The Influence of Tone on the Alignment of Speech and Co-Speech Gesture

Kathryn Franich[1]*, Hermann Keupdjio*[2]

[1]University of Delaware
[2]McGill University

kfranich@udel.edu, hermann.keupdjio@mail.mcgill.ca

Abstract

Evidence continues to accrue suggesting that co-speech gestures form an integrated part of the prosodic system of languages. Several studies have highlighted a tight link between the timing of gestures of the hands and head with syllables bearing prosodic prominence. Most work to date has examined this relationship in Indo-European languages, where gestures appear to be crucially timed with respect to pitch-accented syllables. Less work has examined the timing of co-speech gestures in tonal languages, where pitch plays quite a different role within the phonological system. Here, we examine the influence of tone on the timing of manual co-speech gestures in Medmba, a Grassfields Bantu language spoken in Cameroon. We investigate 1) whether certain tones are more likely than others to associate with manual gestures in the language; and 2) whether the fine timing of the speech-gesture relationship is influenced by the tone or relative f_0 of the syllable it co-occurs with. Our findings indicated no preference for any one tone to occur with co-speech gestures. However, gesture apexes were found to align significantly later with respect to the accompanying syllable's vowel for low-toned syllables as compared with syllables of other tones.

Index Terms: co-speech gesture, tone, prominence, rhythm, Bantu

1. Introduction

1.1. Speech, Gesture, and Prominence

Recent work on the timing of co-speech gestures in several Indo-European languages points to the close link between speech and co-speech gesture as parallel modalities through which elements of prosodic structure and pragmatic information are expressed [1, 2, 3, 4, 5, 6]. In particular, a number of studies have now shown that gestures of the hands and head are preferentially timed to co-occur with prosodically 'prominent' events in speech, such as pitch-accented syllables, in languages such as English [4, 7, 3, 5], French [8], Italian [9], and Catalan [10]. Not only are pitch-accented syllables more likely to align with gestures than non-accented syllables, but the precise timing of speech and manual gesture has been argued to revolve around the relative timing of the gestural *apex*—the point of maximum extension of the articulators (e.g. the fingers)—and the aligning vowel's f_0 peak [3, 7], though other speech-based landmarks, such as the vowel onset or perceptual center have also been posited [7]. More recently, Im & Baumann [11] have identified a probabilistic relationship between specific pitch accents (e.g. L+H*, H*, !H*, and L*) and the likelihood of gesture co-occurrence in English. These findings mirror earlier findings by Baumann & Röhr [12] which suggest that pitch accents can be ordered in their relative prominence based on their pitch attributes. Based on a recent cross-linguistic investigation of perceived prominence, Cole et al. [13] suggest that some aspects of the acoustic signal, including higher peak f_0, may be interpreted as prominence-lending regardless of the relationship between pitch patterns, structural prominence, and pragmatic function in a given language. Findings from this study and others also indicate that increased intensity and duration are associated with perceived prominence cross-linguistically; these two variables have also been found to correlate with co-speech gesture presence [14, 15].

1.2. Gesture and f_0 in a Tone Language

Less is known about the constraints on the alignment of co-speech gesture in languages in which tone plays a lexically-contrastive role, and where the relationship between pitch and prominence is less straightforward. Aside from informing the typology of speech-gesture relations more broadly, understanding the relationship between tone, pitch, and gesture in tone languages has the potential to inform our understanding of the nature of prominence itself, and the relationship between articulatory constraints at the level of speech and gesture. For example, the results from Cole et al. suggest that, in spite of its primary role in cuing lexical contrasts, tone and/or f_0 might still have a role to play in cuing prominence in tonal languages; by extension, we might also expect that co-speech gestures will be more likely to coincide with syllables which bear high tones or relatively higher f_0. Furthermore, recent work by Pouw et al. [16] has suggested that the link between co-speech gesture and higher f_0 (as well as higher amplitude) may be driven by biomechanical factors leading to coupling of manual and articulatory movements. Such a link might be expected to be found across many different languages, perhaps even universally.

In the present work, we seek to understand whether a language which utilizes lexical tone may show a similar bias towards certain tonal and pitch patterns being aligned to co-speech gestures. Specifically, we investigate the alignment of co-speech gestures in Medʉmba, a Grassfields Bantu language spoken in Cameroon, in naturally-occurring conversational contexts. Similar to other Bantu languages, it has a two-tone system, with both high (H) and low (L) tones. Rising and falling 'contour' tones can also arise through a variety of morphological and syntactic processes (see [17] for further detail) but are reduceable to sequences of high + low (for falling) and low + high (for rising) (1).

