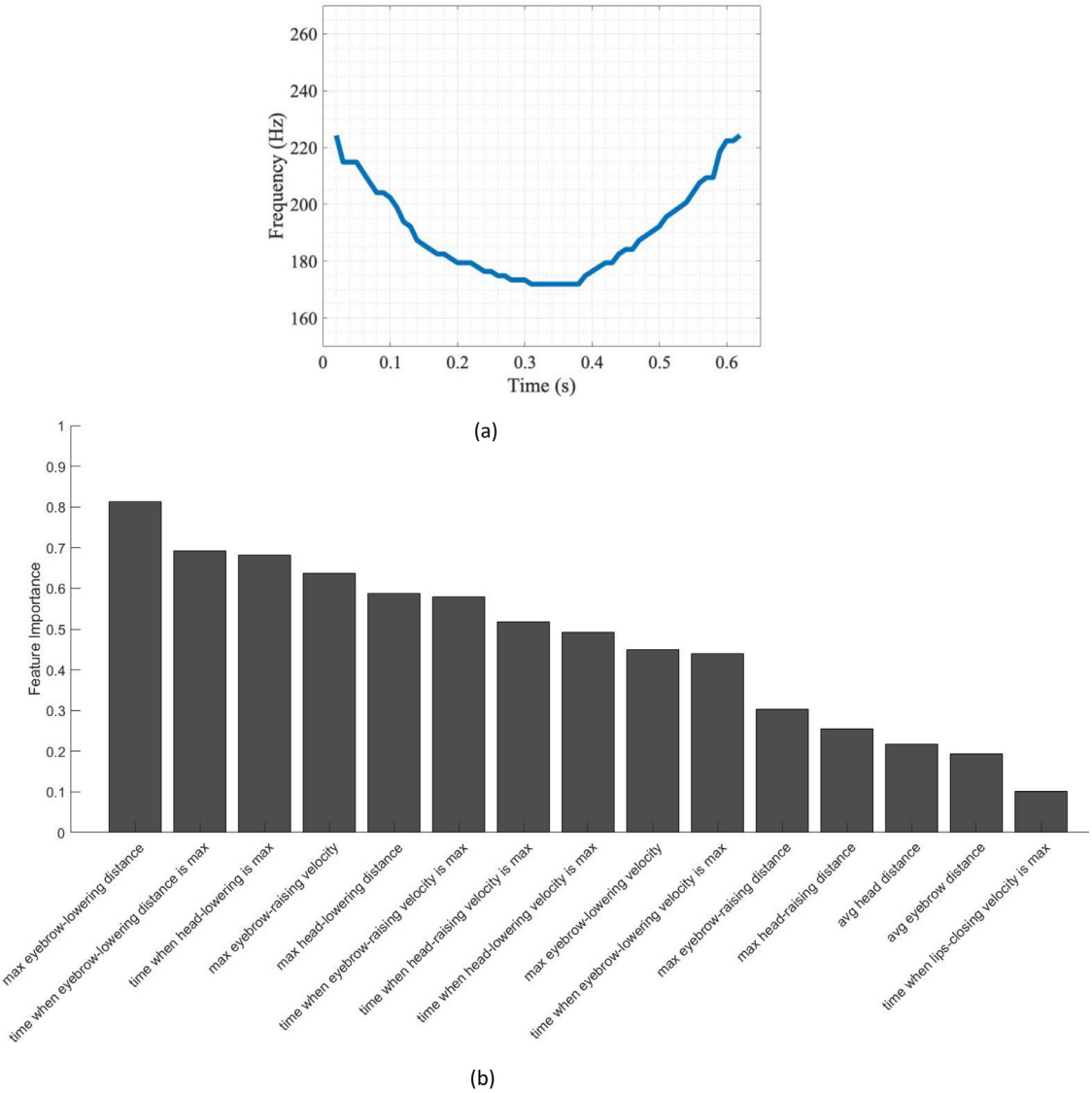


Fig. 6. (a) An example fundamental frequency contour for Tone 3. (b) Importance value of all the features determined to be significant for Tone 3 classification.

lowering distance; (ii) relative time when eyebrow-lowering distance is maximum; (iii) relative time when head-lowering is maximum; (iv) maximum eyebrow-raising velocity; (v) maximum head-lowering distance; (vi) relative time when eyebrow-raising velocity is maximum; (vii) relative time when head-raising velocity is maximum; (viii) relative time when head-lowering velocity is maximum; (ix) maximum eyebrow-lowering velocity; (x) relative time when eyebrow-lowering velocity is maximum; (xi) maximum eyebrow-raising distance; (xii) maximum head-raising distance; (xiii) average head distance; (xiv) average eyebrow distance; and (xv) relative time when lips-closing velocity is maximum. From this figure, one could generally see that features extracted from the eyebrow and the head had high feature importance. In contrast, there was only 1 feature extracted from the lips that had high feature importance.

