CONTROL ID: 3008913

TITLE: Computer-vision analysis shows different facial movements for the production of different Mandarin tones

AUTHORS (FIRST NAME, LAST NAME): Saurabh Garg[1], Lisa Tang[5], Ghassan Hamarneh[3], Allard Jongman[2], Joan A. Sereno[2], Yue Wang[4]
INSTITUTIONS (ALL): 1. Pacific Parkinson's Research Centre , University of British Columbia, Vancouver, BC, Canada.
2. Department of Linguistics, University of Kansas, Kansas City, KS, United States.
3. School of Computer Science, Simon Fraser University, Burnaby, BC, Canada.
4. Department of Linguistics, Simon Fraser University, Burnaby, BC, Canada.
5. School of Computer Science, Simon Fraser University, Buranby, BC, Canada.

ABSTRACT BODY:
Abstract (200 words): We aim to identify visual cues resulting from facial movements made during Mandarin tone production and examine how they are associated with each of the four tones. We use signal processing and computer vision techniques to analyze audio-video recordings of 21 native Mandarin speakers uttering the vowel /ɜ/ with each tone. Four facial interest points were automatically detected and tracked in the video frames: medial point of left-eyebrow, nose tip (proxy for head movement), and midpoints of the upper and lower lips. Spatiotemporal features were extracted from the positional profiles of each tracked point. These features included distance, velocity, and acceleration of local facial movements with respect to the resting face of each speaker. Analysis of variance and feature importance analysis based on random decision 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. Preliminary results suggest alignments between articulatory movements and pitch trajectories, with downward or upward head and eyebrow movements following the dipping and rising tone trajectories, faster lip-closing toward the end of falling tone production, and minimal movements for the level tone.

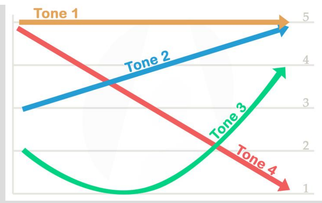
Computer-vision analysis shows different facial movements for the production of different Mandarin tones

Saurabh Garg[a], Lisa Tang[b], Ghassan Hamarneh[b], Allard Jongman[c], Joan Sereno[c], and 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

Overview

Four lexical tones in Mandarin, differing in F0 height & contour as high level (Tone1), mid-high rising (Tone2), low dipping (Tone3), & high falling (Tone4).



- Research including our own has shown that lip and jaw movements provide visual cues and facilitate segmental speech perception [1,2].
- Visual cues shown to be relevant for prosody, including tone, involve head, neck, and eyebrow movements [3,4,5,6], since prosodic production does not rely on vocal tract configurations.
- Pitch has been claimed to be visual spatial in nature [7].
- Research has not been conclusive about which specific movements are used to characterize each tone.

Aim

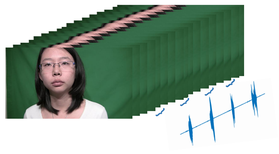
- We aim to identify specific visual cues as induced by facial movements made during Mandarin tone production using state-of-the-art computer-vision and image processing techniques.

We hypothesize that the 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.

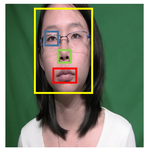
Materials

- 20 native speakers of Mandarin (8 males) who were born and raised in Northern China or Taiwan at least during the first 18 years of their lives.
- From each speaker, 150 pronunciations of the /3/ tone quadruplet words were recorded.

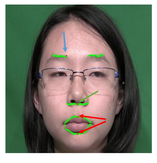
Methods



- Tone utterances were segmented first using automatic tools based on audio amplitude.



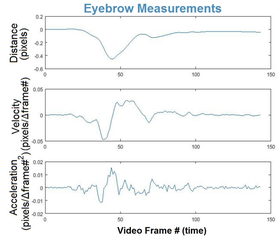
- A rough bounding box was localized using the cascade filter approach [8] and part-specific detectors were used to obtain better localizations of the region of interest.

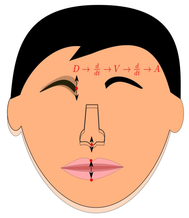


- Four facial points of interest were detected automatically: medial point of left-eyebrow, nose tip (proxy for head movement), and midpoints of the upper and lower lips.

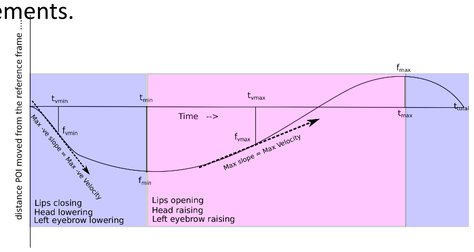
Methods (cont'd)

- Keypoints that were identified on the first frame of each video token were tracked on the rest of the video frames using the Kanade-Lucas-Tomasi method [9].





- A set of 33 features was computed to quantify the motion dynamics of each of the four tracked keypoints and to provide summary statistics of the local (eyebrows and lips) and rigid (head) movements.



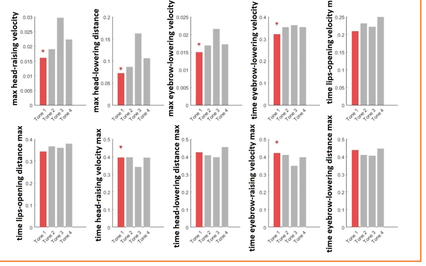
- Critical features such as the distance, velocity, time, and acceleration describing local facial movements with respect to the resting face of each speaker were extracted from the positional profiles of each tracked point.

Results

- 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.

