Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Gestures time to vowel onset and change the acoustics of the word in Mandarin

Patrick Louis Rohrer [1], *Yitian Hong* [2], *Hans Rutger Bosker* [1,3]

[1]Donders Center for Cognition, Radboud University, Nijmegen, The Netherlands
[2]The Hong Kong Polytechnic University, Hong Kong SAR, China
[3]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
patrick.rohrer@donders.ru.nl, ytian.hong@connect.polyu.hk,
hansrutger.bosker@donders.ru.nl

Abstract

Recent research on multimodal language production has revealed that prominence in speech and gesture go hand-in-hand. Specifically, peaks in gesture (i.e., the apex) seem to closely coordinate with peaks in fundamental frequency (F0). The nature of this relationship may also be bi-directional, as it has also been shown that the production of gesture directly affects speech acoustics. However, most studies on the topic have largely focused on stress-based languages, where fundamental frequency has a prominence-lending function. Less work has been carried out on lexical tone languages such as Mandarin, where F0 is lexically distinctive.

In this study, four native Mandarin speakers were asked to produce single monosyllabic CV words, taken from minimal lexical tone triplets (e.g., /pi1/, /pi2/, /pi3/), either with or without a beat gesture. Our analyses of the timing of the gestures showed that the gesture apex most stably occurred near vowel onset, with consonantal duration being the strongest predictor of apex placement. Acoustic analyses revealed that words produced with gesture showed raised F0 contours, greater intensity, and shorter durations. These findings further our understanding of gesture-speech alignment in typologically diverse languages, and add to the discussion about multimodal prominence.

Index Terms: multimodality, temporal alignment, lexical tone, gesture production

1. Introduction

The multimodal nature of face-to-face communication is receiving increasing attention from researchers in the language sciences [1], surfacing pertinently in the fact that gesture and speech are temporally coordinated, both regarding semantics ([2], [3], [4]) as well as prosodic prominence [5]. Specifically, prominent parts of gesture roughly align with prominent syllables in speech. A number of lab-based studies have focused on the apex, or the "point of maximum extension" of a gesture stroke [6], generally finding a close temporal relationship with the peak F0 of the pitch accented syllable. For example, [7] asked a native speaker of English to perform a reading task and to produce beat gestures on specific target words. They found that the timing of the apex of the beat gesture was least variable with regard to the peak of the F0 within the pitch accented syllable (for similar results, see [8]). This close correlation holds even in contexts where pitch production is constrained by phrasal prosodic structure [9].

Earlier studies on the relationship between gesture and speech suggested that this alignment may be done purposely by the speaker to highlight new or important content in speech. Specifically, beat gestures have been described as "highlighters" that function collaboratively with speech prosody in order to lend prominence to certain syllables or words in speech [5]. This notion of gesturing to boost prominence is seen not only in timing, but also in the effects gesture production has on speech. For example, when asking participants to read entire Dutch sentences, [10] showed how the production of a beat gesture increased the duration and boosted the first and second formants (F1 and F2) of the corresponding syllable, which also act as cues to prominence in speech. More recently, however, another plausible explanation for this close relationship may lie in theories of biomechanics, which have shown how the forces of moving one's arm have a direct impact on the muscles surrounding the rib cage, which in turn impact phonation intensity and F0 [11], [12]. More specifically, when asking participants to steadily phonate a single vowel while moving their arm up-and-down, they found that the acoustic signal (i.e., F0 and amplitude envelope) entrained to the movement patterns. However, moving towards speech, such biomechanical effects for F0 seem less clear [13].