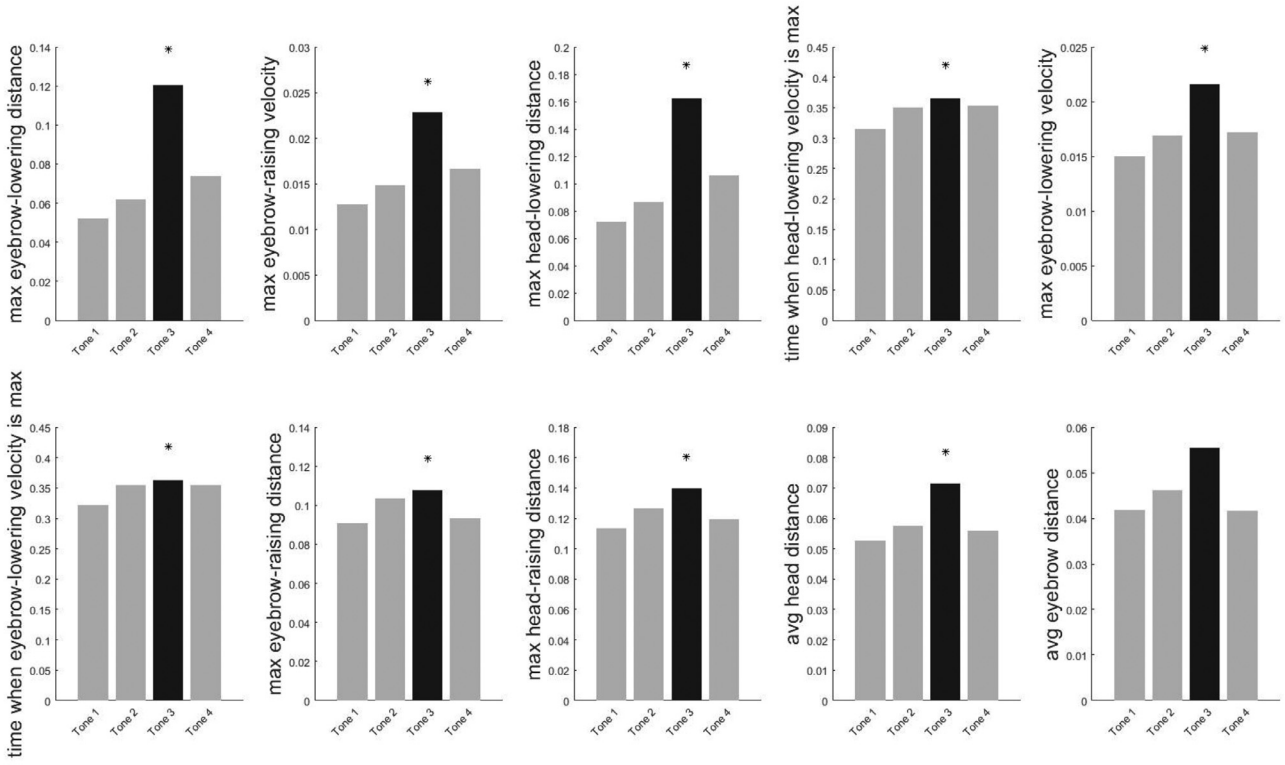
Fig. 7(a) plots the mean values of the top ten significant features for the Tone 3 classification task where the feature value in Tone 3 was largest when compared to the values of the other tones for the same features, with an asterisk showing a significant difference from *t*-test at $p < 0.05$. In summary, Tone 3 has the greatest amount of movement for head-raising, head-lowering, eyebrow-raising and eyebrow lowering as the distance traveled by the corresponding keypoint was the largest. The average distance traveled by head and eyebrow were also the largest. Additionally, the velocity for eyebrow-raising and eyebrow-lowering was the largest for this tone. The times taken for the head-lowering velocity and the eyebrow-lowering velocity to reach maximum values were also the largest for this tone when compared to all other tones.

Fig. 7(b) plots the mean values of the five significant features where the value in Tone 3 was the smallest when compared to the values of the other tones. We can see that the time taken for the lips-closing velocity to reach maximum value was also shortest for this tone. Additionally, the times taken for the head-raising and eyebrow-raising velocity to reach maximum were also shortest for Tone 3. These last two observations suggest that the head and eyebrow movements co-vary.

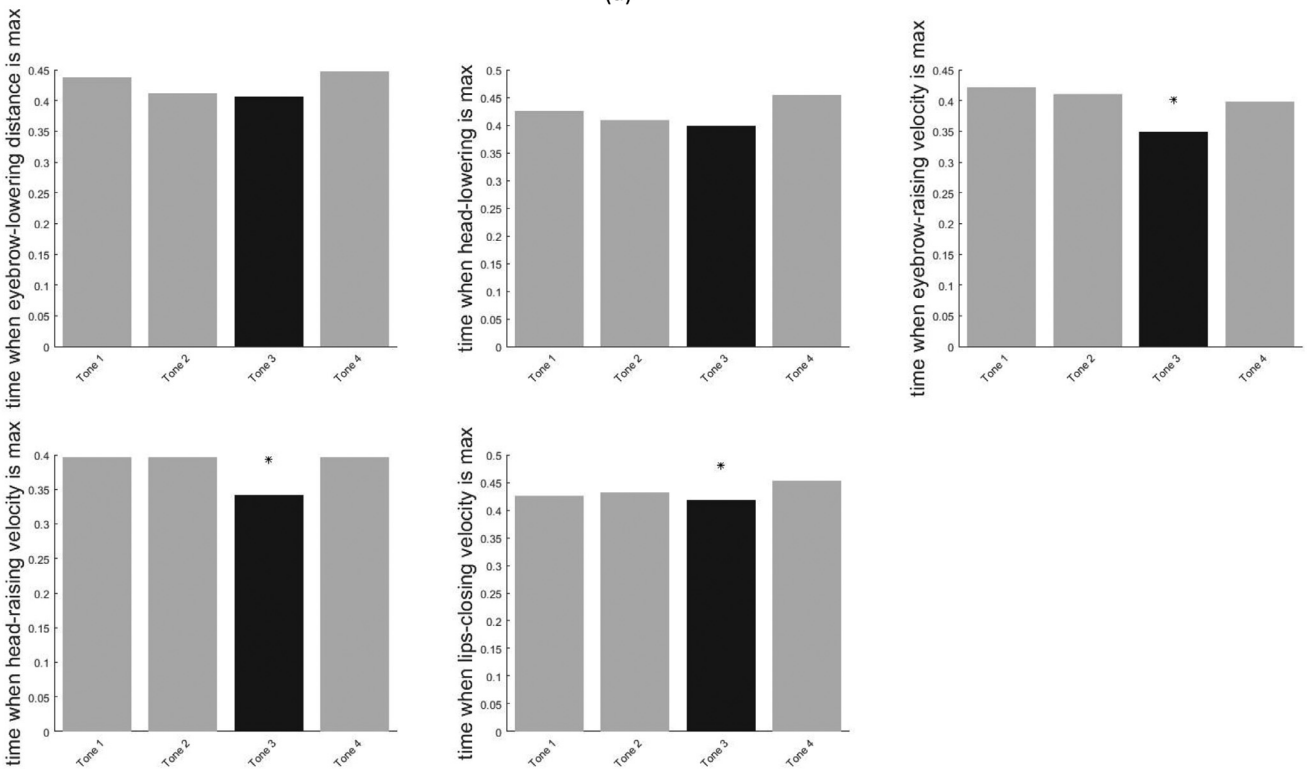
Collectively, the shorter mean times of maximum head-lowering distance and eyebrow-lowering distance suggest that the aforementioned events occurred fairly early in the tone production while the velocity peaks of downward movements of keypoints occurred later, close to the end of tone production. These patterns align with the dipping nature of Tone 3.