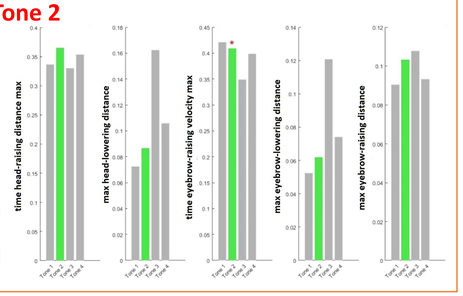
Tone 1

- When compared to the other tones max head velocity for Tone 1 was smallest: reflecting articulation of this tone required minimal head movements
- The times taken by the lip and eyebrow keypoints to reach max velocity were longest for Tone 1, suggesting that the height of motion happened quite late for this tone.
- Overall, Tone 1 generally involved lower mean values (i.e. smaller movements) for these features as compared to other tones.



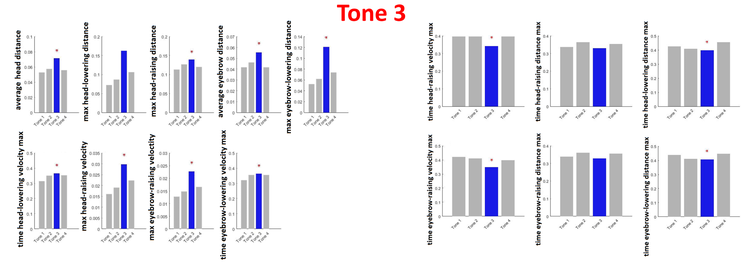
Tone 2

- Max head lowering and eyebrow lowering distance was second smallest, while the time taken to reach the maximum head raising distance was largest and max eyebrow raising velocity was second largest.
  - This suggests that the feature maxima happened at the later part of the tone pronunciation.
- The minimal movement in the beginning and the larger motion in the later part correspond to the F0 contour (lowering and then raising) feature of Tone 2.



Results (cont'd)

Tone 3

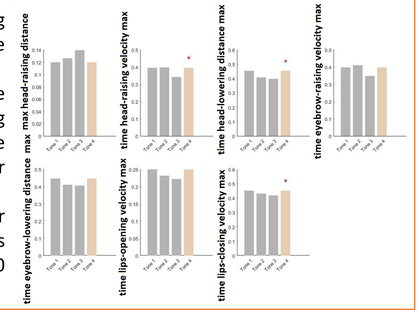


Features with large value in Tone 3 — Features with small value in Tone 3

- Tone 3 has the greatest amount of movement for head-raising, head-lowering, and eyebrow-lowering as the distance travelled by the corresponding keypoint was the largest.
- The times taken for the head and eyebrow-raising distance and velocity to reach max values were shortest for this tone when compared to all other tones, while the times taken for the head velocity and the eyebrow-lowering velocity to reach maximum values were the largest for this tone.
- In summary, Tone 3 exhibits a larger amount of movements, involving slower lowering and faster raising head and eyebrow movements, corresponding to the dipping contour of this tone.

Tone 4

- The times for the velocity of lip-opening and lip-closing to reach maximum value was largest for Tone 4.
- The time required by the head and the eyebrow keypoints to reach max lowering was largest for Tone 4, suggesting that the lowering movement occurred in the later part of the tone production.
- Overall for Tone 4, the time required for critical events during lowering movements was largest, matching with the dipping F0 contour of this tone.



Conclusions

Results suggest alignments between articulatory movements and pitch trajectories, with downward or upward head and eyebrow movements following the dipping and rising tone trajectories, lip closing movement being associated with the falling tone, and minimal movements for the level tone.

References & Acknowledgements

1. Jeesun Kim, Erin Cvejic, and Chris Davis. "Tracking eyebrows and head gestures associated with spoken prosody". In: Speech Communication 57 (2014), pp. 317–330.
2. Lisa Y.W. Tang, Beverly Hannah, Allard Jongman, Joan Sereno, Yue Wang, and Ghassan Hamarneh. "Examining visible articulatory features in clear and plain speech". In: Speech Communication 75 (2015), pp. 1–13.
3. Denis Burnham, Valter Ciocca , and Stephanie Stokes. "Auditory–visual perception of lexical tone", In: EUROSPEECH, (2001) pp. 395–398.
4. Trevor H. Chen and Dominic W. Massaro. "Seeing pitch: Visual information for lexical tones of Mandarin-Chinese". In: The Journal of the Acoustical Society of America 123.4 (2008), pp. 2356–2366.
5. Kevin G. Munhall, Jeffery A. Jones, Daniel E. Callan, Takaaki Kuratate, and Eric Vatikiotis-Bateson. "Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception". In: Psychological Science 15.2 (2004). PMID: 14738521, pp. 133–137.
6. Marc Swerts and Emiel Krahmer. "Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions". In:Journal of Phonetics 38.2 (2010), pp. 197–206.
7. Louise Connell, Zhenguang G. Cai, and Judith Holler. "Do you see what I'm singing? Visuospatial movement biases pitch perception". In: Brain and Cognition 81.1 (2013), pp. 124–130.
8. Rainer Lienhart, Alexander Kuranov, and Vadim Pisarevsky. "Empirical analysis of detection cascades of boosted classifiers for rapid object detection". In: Pattern Recognition (2003), pp. 297–304.
9. Carlo Tomasi and Takeo Kanade. Detection and Tracking of Point Features. Tech. rep. International Journal of Computer Vision, (1991).

Acknowledgments: This study was funded by research grants from SSHRC and NSERC. We also thank Keith Leung, Jane Jian, Charles Turo, and Dahai Zhang from SFU Language and Brain Lab for their assistance, as well as WestGrid and Compute Canada for their IT support.