It is important to note that most of the previous research on the temporal integration of speech and gesture has focused on languages where F0 (along with increased intensity and duration) generally functions to denote prominence conveyed via pitch accentuation. Much less is known about how gestures are temporarily integrated in tonal languages. One study by [14] has investigated the temporal integration of pointing gestures in Cantonese using a picture-naming task. They asked 10 native Cantonese speakers to point to mono- and di-syllabic lexical items under neutral and corrective focus conditions. They found that speakers generally produced pointing gestures so as to coincide with words in focus (monosyllabic words: 76.65%; disyllabic words: 88.15%), however in disyllabic words, speakers' pointing gestures tended to co-occur with the first syllable, regardless of whether the prosodic focus (cued by durational lengthening) was on the first or second syllable. The authors thus conclude that the first syllable of the lexical word carrying prosodic focus may be a key anchoring point for gesture in Cantonese, though there was some individual variation in this pattern. The study did not offer further analysis of the data with regard to specific timing of the apexes within the syllable or effects of gesture on speech acoustics.

Another study by [15] investigated the relationship between gesture and speech in a corpus of conversational speech in Medʉmba, a Bantu language from Cameroon. They found that gesture apexes occurred very close to vowel onsets (mean distance of 3 ms ± 79 ms). Importantly, tone was a significant predictor of the timing of apexes relative to vowel onset, where L-tone vowels had apexes significantly later relative to vowel

10.21437/SpeechProsody.2024-175

onset than H-tone or falling tone vowels. Additionally, the authors found a significant effect of vowel duration, where longer vowels were produced with apexes occurring later relative to vowel onset. Interestingly, the authors also compared syllables that co-occurred with gesture to those that did not. They found that greater vowel intensity and longer duration were also significant predictors of whether a syllable was produced with a gesture or not. The authors also mention a marginally significant effect of tone (p = 0.08) where gestures tended to co-occur with syllables that had a *lower* mean F0.

In order to better understand multimodal speech production in Mandarin, we collected data from 4 speakers uttering 60 CV monosyllabic words (e.g., /pi/) sampled from 20 minimal tone triplets (henceforth, "items"), with Tone 1 (a high, flat tone, e.g., /pi1/ meaning 'to threaten'), Tone 2 (a rising tone, e.g., /pi2/ meaning 'nose') and Tone 3 (a dip followed by a rise, e.g. /pi3/ meaning 'pen'). Participants produced these 60 words in two conditions, namely with and without a corresponding beat gesture. An exploratory analysis of this data will assess two crucial aspects of multimodal language production: the timing patterns of gesture-speech alignment and the effects of gesture production on speech acoustics in a tonal language.

We are particularly interested in the timing relationship between gesture and speech, as most previous studies have identified the F0 peak within a pitch accented syllable as a prosodic anchor for gesture production in pitch accent languages. However, prior evidence on tonal languages seem to suggest that F0 has a rather minimal role in gesture timing, and other landmarks (e.g., onset of the word or specifically the vowel) or acoustic factors (e.g., vowel duration) may play a crucial role in the timing of gesture and speech in those languages. Investigating the effect gesture production has on speech in a tonal language may also be particularly insightful with regards to phonological and biomechanical accounts of speech production. A phonological account would suggest that gesture production in a tonal language should reinforce cues to prominence in speech, namely for Mandarin via increasing syllabic duration and pitch height/range in lexical tones (e.g., [16], [17], [18]). This also presented as interesting testing ground for biomechanical accounts for intensity and (particularly) F0 modulation, as effects of muscle tension due to movement deceleration should be relatively short-lived compared to the duration of a vowel, producing an F0 peak within the syllable at moments near the gestural apex. However, this would present a problem in Mandarin, as there is a general need to maintain vocal integrity for lexical tone production.