4.3.4. Tone 4

Lastly, we repeated the above analyses for Tone 4. Fig. 8(b) ranks the other significant features relative to this feature. These are: (i) relative time when head-lowering is maximum; (ii) relative time when



(a)



(b)

Fig. 7. (a) Comparisons of the group means of the 10 features whose feature value was larger in Tone 3 as compared to other tones and deemed significant for Tone 3 classification. Note that 9 features were found to exhibit statistically significant differences between Tone 3 and all the others combined. (b) Comparisons of the group means of the 5 features whose values were smallest for Tone 3 and deemed significant for Tone 3 classification. Note that 3 features were found to exhibit statistically significant differences between Tone 3 and all the others combined. A $*$ above a bar represents a p-value smaller than 0.05.

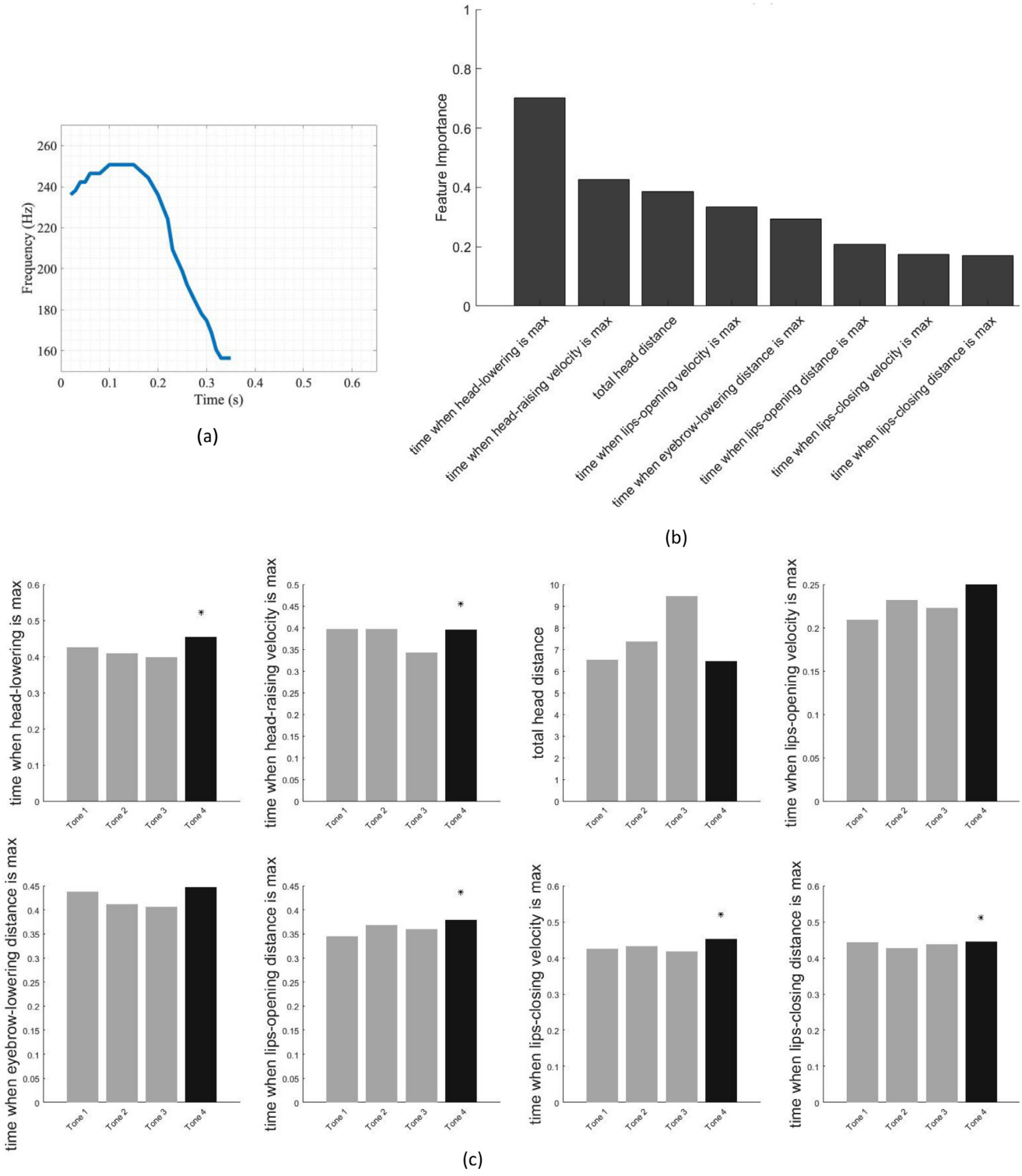


Fig. 8. (a) An example fundamental frequency contour for Tone 4. (b) Importance value of all the features determined to be significant for Tone 4 classification. (c) Comparisons of the group means of the 8 features that were significant for Tone 4 classification. 5 features were found to exhibit statistically significant differences between Tone 4 and all the others combined. A $*$ above a bar represents a p-value smaller than 0.05.

head-raising velocity is maximum; (iii) total head distance; (iv) relative time when lips-opening velocity is maximum; (v) relative time when eyebrow-lowering distance is maximum; (vi) relative time when lips-opening distance is maximum; (vii) relative time when lips-closing velocity is maximum; and (viii) relative time when lips-closing distance is maximum. The most important significant feature for this tone was the relative time when head lowering was largest. Overall, most of the features that were found to be significant were time-based.