(1)	H tone	mə́n	'child'
	L tone	mə̀n	'person/someone'
	HL falling	mə̂n sáŋɔ́	'chief's child'
	LH rising	mə̌n ná	'that person'

Table 1: *Four tonal patterns found in Medʉmba*

If higher pitch is associated with greater prominence for Medumba speakers, as was found for the languages investigated by Cole et al., then we might expect that gestures would be more likely to align with high tones, or with vowels with higher f_0, than with low tones, or vowels with lower f_0, in the language. Similarly, given that rises have been identified as more perceptually prominent than falls [12], it may be that syllables bearing LH rising tones will be more likely to occur with gestures in Medumba than those bearing HL falling contours or level H or L tones.

1.3. Tone, f_0, and Timing

Aside from the broad tendency for certain pitch accent types to associate with co-speech gestures cross-linguistically, the possibility also exists that more subtle patterns in timing will arise based on the tone or fundamental frequency of vowels. For example, size of pitch excursion has been found to influence perceived prominence (with larger excursion size associated with greater perceived prominence) [18], and therefore, may also exert an influence on co-speech gesture timing. Previous work has also shown that tone influences the *perceptual center*, or 'perceived moment of occurrence' [19] of a syllable in Medumba, such that p-centers occur later in low-toned syllables than high-toned syllables [20]. This pattern is thought to relate to the tendency for low tones to have pitch contours which are slightly falling in certain positions, leading to the illusion that their vowels are somewhat longer than those found with high tones (vowel duration is one factor which has been found to influence p-center location; [21]). Finally, as mentioned previously, pitch peak has been argued to be the acoustic landmark to which gesture apexes align most closely in some languages; if this is also the case in Medumba, then we would expect that the timing of the pitch accent peak would predict gesture timing.

1.4. Other Acoustic Variables

In addition to tone and f_0, we also examine how other acoustic variables, including vowel intensity and duration, predict gesture occurrence. Both of these variables have been found to associate with production of lexical stress, phrase-level accent, and contrastive focus across a number of languages [22, 23], making them good candidates as predictors of gesture alignment.

2. Method

2.1. Data Collection

Results are drawn from a corpus of naturally-occurring speech from four Medumba speakers (2 identifying as men and two as women) collected through interviews in Bangangté, Cameroon, in January of 2020. Participants responded to a series of questions about local customs around major events such as marriage ceremonies or the birth of a child. Participants were video- and audio-recorded in a quiet room on a Zoom Q8 recorder positioned approximately 2.5 meters from the participant. A separate time-aligned audio track was recorded of each participant's speech with a AKG C520 head-mounted microphone. The camera captured the participant's head, upper body, and lap. Participants were recorded for an average of 45 minutes. Data analyzed in the present work is based on 4565 vowels, 669 of which aligned to a gesture.

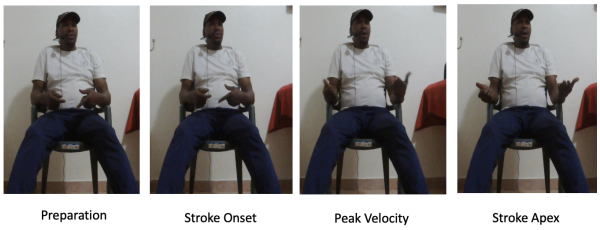

Figure 1: *Phases of a sample bimanual gesture*

2.2. Data Annotation and Preparation

Audio data were transcribed and glossed by the second author and subject to forced alignment using FAVE-align [24]; alignment was subsequently checked for accuracy and adjusted as necessary by the first author. Video data were annotated for several manual gestural landmarks by trained student annotators from the University of Delaware according to the MIT gesture studies coding manual [24]. Gestures were annotated with the sound muted so as to avoid possible bias from the audio signal. Landmarks included gesture preparation, stroke, hold, and recovery. In addition, the point of peak velocity (period within the stroke in which the hands moved with greatest velocity) was coded based on visualized extent of change in position of the hands between video frames and amount of blurring of the hands (Figure 1). Peak velocity of the hands has been found to immediately precede the timing of the gesture apex [7]—typically the point of maximum extension of the fingers (Figure 1)—and was a more viable landmark to use for apex calculation than point of maximum extension, which was harder to consistently assess through video data. We henceforth refer to this landmark as the gestural 'apex.' Inter-annotator reliability was achieved by having annotators work in pairs and checking to ensure that annotations between partners differed by no more than one 30 ms video frame from one another.