2. Methods

2.1. Data collection

Four native Mandarin speakers (all female speakers, born and raised in Northern China and did not report acquiring any other Chinese dialects) were recruited at Radboud University (mean age of 29 ± 3.74). Speakers were recorded in the DCC lab with a Canon XF405 camera and a Shure 16A overhead microphone. The video was recorded in 1920x1080 at 50 frames per second in MP4 format. The speakers were asked to read from a list of 60 words that contrasted across three lexical tones: Tone 1 (T1), Tone 2 (T2), and Tone 3 (T3). Tone 4 was not included as the materials were originally elicited for a perception study [19], testing T1-T2 and T2-T3 continua. The 60 words were randomly ordered into a PowerPoint (PPT) presentation. We presented one word per slide, with the PPT automatically advancing at 4-second intervals. Speakers were initially instructed to read the words on the slides. They were given no further instructions, as this constituted the baseline (i.e., No Gesture) condition. After going through all of the slides two times, the speakers were then asked to read the list two more times, and were specifically asked to produce a "forceful downward beat gesture" while uttering each word (i.e., the Gesture condition), so as to maximize any potential biomechanical effects from gesture production. Gestures were always produced with the right hand in a fist handshape starting by the speaker's side, then moving up to shoulder level before coming down forcefully and stopping in front of them. This resulted in a total of 960 one-word utterances produced across speakers and conditions for analysis (i.e., 4 speakers x 20 items x 3 tones x 2 conditions x 2 recordings). All participants of the experiments described here gave informed consent as approved by the Ethics Committee of the Social Sciences department of Radboud University (ECSW-LT-2023-3-24-97246).

2.2. Data processing

Following the recording, speech and gesture annotation were carried out independently of one another. These independent annotations were then merged in ELAN [20] and exported for further descriptive and statistical analysis in R [21].

2.2.1. Gesture annotation

Gestures were annotated manually in ELAN following the M3D guidelines for gesture phase and apex annotation [22]. For each gesture, three to four gesture phases were annotated, namely the preparation (movement of the hand from rest to the most upward position before changing direction), the stroke (the downward movement of the hand), and the recovery (the return of the hand to the rest position). Sometimes, the speakers produced a post-stroke hold (momentary pauses after the stroke, where movement is minimal) which were annotated when present. To facilitate comparability with earlier studies, the point of interest for the current analysis is the apex, which was annotated as the point of maximum downward extension identified in frame-by-frame analysis.

2.2.2. Speech annotation and acoustic extraction

Speech was annotated in Praat [23] where word boundaries were identified and transcribed in pinyin. Additionally, each consonant and vowel was identified and annotated, and the consonants were further classified based on consonant type (nasals, aspirated vs. unaspirated stops, aspirated vs. unaspirated affricates, fricatives, and epenthetic vs. non-epenthetic glides). The F0 contour for each word was then extracted by sampling the F0 value at 10 ms intervals throughout the entire duration of the vowel. This was then time-normalized by calculating average F0 in consecutive bins of 10% of the total vowel duration.

2.2.3. Statistical analysis

In order to assess the temporal alignment of gesture with speech, an initial visual inspection of the data was carried out. This visual inspection allowed us to determine the relevant landmarks in speech to which gesture may be associating with. Then, a set of linear mixed effect models were run using the *lme4* package [24] to assess which factors best predicted apex placement, as well as to assess the effects of gesture production on speech acoustics in terms of F0, intensity, and vowel duration. The random effects structure for each model was

determined by the *buildmer* package [25], which takes as input the most complex random effects structure and returns the random effects that best fit the data. All models are specified in the relevant results sections. The data and scripts are available online at https://osf.io/w4czh/.

3. Results

3.1. Patterns of gesture-speech temporal integration

Initial visualization of the data found many apexes to be produced quite early within the word. Specifically, the apexes were found to occur on average 12 ms after vowel onset (SD = 95 ms), and around 125 ms after word onset (SD = 101 ms). Figure 1 (left panel) shows the distribution of the distance of the apex from two landmarks, vowel onset and word onset. An initial inspection of the data suggested that the duration of the consonant may play a role in this timing relationship. The upper right panel of Figure 1 shows the distance of the apex from word onset as a function of consonantal duration, while the lower right panel shows the distance from the vowel onset as a function of consonantal duration. Taken together, it appears as though vowel onset seems to be a more stable landmark in speech for apex production: as the duration of the consonant increases, the apex is produced further away after word onset, and relatively closer to vowel onset. However, some individual variability did surface in the data, as illustrated in Figure 2, where each speaker's timing significantly differed from zero (see OSF for these analyses).