Fig. 8(c) compares the mean values of each significant feature for Tone 4 relative to those for the other tones, based on the t-test results. Several observations can be seen. Firstly, the relative time for the velocity of lip-closing to reach maximum value was largest for Tone 4.

Secondly, the relative time required by the head keypoint to reach maximum lowering was largest for Tone 4. This suggests that the lowering movement occurred in the later part of the vowel pronunciation. Thirdly, the relative time it took for the head-raising keypoints to reach maximum velocity had the second smallest value (Tone 3 had the smallest value). Three of these five significant features relate to those from the lips regions, while the remaining two are related to the head. Hence, this is the only other tone that seems to be characterized well by lip-related features beside Tone 1. Overall, the relative times it took for critical events to happen during lowering movements were the largest for Tone 4.

5. Discussion

5.1. Technical contribution

This study used a synergistic pipeline of LBP-based cascade classifiers and optical-flow algorithms (Beauchemin et al. 1995) to carefully track facial features such as head, eyebrow, and lips. In terms of motion tracking and feature extraction, our image analysis algorithms freed us from the use of physical markers that had various limitations as explained in Section 1.3, albeit it was a commonly adopted solution in several linguistic studies e.g., Kim et al., 2014a; Attina et al., 2010; Yehia et al., 2002. In terms of accuracy and precision, smooth tracking of motion trajectories allowed us to robustly measure distance, time, and kinematic features from each of the keypoints. This contrasts with the use of a generic face classifier that was employed in our previous work (Tang et al., 2015). Analyses of these additional feature dimensions allowed us to extend our findings from existing literature pertaining to motion- and velocity- related tone characteristics.

In terms of analyses, tones were compared using a 2-step discriminant analysis approach based on RF and statistical analysis proposed by Paul et al. (2015) to evaluate each feature importance in discriminating each tone from all other tones in this work. Our 2-step approach aided us to systematically identify 4 subsets of features, each of which is critical in characterizing a tone, thereby allowing us to relate our results to previous findings (e.g., Attina et al., 2010; Munhall et al., 2004; Scarborough et al., 2009; Yehia et al., 2002) and draw conclusions from new findings for each tone with confidence built on machine learning algorithms and statistical basis. Our results show overall alignment of head and eyebrow movements with the tone contour.

In summary, our extended approach adopted in the current study facilitates identification of visual articulatory features characterizing each tone, as discussed below.

5.2. Features characterizing each tone

5.2.1. Tone 1

Firstly, as presented in Section 4.3, we found that Tone 1 has the smallest amount of head-lowering displacement as well as the lowest velocities in head-raising and eyebrow-lowering as compared to other tones. These relatively small movements observed for Tone 1 may be explained by its small tone variation in F0, which would be consistent with two sets of prior works that showed that (1) head movements are larger in prosodic constituents with a larger amount of variance in F0 (Munhall et al., 2004; Yehia et al., 2002); and (2) larger eyebrow displacements are associated with focused words that involve greater F0 variation (Kim et al., 2014a).

Additionally, our result on the kinematic differences between Tone 1 and the other tones is an observation that has not been reported in previous work. For instance, we observed that the time taken by the eyebrow to reach maximum velocity during eyebrow raising was the longest for Tone 1 when compared to all other tones. This echoes the lack of fluctuations in the contour of this tone due to its level nature. Furthermore, our analysis also suggests that small head and eyebrow movements maybe coordinated and co-vary.

5.2.2. Tone 2

Tone 2 exhibits the longest time to reach maximum head raising distance relative to the other tones. Furthermore, we observed a trend of larger eyebrow raising in the later part of the motion. These movements consistently correspond to the rising F0 contour for this tone. Similar observations on F0 and eyebrow were also reported for other pitch-related prosodic variations (Cavé et al., 1996; Dohen et al., 2006 and Scarborough et al., 2009). For example, as discussed previously, Cavé et al. (1996) and Cavé et al. (2002) reported a significant link between eyebrow movements and F0 and also showed that the rising-falling F0 curve was similar in shape to the eyebrow movement curve; Dohen et al. (2006) found a relationship between eyebrow motion (rising) and the production of prosodic focus for three out of the five French speakers; and Scarborough et al. (2009) reported that their speakers raised an eyebrow on almost all stressed words.

Our result of the link between Tone 2 and kinematic features such as the time when eyebrow-raising velocity was maximum also finds support in previous studies on prosody. As Scarborough et al. (2009) had reported that displacement-based and velocity-based measures were highly correlated, the present finding that Tone 2's rising velocity happened near the end of the pronunciation was expected, as it is a rising tone.

5.2.3. Tone 3

Total distance travelled by a keypoint during the utterance of a tone is an indirect measure of the duration of the utterance. Hence, the larger the total distance, the longer is the duration of the utterance. For Tone 3, total head distance was largest, suggesting that it is a longer tone which is consistent with what is reported in the literature (e.g., Attina et al., 2010).