Annotated apexes were time-aligned to the speech signal in Praat [25]. Given the existing cross-linguistic evidence that vowels, rather than syllable onsets, more closely approximate the timing of gesture apexes, the time between each manual apex and the onset of the temporally-closest vowel was calculated. Only results for monosyllabic words, which accounted for around 70% of the gesture-aligned data overall, were considered here. Gestures occurring more than two standard deviations (around 200 ms) from a vowel were considered outliers and excluded. Target vowels were then analyzed for several f_0-related variables using ProsodyPro [26]. We note that all gesture types were considered for the present study, including 'beat gestures,' and more meaningful iconic and deictic gestures, since all of these gesture types can potentially be associated with prosodic prominence [27].

3. Results

3.1. Proportion of Tones Aligning with Apexes

Table 2 below provides the percentages of each tone found among syllables aligned to apexes compared with the percentages of tones not aligned with apexes. A Chi-square test of independence revealed that there was no significant difference between proportions (χ^2 (3, N = 5685) = .03, p = .99).

Table 2: *Proportions of gestures aligned to different tones*

	High	**Low**	**Falling**	**Rising**
Non-Gesture-Aligned	50.7	23	18.2	8.1
Gesture-Aligned	49.8	23.5	18.2	8.5

3.2. Tone, f_0, Intensity, and Duration as Predictors of Gesture Presence

Gesture presence vs. absence was modeled using mixed effects logistic regression with the `lme4` package in *R* statistical software. Predictor variables included Tone, Mean and Max f_0 (converted from raw Hz to semitones), Mean Intensity, and Duration of vowels. Tone was treated as a categorical variable with four levels (High, Low, Rising, Falling) and sum-coded. The other variables were treated as continuous and mean-centered to avoid collinearity. Two-way interactions between Tone and each of the four continuous predictors were also included. By-subject random intercepts were included for all models. Results revealed that that gesture presence was predicted to a significant degree by both vowel Intensity (β = 0.32, z = 3.39, $p < .001$) and Duration (β = 0.24, z = 4.11, $p < .001$) (Figures 2 and 3). No significant effects were found for Tone, Mean or Max f_0, or any of the interactions ($ps > .05$), though there was a marginal effect of Mean f_0 (β = -0.63, z = -1.76, p = .08); interestingly, vowels aligned to gestures trended toward having *lower* mean f_0, rather than higher f_0.