Given the results reported by [15] regarding vowel duration as a significant predictor of gesture-speech temporal alignment in Medumba, a linear mixed-effects model was run to see if tone, consonant duration, and vowel duration would significantly predict the distance of the apex to vowel onset. The fixed factor of Tone was run as a categorical variable (T1, T2, and T3; T1 mapped onto intercept), and with consonant and vowel duration as z-scaled continuous variables. The model included by-speaker, by-item, and by-recording random intercepts. The model showed a significant effect of consonant duration (β = −41.264, s.e. = 4.162, t = −9.915, p < .001) indicating that as the consonant became longer, the apex tended to occur earlier with respect to vowel onset. Similarly, there was also a significant yet slightly more modest main effect of vowel duration (β = 9.982, s.e. = 4.346, t = 2.297, p = .022) which suggests that as the vowel became longer, the apex tended to occur later with respect to vowel onset. Finally, a significant main effect of tone was also found suggesting that compared to T1 words, apexes occurred slightly earlier with respect to vowel onset in T2 words (β = −8.258, s.e. = 4.169, t = −1.981, p = .048), as well as in T3 words (β = −17.916, s.e. = 6.194, t = −2.893, p = .004). Note, however, that these effects of tone and vowel duration were very small, falling within the error margin for frame-by-frame apex annotation (see Section 4).

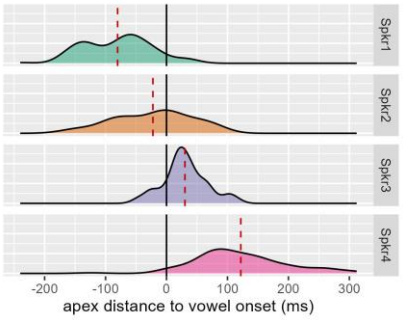

Figure 2: *Distribution of distance of apex to vowel onset by speaker. Solid line represents vowel onset.*

3.2. Effects of gesture production on speech

The second objective of this current study is to assess the effects of gesture on speech production. In order to do so, three linear mixed effects models were run. The first model was run with the time-normalized F0 contour (in 10 % bins) as the dependent variable, fixed factors of Gesture condition (categorical variable; Gesture mapped onto intercept), Tone (categorical variable; T1 mapped onto intercept), and Normalized time (continuous variable), as well as their interaction, and a random effects structure of by-speaker and by-item random intercepts, and by-item random slopes for Tone and Gesture condition. The model returned a significant effect of Tone and an interaction between Tone and Normalized time (as to be expected given the different tonal contours). Importantly, a significant main effect of Gesture condition was found (β = −10.28, s.e. = 2.042, t = −5.036, p < .001), suggesting that producing a gesture tended to raise the F0 contour by approximately 10 Hz. No 2-way interactions were found between Gesture and Tone or Normalized time, suggesting that all tones were equally affected by gesture, across the entire contour (see Figure 3). It should be noted that the same analysis was run with mean F0 as the dependent variable (averaging across time), and though the trend still existed, it was not shown to be significant (see https://osf.io/w4czh/ for further results).

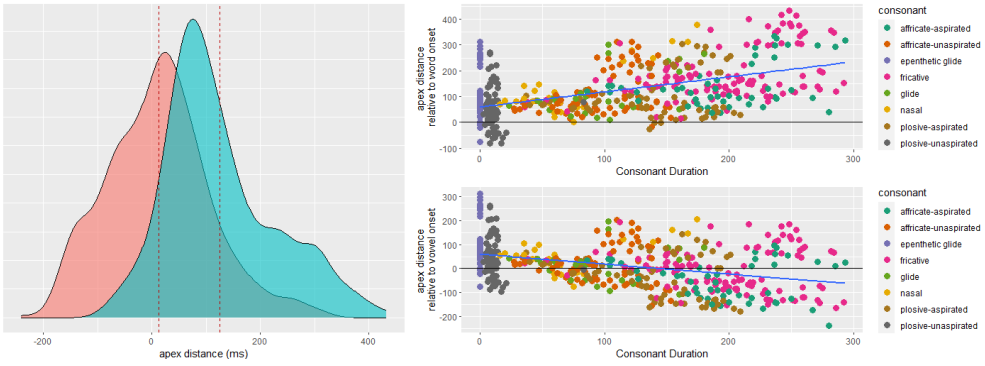

Figure 1 (left panel): *Distribution of the distances of the apex to two different speech landmarks, the vowel onset (red) and word onset (blue). Dotted red lines illustrate the mean.* (right panel): *The distance of the apex to word onset (upper panel) and vowel onset (lower panel) as a function of consonantal duration. Colors represent different consonant types.*