Further, Tone 3 is known to be the most dynamic tone in terms of F0 variation. Munhall et al. (2004) and Yehia et al. (2002) observed that the head movements are larger in prosodic constituents with a larger amount of variance in F0. These observations are consistent with our findings of larger movements seen in the head and eyebrow for Tone 3 as compared to the other tones.

However, Attina et al. (2010) found lip-closing and lip-raising to be significantly negatively correlated with the F0 contour of Tone 3, while in our analysis none of the features related to lips showed up as significant.

Another set of new observations made for Tone 3 is that we found certain dynamic movements of head and eyebrow such as eyebrow-raising and eyebrow-lowering velocity to be largest for Tone 3 as it is the most dynamic tone with larger variations. Also, the time when head- and eyebrow- lowering velocity is maximum is longest for Tone 3 and the time when head- and eyebrow- raising velocity is maximum is shortest for Tone 3. This suggests that the raising velocity was larger in the beginning of the pronunciation and the lowering velocity was larger at the end of the pronunciation compared to the other tones.

5.2.4. Tone 4

For Tone 4, features related to lip-movement seem to play an important role in differentiating this tone from the other tones. Previously, Attina et al. (2010) have reported that lip-closing is positively correlated with F0 in Tone 4. Our analysis also shows two lip-closing features significant in differentiating Tone 4 from the other three tones, i.e., times when distance and velocity were largest during lip-closing. Further, Huron et al. (2013) reported that the average F0 correlated positively with eyebrow position, with higher vocal pitch associated with higher eyebrow placement. Munhall et al. (2004) reported a similar correlation between head movement and F0 contour. These patterns are consistent with our findings: it took longer for head-lowering distance to reach maximum, suggesting head-lowering happened later in the tone. Since this is a falling tone, the head-lowering followed the F0 contour.

In addition, our analyses found several new features that are characteristic of Tone 4: the time when head-raising velocity was largest was longest for this tone. This could reflect the head quickly coming to a resting position after the head lowering movement of the tone. Moreover, the time when lip-opening distance was largest was also longest, suggesting the movements happened later in the tone.

5.2.5. Summary of characteristic features

In summary, the current results show three general trends. Firstly, the spatial and temporal dynamics of the head, eyebrow and lip features generally followed the pitch contour. For example, head lowering occurred close to the end of the articulation of Tone 4, which agrees with the falling pitch contour of Tone 4. Likewise, Tone 2 exhibits a long time to reach maximum eyebrow raising, aligning with the rising nature of this tone. Secondly, head and eyebrow features covary in their movement trajectories. For instance, temporal and kinematic features of head and eyebrow movements covary for Tone 3. Lastly, in terms of discriminatory ability of the features, we observed that both displacement-based and kinematic (especially velocity-based) features were powerful measures in guiding the discrimination of each tone.

5.3. General discussion

Findings of the current study provide new insights into defining the nature of visual tonal cues and determining their linguistic relevance. As discussed earlier, unlike phonemes, the production of lexical tones is driven by laryngeal activities independent of vocal tract configurations (Howie 1976; Lehiste 1970; Yip 2002). It is thus an open question whether facial and mouth movements in tone production are articulatorily required cues and how they are used to signal tonal category distinctions.

The current results show that some of the visual tonal cues may indeed be articulatorily motivated, arguably due to movements of the laryngeal muscles that control the vocal folds when pitch is varied (Burnham et al., 2015; Yehia et al., 2002). For instance, the significantly large head-lowering and head raising movements for Tone 3 (associated with the dipping nature of this tone) may find support from the previous claim that head lowering or raising can be triggered by the reduced or tightened vocal folds, resulting in low- or high-pitched tones (Moisik et al., 2014; Smith et al., 2012). However, such a link to laryngeal activities cannot account for the current results of eyebrow and lip movements associated with the production of specific tones.

A more cogent explanation lies in the visuospatial-acoustic link between articulatory movements and pitch trajectories, showing that head, eyebrow, and lip movements in terms of spatial and temporal changes in distance, direction, velocity, and timing are aligned with acoustic features of tonal changes in height, contour, and duration. This cross-modal association between spatial and pitch changes in tone is in line with the previous claim that pitch is audio-spatial in representation (Connell et al., 2013; Hannah et al., 2017), in that head and facial movements accompanied by tone productions provide spatial equivalence to pitch trajectories. For example, an elevation in space (e.g., hand or eyebrow raising) is generally found to correlate with an elevation in pitch (Connell et al., 2013; Huron et al., 2013; Küssner et al., 2014).

While such cross-modal spatial-acoustic binding has been claimed to exist universally in the human sensory-motor system (Barsalou 2008; Borghi et al., 2013; Hannah et al., 2017), the current results further suggest linguistically meaningful associations of the above-mentioned visual cues with tone articulation. This is evidenced by tone-specific alignments between head, eyebrow, and lip movements and the pitch trajectories characterizing different tones. In particular, the downward or upward head and eyebrow movements follow the dipping (Tone 3) and rising (Tone 2) pitch trajectories, the timing and dynamicity of lip closing movements are associated with the falling pitch trajectory of Tone 4, and minimal movement is characteristic of Tone 1, the level tone with a lack of pitch change. These patterns demonstrate how specific visual cues are used in characterizing individual tones, making each tone distinct from the other contrasting tones. Thus, findings from the current study indicate linguistically salient, category-defining tonal articulatory cues, suggesting language-specific mechanisms underlying cross-modal tone production, above and beyond a general language-universal system.