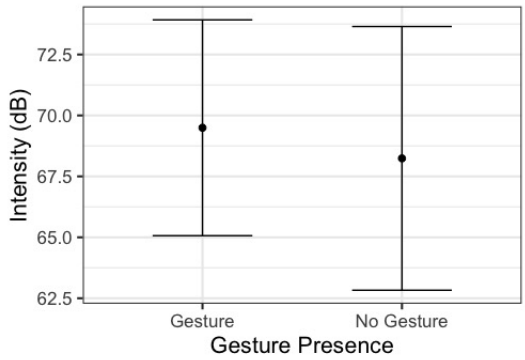

Figure 2: *Mean intensity by gesture presence*

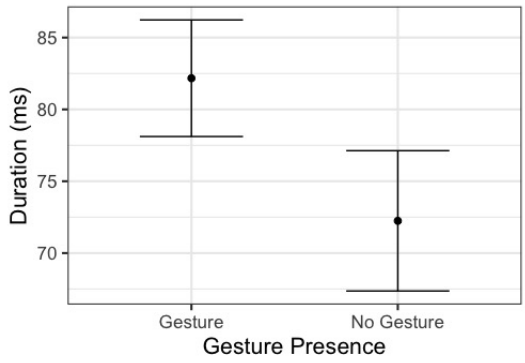

Figure 3: *Duration by gesture presence*

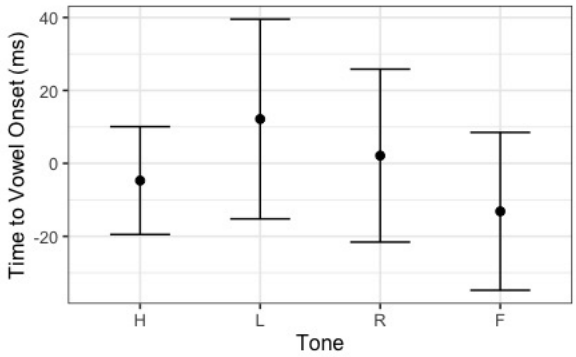

Figure 4: *Time from apex to vowel onset by tone*

3.3. Effects of Tone and f_0 on Apex-to-Vowel Timing

Six separate linear mixed effects models investigated the effects of tonal and pitch-related acoustic variables on the timing between gesture apexes and vowel onsets. These variables included Tone, treated as a categorical variable with four levels (High, Low, Rising, Falling), as well as Maximum f_0, Minimum f_0, f_0 Excursion Size, Time to f_0 Peak, and f_0 Peak Velocity. Raw Hz values of f_0 were converted to semitones. For all models, Vowel Duration was also included as a co-variate. Continuous predictors were all mean-centered and Tone was sum-coded. By-subject random intercepts were also included in all models.

Post-hoc comparisons were conducted using the `emmeans` package for *R*. Results revealed there was a significant effect of Tone on apex-to-vowel timing (Figure 4): gestures coinciding with L-tone vowels had apexes which were initiated significantly later relative to vowel onsets than gestures coinciding with H-tone vowels (β = 14.78, t = 2.03, $p < .05$) or than those associated with HL falling-tone vowels (β = 23.72, t = 2.61, $p < .01$). There was no significant difference in timing between H, LH rising, and HL falling tones in terms of timing ($ps > .05$). There was a significant effect of Vowel Duration, with longer vowels eliciting later apexes overall (β = 22.09, t = 7.56, $p < .001$). No significant effects were found for Maximum f_0, Minimum f_0 f_0 Excursion Size, Time to f_0 Peak, or f_0 Peak Velocity on apex-to-vowel timing ($ps > .05$). A marginal effect of Minimum f_0 was found.

4. Discussion

Our findings have revealed several patterns of interest in the alignment of speech and co-speech gesture in Medumba. In contrast with results from many Indo-European languages, co-speech gestures in Medumba are not preferentially aligned to vowels bearing higher f_0, or to vowels bearing high or rising tones. Indeed, the proportions of the tones found among the vowels with gestures aligned to them were almost identical to those of vowels with no gestures aligned to them. If anything, the trending effect of Mean f_0 found in Section 3.2 would seem to suggest that gestures are preferred to occur with lower f_0, rather than higher f_0. We can therefore conclude that there is no universal notion of pitch 'prominence' when it comes to the alignment of co-speech gesture across languages; the interpretation of pitch in relation to prominence appears to be conditioned by language-specific factors. We did, however, find that both Mean Intensity and Duration predicted gesture occurrence,

in the same direction as has been found for other languages: vowels which were louder and longer were more likely to occur with gestures. Indeed, duration has been identified as a correlate of phrase-level prominence previously in Medʉmba [27], and though the specific prosodic and information-structural factors which influence intensity in the language are not yet clear, it is unsurprising that this variable should be associated with gesture occurrence.

Our findings also call into question the notion that the relationship between gesture and pitch is conditioned by universal biomechanical factors related to the coupling of manual and articulatory variables, although our results do support the proposal by Pouw et al. [16] that manual gesture may entrain some aspects of phonation captured in the speech amplitude envelope. Though the lack of a significant relationship between f_0 and gesture occurrence in our study may be attributable to the less constrained nature of our conversational sample (as compared with the more tightly-controlled experiments conducted by Pouw et al.), the fact that our finding trended in the opposite direction from what is predicted from that study seems to suggest a genuine difference between Medʉmba and English in that regard. The finding that gesture is associated with increased intensity but marginally decreased f_0 in Medʉmba is also consistent with work which has suggested that the relationship between intensity and f_0 cannot be boiled down solely to increased subglottal pressure causing faster vocal fold vibration [28, 29].