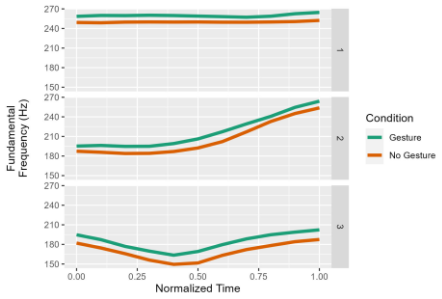

Figure 3: *The effect of gesture production on F0 contour for Tone 1 (upper panel), Tone 2 (middle panel) and Tone 3 (lower panel).*

The second model was run with the intensity measured at vowel midpoint (in dB) as the dependent variable, fixed factors of Gesture condition, Tone, and their interaction, and a random effects structure of by-speaker and by-item random intercepts and by-speaker random slopes for Tone. The model returned a significant effect of Tone indicating that T3 showed significantly lower intensity than T1 ($\beta = -12.169$, s.e. = 3.354, $t = -3.74$, $p = .032$). A significant main effect of Gesture condition was found ($\beta = -2.056$, s.e. = 0.446, $t = -4.612$, $p < .001$), suggesting that words produced with gestures showed greater intensity by approximately 2 dB. The model showed no significant interactions. A model with the same structure was run on the vowel duration data (in ms). This model returned a significant effect of Tone indicating that T2 and T3 showed significantly longer vowels than T1 (T2: $\beta = 33.814$, s.e. = 10.19, $t = 3.318$, $p = .036$; T3: $\beta = 117.301$, s.e. = 33.299, $t = 3.523$, $p = .038$). Also, a significant main effect of Gesture was found ($\beta = 17.957$, s.e. = 3.945, $t = 4.552$, $p = .002$), suggesting that words produced with gestures were significantly *shorter* than words produced without gestures by approximately 18 ms. The model showed no significant interactions.

4. Discussion & Conclusions

The current study is the first to assess gesture-speech alignment in Mandarin, specifically with regards to the temporal alignment of beat gestures with monosyllabic CV words that differ in lexical tones, as well as the effects of these gestures on speech acoustics. In terms of timing, our exploratory analyses suggest gestures stably anchor to vowel onset. This finding corroborates findings by [15] for Medumba, and adds nuance to those reported by [14] for Cantonese by specifying where within the syllable the apex is produced. With regards to the acoustic features that may predict the timing data, our results suggest that consonantal duration is the largest predictive factor, followed by small influences of tone and vowel duration. It is rather surprising that the effect of tone and vowel duration show opposite effects (i.e., T2 and T3 shift the apex occurrence earlier in time, while longer vowels shift the apex occurrence later in time), given the vocalic duration of T2 and T3 are longer, respectively. However, it should be noted that the effect sizes for vowel duration and tone are rather small (e.g., the largest effect being for T3 at approximately −22 ms), the two co-vary, and both are within the margin of error for frame-by-frame apex annotation based on the video framerate (50 frames per second, each frame lasting 20 ms). Thus, more precise measurements of movements (such as those afforded by motion tracking technology) may help clarify the phenomena occurring at such small timescales.