6. Concluding remarks

The current finding of the linguistic relevance of visual tonal cues has significant implications for tone perception. It has been claimed that patterns of visual cues can be used to make specific predictions for perception. For example, Scarborough et al. (2009) predicted more accurate perception for stress contrasts with more distinctive visual features and better perception for more prominent visual features. Extending these predictions to the present results, Tone 3, which contains a greater number of distinctive visual features and more prominent visual features than Tone 2, should be more accurately perceived than Tone 2. Given that little research has mapped specific visual tonal cues to tone perception (Burnham et al., 2007; Chen et al., 2011; Mixdorff et al., 2005), further evidence from tone perception is needed to determine how the visual tonal cues are used to facilitate perception of categorical tonal distinctions, and the extent to which perception is based on the linguistic relevance of these cues.

Understanding the visual correlates of tone production and perception not only advances research on cross-modal integration of sensory-motor information in speech processing, but also has important applications for the development of effective tools for tone language learning, visual aids for hearing-impaired conditions, audio-visual tonal speech synthesis for virtual communication, as well as speech simulation tools that will amplify the identified tone-specific visual articulatory cues to aid speech perception in noisy environments.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was funded by a research grants from the Social Sciences and Humanities Research Council of Canada (SSHRC Insight Grant 435-2012-1641) and Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant 2017-05978). We thank Lisa Tang, Keith Leung, Jane Jian, Charles Turo, and Dahai Zhang from the SFU Language and Brain Lab for their assistance, as well as WestGrid and Compute Canada for their IT support. Portions of this project have been presented at the 176th Meeting of the Acoustical Society of America and 2018 Acoustics Week in Canada, November 2018, Victoria, BC.

Appendix A. EXTRACTED FEATURES

Table A.1.

Table A.1
Feature measurements obtained in mm and in pixels (px).

Index	Feature	Male mean	Std	Mean	Std	Female mean	Std	Mean	Std
1	Maximum displacement of the head while head-raising from its starting position	1.58 mm	1.04	4.75 px	3.13	5.88 mm	3.46	16.14 px	9.43
2	Maximum displacement of the head while head-lowering from its starting position	−0.60 mm	0.63	− 1.82 px	1.88	−1.33 mm	1.79	−3.65 px	4.90
3	Average distance head moved during the utterance	0.54 mm	0.39	1.63 px	1.16	1.71 mm	1.09	4.68 px	2.95
4	Total distance traveled by head during the utterance	59.70 mm	42.27	179.57 px	125.12	213.67 mm	140.37	586.16 px	381.47
5	Maximum displacement of the eyebrow keypoint from its starting position	1.24 mm	0.82	3.72 px	2.45	4.10 mm	2.46	11.24 px	6.71
6	Maximum displacement of the eyebrow while eyebrow-lowering from its starting position	−0.52 mm	0.51	−1.57 px	1.53	−1.09 mm	1.38	−3.00 px	3.79
7	Average distance eyebrow moved during utterance	0.41 mm	0.30	1.23 px	0.88	1.23 mm	0.78	3.38 px	2.13
8	Total distance eyebrow moved during the utterance	45.175 mm	32.05	135.73 px	94.21	154.8 mm 102	.56	424.59 px	278.76
9	Maximum lip-opening distance	29.08 mm	2.35	62.55 px	3.89	31.39 mm	3.12	62.87 px	6.66
10	Maximum lip-closing distance	20.76 mm	1.25	87.61 px	7.21	22.81 mm	2.32	86.52 px	8.83
11	Average distance lips moved during utterance	22.85 mm	1.32	68.83 px	4.08	24.94 mm	2.29	68.76 px	6.69
12	Total distance lip moved during the utterance	2520.6 mm	210.07	7593.90 px	653.33	3115.3 mm	469.43	8587.1 px	1315.60
13	The relative time at which the displacement of the head while head-raising was maximum	0.37	0.15	0.37	0.15	0.28	0.11	0.28	0.11
14	The relative time at which the displacement of the head while head-lowering was maximum	0.45	0.31	0.45	0.31	0.35	0.29	0.35	0.29
15	The relative time at which the head velocity was maximum during head-raising	0.40	0.14	0.40	0.14	0.33	0.11	0.33	0.11
16	The relative time at which the head velocity was maximum during head-lowering	0.33	0.15	0.33	0.15	0.24	0.13	0.24	0.13
17	The relative time at which the displacement of the eyebrow while eyebrow-raising was maximum	0.37	0.17	0.37	0.17	0.29	0.13	0.29	0.13
18	The relative time at which the displacement of the eyebrow while eyebrow-lowering was maximum	0.47	0.30	0.47	0.30	0.33	0.27	0.33	0.27
19	The relative time at which the eyebrow key-point reached maximum velocity during eyebrow-raising	0.43	0.19	0.43	0.19	0.33	0.10	0.33	0.10
20	The relative time at which the eyebrow velocity during eyebrow-lowering was maximum	0.36	0.16	0.36	0.16	0.25	0.13	0.25	0.13
21	The relative time when the amount of lipopening reached maximum	0.34	0.12	0.34	0.12	0.26	0.12	0.26	0.12
22	The relative time when the amount of lipclosing reached maximum	0.47	0.30	0.47	0.30	0.42	0.33	0.42	0.33
23	The relative time at which the lip velocity during lip-opening was maximum	0.26	0.10	0.26	0.10	0.19	0.09	0.19	0.09
24	The relative time at which the lip velocity during lip-closing was maximum	0.45	0.12	0.45	0.12	0.37	0.12	0.37	0.12
25	Maximum head velocity during head-raising	0.32 mm/s	0.15	0.98 px/s	0.46	0.67 mm/s	0.40	1.83 px/s	1.08
26	Maximum head velocity during headlowering	−0.36 mm/s	0.16	−1.07 px/s	0.48	−0.83 mm/s	0.46	−2.28 px/s	1.24
27	Maximum absolute acceleration of the head	0.23mm^2/s	0.10	0.71 px^2/s	0.29	0.28mm^2/s	0.11	0.76 px^2/s	0.30
28	Maximum eyebrow velocity during eyebrowraising	0.29 mm/s	0.14	0.87 px/s	0.42	0.52 mm/s	0.30	1.42 px/s	0.81
29	Maximum eyebrow velocity during eyebrowlowering	−0.28 mm/s	0.14	−0.84 px/s	0.42	−0.61 mm/s	0.33	−1.68 px/s	0.90
30	Maximum absolute acceleration of the eyebrow	0.19 mm^2/s	0.09	0.58 px^2/s	0.26	0.23mm^2/s	0.08	0.63 px^2/s	0.22
31	Maximum lip velocity during lip opening	3.12 mm/s	1.07	9.39 px/s	3.20	1.67 mm/s	0.59	4.61 px/s	1.64
32	Maximum lip velocity during lip closing	−0.87 mm/s	0.30	−2.62 px/s	0.90	−0.91 mm/s	0.30	−2.49 px/s	0.81
33	Maximum absolute acceleration of the lips	2.14 mm^2/s	0.84	6.44 px^2/s	2.53	0.84 mm^2/s	0.48	2.33 px^2/s	1.33