Our findings furthermore highlight the more subtle influence that tone has on speech-gesture timing. Specifically, gestures accompanying low-toned syllables were found to be initiated later with respect to the vowel onset as compared with those accompanying high- or falling-toned syllables. These findings mirror previous findings suggesting that the p-center of low toned syllables in Medʉmba is later than that for high toned syllables [20]. The lateness of low tone p-centers was previously explained due to the slight fall in f_0 which occurs on low-toned syllables, particularly before pause. In light of this, it is interesting that falling toned syllables patterned in the opposite way from low-toned syllables with respect to gesture alignment, showing quite early alignment (even before the vowel was initiated). This suggests that manual gestures in Medʉmba are not universally timed with respect to the syllable/vowel, but to some other landmark. Surprisingly, no other pitch-related landmarks, including Excursion Size and Time to f_0 Peak, predicted gestural apex timing independent of Vowel Duration. Another possibility not explored here is that the Time to Peak Velocity of pitch movement could be an important factor which differentiates the two types of tones, as falling tones tend to demonstrate a steep fall shortly after the vowel is initiated, while low tones have a more gradual fall which occurs later in the vowel. Future work will need to explore this possibility.

There are, of course, many outstanding questions as to the nature of co-speech gesture alignment in Medʉmba. Evidence suggests that Medʉmba exhibits stem-initial prominence which is independent of tone [29]; stem-initial syllables are likely good candidates for gesture alignment, and future work will need to investigate this relationship, as well as the influence of information-structural factors on gesture occurrence. Going forward, we propose that co-speech gesture may provide an important tool for investigating the notion of prominence cross-linguistically. Given the observed links between perceived prominence and gesture, this relationship could prove particularly helpful for understanding prominence in languages, such as Medʉmba, which lack canonical acoustic correlates of word stress, and where contrastive focus and other aspects of information structure are encoded through means other than acoustic marking.

5. Conclusions

This work presents a study of tone, f_0, and other acoustic factors in conditioning the occurrence of co-speech gesture in Medʉmba. We have shown that the occurrence of manual gestures in Medʉmba is not strongly mediated by either tone or f_0 of accompanying speech, as has been found for many better-studied Indo-European languages. Nonetheless, the fine timing of manual gestures is found to be influenced by the tone of the accompanying vowel, with gestures initiated later relative to the onset of low-toned vowels as compared with vowels of other tones. We also find evidence that gesture occurrence is predicted by both increased duration and greater average intensity, similar to findings from other languages.

6. Acknowledgements

The authors would like to thank the Medʉmba speakers who participated in the study. This work was supported by National Science Foundation Linguistics Program Grant No. BCS-2018003 (PI: Kathryn Franich). The National Science Foundation does not necessarily endorse the ideas and claims in this paper. All errors are our own.

7. References

[1] A. Kendon, "Gesticulation and speech: two aspects of the process of utterance," in *The Relationship of Verbal and Nonverbal Communication*, M.R. Key, Ed. The Hague: Mouton, 1980, pp. 207-227.

[2] A. Kendon, *Gesture—visible action as utterance*. Cambridge, UK: Cambridge University Press, 2004.

[3] D. Loehr, *Gesture and Intonation*. Washington, D.C.: Georgetown University dissertation, 2012.

[4] D. McNeill, *Hand and Mind: What Gestures Reveal About Thought*. Chicago: University of Chicago Press, 1992.

[5] S. Shattuck-Hufnagel and A. Ren, "Preliminaries to a Kinematics of Gestural Accents," in *International Society for Gesture Studies 6*, 2018, San Diego.

[6] I. Hūbschler and P. Prieto, "Gestural and Prosodic Development Act as Sister Systems and Jointly Pave the Way for Children's Sociopragmatic Development." *Frontiers in Psychology*, 2019. https://doi.org/10.3389/fpsyg.2019.01259

[7] T. Leonard and F. Cummins, "The temporal relation between beat gestures and speech." *Language and Cognitive Processes*, 26, 10, 2011, pp. 1457-1471.

[8] N. Esteve-Gibert and P. Prieto, "Prosodic structure shapes the temporal realization of intonation and manual gesture movements." *JSLR*, 56, 3, 2013, pp. 850-64.

[9] D. Brentari, G. Marotta, I. Margherita, and A. Ott "The interaction of pitch accent and gesture production in Italian and English." in *SSL*, LI, 1, 2013, pp. 79-97.

[10] N. Esteve-Gibert, J. Borràs-Comes, E. Asor, M. Swerts, P. Prieto, "The timing of head movements: The role of prosodic heads and edges." in *JASA*, 141, 6, 2017, pp. 4727–4739.