In terms of production effects, we found that producing beat gestures slightly raised the entire F0 contour and boosted intensity at vowel midpoint. Interestingly, however, words that were produced with a gesture showed shorter vocalic durations. Taken together, these results suggest that neither a purely phonological nor a purely biomechanical account alone is sufficient in explaining the effects of gesture production on speech. As prominence in Mandarin is encoded by increased pitch range and duration ([16], [17], [18]), under a phonological account the gesture would be expected to boost prominence by showing a higher pitch excursion in T2 and lowering the dip in T3, as well as lengthening syllabic duration. The current data do not support such a view. Effects in F0 could relate to a biomechanical effect (where muscle tensioning from arm movement increases F0), however, these are typically very short-lived. Perhaps the biomechanics primarily affects F0 near vowel onset, causing speakers to modulate (i.e., raise) their entire F0 contour in an effort to maintain the integrity of the lexical tone. Such a case would indeed suggest an interplay between biomechanics and phonology, deserving further attention in future studies. It is also important to note that when recording, the gesture condition always followed the no-gesture condition. This may in part explain the unexpected shorter vocalic duration in the gesture condition, but could also entail that the F0 and intensity effects are in fact underestimated.

Finally, it is important to highlight the small effects reported here, sampled from only 4 speakers who show a vast amount of variation in their production patterns (cf. Figure 2). The timing results echo the findings described in [14] regarding by-speaker variability in alignment with disyllabic Cantonese words: our data show very clear patterns inherent to individual speakers (i.e., speaker 1 regularly gestured early, while speaker 4 regularly gestured relatively late). So while cross-individual distributions from our data seem to suggest that vowel onset is a stable anchoring point for gesture-speech production in Mandarin, it could be that specific speakers tend to follow their own individual patterns of gesture-speech alignment. In addition to individual differences, it is also important to note that the context in which these gestures were produced was highly constrained (producing single words and being explicitly instructed to do a forceful beat gesture in a lab). Much less is known about the timing relationship between speech and different types of gestures in more spontaneous conversation, particularly for lexical tone languages. All in all, this first exploration of multimodal Mandarin production adds to the growing body of literature suggesting that gesture-speech alignment patterns vary according to different language typologies (i.e., pitch-accent vs. tonal languages) and even among individual speakers. However, more data is necessary in order to fully understand how context and speaker-specific differences may influence multimodal speech production.

5. Acknowledgements

The authors would like to acknowledge funding by an ERC Starting Grant [PI: H.R.B.] (HearingHands, 101040276) from the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. We would also like to thank our Research Assistant Beatrice Caddeo for her help with the annotations, as well as the other members of the SPEAC research group for their helpful feedback on the study.

6. References

[1] P. Perniss, "Why We Should Study Multimodal Language," *Front. Psychol.*, vol. 9, p. 1109, Jun. 2018, doi: 10.3389/fpsyg.2018.01109.

[2] M. ter Bekke, L. Drijvers, and J. Holler, "Hand Gestures Have Predictive Potential During Conversation: An Investigation of the Timing of Gestures in Relation to Speech," *Cogn. Sci.*, vol. 48, no. 1, p. e13407, 2024, doi: 10.1111/cogs.13407.

[3] E. Donnellan, L. E. Özder, H. Man, B. J. Grzyb, Y. Gu, and G. Vigliocco, "Timing relationships between representational gestures and speech: A corpus based investigation," in *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, J. Culbertson, A. Perfors, H. Rabagliati, and V. Ramenzoni, Eds., 2022, pp. 2052–2058.

[4] M. Chu and P. Hagoort, "Synchronization of Speech and Gesture: Evidence for Interaction in Action," *J. Exp. Psychol. Gen.*, vol. 143, no. 4, pp. 1726–1741, 2014, doi: https://doi.org/10.1037/a0036281.

[5] D. McNeill, *Hand and Mind: What gestures reveal about thought*. Chicago, IL: University of Chicago Press, 1992.

[6] D. P. Loehr, "Gesture and Intonation," Doctoral dissertation, Georgetown University, 2004.

[7] T. Leonard and F. Cummins, "The temporal relation between beat gestures and speech," *Lang. Cogn. Process.*, vol. 26, no. 10, pp. 1457–1471, Dec. 2011, doi: 10.1080/01690965.2010.500218.