References

Attina, V., Gibert, G., Vatikiotis-Bateson, E., Burnham, D., 2010. Production of Mandarin lexical tones: auditory and visual components. Auditory-Visual Speech Processing 4–2.

Barsalou, L.W., 2008. Grounded cognition. Annu. Rev. Psychol. 59 (1), 617–645.

Beauchemin, S., Barron, J., 1995. The computation of optical flow. ACM Comput. Surv. 27 (3), 433–466.

Blicher, D.L., Diehl, R.L., Cohen, L.B., 1990. Effects of syllable duration on the perception of the Mandarin tone 2/tone 3 distinction: evidence of auditory enhancement. J. Phon. 18 (S1), 37–49.

Borghi, A., Scorolli, C., Caligiore, D., Baldassarre, G., Tummolini, L., 2013. The embodied mind extended: using words as social tools. Front. Psychol. 4, 214.

Burnham, D., Brooker, R., Reid, A., 2015. The effects of absolute pitch ability and musical training on lexical tone perception. Psychol. Music. 43 (6), 881–897.

Burnham, D., Ciocca, V., Stokes, S., 2001. Auditory-visual perception of lexical tone. In: EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, pp. 395–398.

Burnham, D., Reynolds, J., Vignali, G., Bollwerk, S., Jones, C., 2007. Rigid vs non-rigid face and head motion in phone and tone perception. In: INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, pp. 698–701.

Cai, Y., Huang, Y., Zhang, S., 2012. A method for nose tip location and head pose estimation in 3d face data. In: IET Conference Publications, pp. 115–118.

Castrillón, M., Déniz, O., Guerra, C., Hernández, M., 2007. Encara2: real-time detection of multiple faces at different resolutions in video streams. J. Vis. Commun. Image. Represent. 18 (2), 130–140.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., Espesser, R., 1996. About the relationship between eyebrow movements and f0 variations. In: Proceeding of Fourth International Conference on Spoken Language Processing, 4, pp. 2175–2178.

Cavé, C., Guaïtella, I., Santi, S., 2002. Eyebrow movements and voice variations in dialogue situations: an experimental investigation. 7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH.

Chan, T.F., Vese, L.A., 2001. Active contours without edges. IEEE Trans. Image Process. 10 (2), 266–277.

Chao, Y.R., 1948. *Mandarin Primer*. Harvard University Press.

Chen, C.-C., Lu, P.-T., Hsia, M.-L., Ke, J.-Y., T.-C. Chen, O., 2011. Gender-to-age hierarchical recognition for speech. In: 2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS), pp. 1–4.

Chen, T.H., Massaro, D.W., 2008. Seeing pitch: visual information for lexical tones of Mandarin-Chinese. J. Acoust. Soc. Am. 123 (4), 2356–2366.

Connell, L., Cai, Z.G., Holler, J., 2013. Do you see what I'm singing? isuospatial movement biases pitch perception. Brain. Cogn. 81 (1), 124–130.

Cvejic, E., Kim, J., Davis, C., 2010. Prosody off the top of the head: prosodic contrasts can be discriminated by head motion. Speech Commun. 52 (6), 555–564.

Dohen, M., Loevenbruck, H., 2005. Audiovisual production and perception of contrastive focus in french: a multispeaker study. Interspeech/Eurospeech 2005 2413–2416.

Dohen, M., Loevenbruck, H., Hill, H., 2006. Visual correlates of prosodic contrastive focus in french: description and inter-speaker variability. Proc. Speech Prosody 1, 221–224.

Dreher, J.J., Lee, P.-.C., 1968. Instrumental investigation of single and paired Mandarin tonemes. Monumenta Serica 27 (1), 343–373.

Flecha-García, M.L., 2010. Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. Speech. Commun. 52 (6), 542–554.

Gandour, J.T., 1983. Tone perception in far eastern languages. J. Phon. 11 (2), 149–175.

Garg, S., Hamarneh, G., Jongman, A., Sereno, J.A., Wang, Y., 2018. Joint gender-, tone-, vowel- classification via novel hierarchical classification for annotation of monosyllabic Mandarin word tokens. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5744–5748.

Hannah, B., Wang, Y., Jongman, A., Sereno, J.A., Cao, J., Nie, Y., 2017. Cross-modal association between auditory and visuospatial information in Mandarin tone perception in noise by native and non-native perceivers. Front. Psychol. 8, 2051.

Howie, J.M., 1976. *Acoustical studies of Mandarin vowels and tones*. Princeton-Cambridge Studies in Chinese linguistics. Cambridge University Press.

Huron, D., Shanahan, D., 2013. Eyebrow movements and vocal pitch height: evidence consistent with an ethological signal. J. Acoust. Soc. Am. 133 (5), 2947–2952.