[11] S. Im and S. Baumann, "Probabilistic relation between co-speech gestures, pitch accents and information status." In *Proceedings of the Linguistic Society of America*, 5, 1, 2020, pp. 685-697.

[12] S. Baumann, and F. Röhr "The perceptual prominence of pitch accent types in German," in *Proceedings of ICPHS*, 2015.

[13] J. Cole, J. Hualde, C.L. Smith, C. Eager, T. Mahrt, and R.N. de Souza, "Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish." *Journal of Phonetics*, 75, 2019, pp. 113–147.

[14] E. Krahmer and M. Swerts, "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception." *Journal of Memory and Language*, 57, 3 2007, pp. 396–414.

[15] A. Cravotta, M. Grazia Busà, and P. Prieto, "Effects of encouraging the use of gestures on speech." *JSLR*, 62, 2019, pp. 3204–3219.

[16] W. Pouw, S.J. Harrison, J.A. Dixon, "Gesture-speech physics: The biomechanical basis for the emergence of gesture-speech synchrony." *J Exp Psychol Gen.*, 149. 2. 2020, pp. 391-404.

[17] H.S. Keupdjio, *The syntax of A'-dependencies in Bamileke Medumba.* University of British Columbia dissertation.

[18] A.C. Rietveld and C. Gussenhoven. "On the relation between pitch excursion size and prominence." *Journal of Phonetics*, 13, 3, 1985, 299–308.

[19] J. Morton S. Marcus, and C. Frankish, "Perceptual centers (p-centers)." *Psychological Review*, 83, 1976, pp. 405–408

[20] K. Franich, "Tonal and morphophonological effects on the location of perceptual centers (p-centers): Evidence from a Bantu language." *Journal of Phonetics*, 67, 2018, pp. 21—33.

[21] A.M. Cooper, D.H. Whalen, and C.A. Fowler. "The syllable's rhyme affects its P-center as a unit." *Journal of Phonetics*, 16, 1988, pp. 231–241.

[22] M. Gordon and T. Roettger, "Acoustic correlates of word stress: A cross-linguistic survey." *Linguistics Vanguard*, 3,1, 1988, https://doi.org/10.1515/lingvan-2017-0007

[23] I. Vogel, A. Athanasopoulou, and N. Pincus, "Prominence, Contrast, and the Functional Load Hypothesis: An Acoustic Investigation." In J. Heinz, R. Goedemans, H. Van der Hulst (Eds.), *Dimensions of Phonological Stress*, Cambridge: Cambridge University Press, 2016, pp. 123-167.

[24] I. Rosenfelder, J. Fruehwald, K. Evanini, and J. Yuan. *FAVE (Forced Alignment and Vowel Extraction) Program Suite*, 2011, http://fave.ling.upenn.edu.

[24] MIT Speech Communications Group *Gesture Studies Coding Manual.* Retrieved June 2020 from http://web.mit.edu/pelire/www/manual/

[25] P. Boersma and D. Weenink, *Praat: doing phonetics by computer [Computer program].* Version 6.0, retrieved 2020 from https://www.praat.org

[26] Y. Xu. "ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis." In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, Aix-en-Provence, France. 7-10.

[27] P. Prieto, A. Cravotta, O. Kushch, P.L., and I. Vila-Gimenez, "Deconstructing beats: A labelling proposal," in *Proceedings of the Speech Prosody*, 2018, Poznan.

[27] K. Franich, "Uncovering tonal and temporal correlates of phrasal prominence in Medmba." *Language and Speech.* https://doi.org/10.1177/0023830919887994

[28] S. Tilsen. "A shared control parameter for F0 and intensity." In *Proceedings of Speech Prosody*, 2016, pp. 1066–1070. https://doi.org/10.21437/SpeechProsody.2016-219

[29] Z. Zhang, "Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model." *J. Acoust. Soc. Am.* 139, 2016, pp. 1493–1507. https://doiorg.udel.idm.oclc.org/10.1121/1.4944754

[29] K. Franich, Metrical prominence asymmetries in Medumba, a Grassfields Bantu language. Language 97, 2, 2021, pp. 365-402. doi:10.1353/lan.2021.0021.

[4] K. Franich and H.S Keupdjio, "The Influence of Tone on the Alignment of Speech and Co-Speech Gesture," in *Proceedings of Speech Prosody*, 2022.