[8] W. Pouw and J. A. Dixon, "Quantifying Gesture-Speech Synchrony," in *Proceedings of the 6th Gesture and Speech in Interaction Conference (GESPIN 6)*, K. J. Rohlfing, A. Grimminger, and U. Mertens, Eds., Paderborn: Universitaetsbibliothek Paderborn, 2019, pp. 75–80. doi: 10.17619/UNIPB/1-815.

[9] N. Esteve-Gibert and P. Prieto, "Prosodic Structure Shapes the Temporal Realization of Intonation and Manual Gesture Movements," *J. Speech Lang. Hear. Res.*, vol. 56, no. 3, pp. 850–864, Jun. 2013, doi: 10.1044/1092-4388(2012/12-0049).

[10] E. Krahmer and M. Swerts, "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception," *J. Mem. Lang.*, vol. 57, no. 3, pp. 396–414, Oct. 2007, doi: 10.1016/j.jml.2007.06.005.

[11] W. Pouw, S. J. Harrison, and J. A. Dixon, "Gesture–speech physics: The biomechanical basis for the emergence of gesture–speech synchrony," *J. Exp. Psychol. Gen.*, vol. 149, no. 2, pp. 391–404, Feb. 2020, doi: 10.1037/xge0000646.

[12] W. Pouw and S. Fuchs, "Origins of vocal-entangled gesture," *Neurosci. Biobehav. Rev.*, vol. 141, p. 104836, Oct. 2022, doi: 10.1016/j.neubiorev.2022.104836.

[13] W. Pouw, S. J. Harrison, N. Esteve-Gibert, and J. A. Dixon, "Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures," *J. Acoust. Soc. Am.*, vol. 148, no. 3, pp. 1231–1247, Sep. 2020, doi: 10.1121/10.0001730.

[14] H. S. H. Fung and P. P. K. Mok, "Temporal coordination between focus prosody and pointing gestures in Cantonese," *J. Phon.*, vol. 71, pp. 113–125, Nov. 2018, doi: 10.1016/j.wocn.2018.07.006.

[15] K. Franich and H. Keupdjio, "The Influence of Tone on the Alignment of Speech and Co-Speech Gesture," presented at the Speech Prosody 2022, May 2022, pp. 307–311. doi: 10.21437/SpeechProsody.2022-63.

[16] Y. Chen and C. Gussenhoven, "Emphasis and tonal implementation in Standard Chinese," *J. Phon.*, vol. 36, no. 4, pp. 724–746, Oct. 2008, doi: 10.1016/j.wocn.2008.06.003.

[17] J. Cao, "Prosodic prominence and phrasing in spoken Mandarin: The case of the 3rd tone," in *Speech Prosody 2012*, May 2012.

[18] Y. Chen and B. Braun, "Prosodic Realization of Information Structure Categories in Standard Chinese," in *Speech Prosody 2006*, ISCA Archive, 2006.

[19] Y. Hong, P. L. Rohrer, and H. R. Bosker, "Do beat gestures influence audiovisual lexical tone perception in Mandarin?," presented at the AMLaP Asia, Hong Kong, 2023.

[20] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: a Professional Framework for Multimodality Research," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, Eds., Genoa, Italy: European Language Resources Association (ELRA), 2006, pp. 1556–1559. [Online]. Available: https://hdl.handle.net/11858/00-001M-0000-0013-1E7E-4

[21] R Core Team, "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria, 2023. [Online]. Available: https://www.R-project.org/

[22] P. L. Rohrer *et al.*, "The MultiModal MultiDimensional (M3D) labeling system." Accessed: Mar. 22, 2023. [Online]. Available: https://osf.io/ankdx/

[23] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2022 1992. [Online]. Available: https://www.praat.org

[24] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using **lme4**," *J. Stat. Softw.*, vol. 67, no. 1, 2015, doi: 10.18637/jss.v067.i01.

[25] C. C. Voeten, "buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression." 2023. [Online]. Available: https://CRAN.R-project.org/package=buildmer