Ishi, C.T., Ishiguro, H., Hagita, N., 2007. Analysis of head motions and speech in spoken dialogue. In: INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, 2, pp. 670–673.

Kim, J., Cvejic, E., Davis, C., 2014a. Tracking eyebrows and head gestures associated with spoken prosody. Speech. Commun. 57, 317–330.

Kim, J., Davis, C., 2014b. Comparing the consistency and distinctiveness of speech produced in quiet and in noise. Comput. Speech Lang. 28 (2), 598–606.

Krishnan, A., Swaminathan, J., Gandour, J.T., 2009. Experience-dependent enhancement of linguistic pitch representation in the brainstem is not specific to a speech context. J. Cogn. Neurosci. 21 (, 6), 1092–1105.

Küssner, M.B., Tidhar, D., Prior, H.M., Leech-Wilkinson, D., 2014. Musicians are more consistent: gestural cross-modal mappings of pitch, loudness and tempo in real-time. Front. Psychol. 5, 789.

Lehiste, I., 1970. *Suprasegmentals*. MIT Press.

Lienhart, R., Kuranov, A., Pisarevsky, V., 2003. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. Pattern Recognit. 297–304.

Lin, M.-.T., 1965. The pitch indicator and the pitch characteristics of tones in standard Chinese. Acta Acustica 2, 8–15.

Liu, X.-.Y., Wu, J., Zhou, Z.-.H., 2009. Exploratory undersampling for class-imbalance learning. IEEE Trans. Syst. Man Cybern. Part B 39 (2), 539–550.

Lucas, B.D., Kanade, T., 1981. An iterative image registration technique with an application to stereo vision. In: *7th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'81, pp. 674–679.

Mixdorff, H., Hu, Y., Burnham, D., 2005. Visual cues in Mandarin tone perception. In: Ninth European Conference on Speech Communication and Technology, pp. 405–408.

Moisik, S.R., Lin, H., Esling, J.H., 2014. A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (sllus). J. Int. Phon. Assoc. 44 (1), 21–58.

Moore, C.B., Jongman, A., 1997. Speaker normalization in the perception of Mandarin Chinese tones. J. Acoust. Soc. Am. 102 (3), 1864–1877.

Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., Vatikiotis-Bateson, E., 2004. Visual prosody and speech intelligibility: head movement improves auditory speech perception. Psychol. Sci. 15 (2), 133–137.

Noroozi, F., Sapinski, T., Kami´nska, D., Anbarjafari, G., 2017. Vocal-based emotion recognition using random forests and decision tree. Int. J. Speech Technol. 20 (2), 239–246.

Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24 (7), 971–987.

Paul, J., Dupont, P., 2015. Inferring statistically significant features from random forests. Neurocomputing 150, 471–480.

Perkell, J.S., Zandipour, M., 2002. Economy of effort in different speaking conditions. II. kinematic performance spaces for cyclical and speech movements. J. Acoust. Soc. Am. 112 (4), 1642–1651.

Prom-on, S., Liu, F., Xu, Y., 2012. Post-low bouncing in Mandarin Chinese: acoustic analysis and computational modeling. J. Acoust. Soc. Am. 132 (1), 421–432.

Prom-on, S., Xu, Y., Thipakorn, B., 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. J. Acoust. Soc. Am. 125 (1), 405–424.

Reid, A., Burnham, D., Kasisopa, B., Reilly, R., Attina, V., Rattanasone, N.X., Best, C.T., 2015. Perceptual assimilation of lexical tone: the roles of language experience and visual information. Atten., Percept., & Psychophys. 77 (2), 571–591.

Scarborough, R., Keating, P., Mattys, S.L., Cho, T., Alwan, A., 2009. Optical phonetics and visual perception of lexical and phrasal stress in English. Lang. Speech 52 (2–3), 135–175.

Shaw, J.A., Chen, W.-r., Proctor, M.I., Derrick, D., Dakhoul, E., 2014. On the interdependence of tonal and vocalic production goals in Chinese. In: 10th International Seminar on Speech Production, pp. 395–398.

Smith, D., Burnham, D., 2012. Facilitation of Mandarin tone perception by visual speech in clear and degraded audio: implications for cochlear implants. J. Acoust. Soc. Am. 131 (2), 1480–1489.

Srinivasan, R.J., Massaro, D.W., 2003. Perceiving prosody from the face and voice: distinguishing statements from echoic questions in English. Lang. Speech. 46 (1), 1–22.

Swerts, M., Krahmer, E., 2010. Visual prosody of newsreaders: effects of information structure, emotional content and intended audience on facial expressions. J. Phon. 38 (2), 197–206.

Tang, L.Y., Hannah, B., Jongman, A., Sereno, J., Wang, Y., Hamarneh, G., 2015. Examining visible articulatory features in clear and plain speech. Speech Commun. 75, 1–13.

Tomasi, C., Kanade, T., 1991. Detection and tracking of point features. Int. J. Comput. Vis. 9, 137–154.

Traunmüller, H., Öhrström, N., 2007. Audiovisual perception of openness and lip rounding in front vowels. J. Phon. 35 (2), 244–258.

Tu, J., Fu, Y., Huang, T.S., 2009. Locating nose-tips and estimating head poses in images by tensorposes. IEEE Trans. Circuits Syst. Video Technol. 19 (1), 90–102.

Wang, Y., Jongman, A., 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.

Xu, Y., Gandour, J.T., Francis, A.L., 2006. Effects of language experience and stimulus complexity on the categorical perception of pitch direction. J. Acoust. Soc. Am. 120, 1063–1074.

Yehia, H.C., Kuratate, T., Vatikiotis-Bateson, E., 2002. Linking facial animation, head motion and speech acoustics. J. Phon. 30 (3), 555–568.

Yip, M., 2002. *Tone*. Cambridge University Press.