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

The timing of beat gestures affects lexical stress perception in Spanish

Patrick Louis Rohrer [1], *Ronny Bujok* [2], *Lieke van Maastricht* [3], *Hans Rutger Bosker* [1,2]

[1]Donders Center for Cognition, Radboud University, Nijmegen, The Netherlands
[2]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[3]Centre for Language Studies, Radboud University, Nijmegen, The Netherlands
patrick.rohrer@donders.ru.nl, ronny.bujok@mpi.nl, lieke.vanmaastricht@ru.nl, hansrutger.bosker@donders.ru.nl

Abstract

It has been shown that when speakers produce hand gestures, addressees are attentive towards these gestures, using them to facilitate speech processing. Even relatively simple "beat" gestures are taken into account to help process aspects of speech such as prosodic prominence. In fact, recent evidence suggests that the timing of a beat gesture can influence spoken word recognition. Termed the *manual* McGurk Effect, Dutch participants, when presented with lexical stress minimal pair continua in Dutch, were biased to hear lexical stress on the syllable that coincided with a beat gesture.

However, little is known about how this manual McGurk effect would surface in languages other than Dutch, with different acoustic cues to prominence, and variable gestures. Therefore, this study tests the effect in Spanish where lexical stress is arguably even more important, being a contrastive cue in the regular verb conjugation system. Results from 24 participants corroborate the effect in Spanish, namely that when given the same auditory stimulus, participants were biased to perceive lexical stress on the syllable that visually co-occurred with a beat gesture. These findings extend the manual McGurk effect to a different language, emphasizing the impact of gestures' timing on prosody perception and spoken word recognition.

Index Terms: speech perception, gesture, lexical stress

1. Introduction

Communication is a multimodal affair, whereby interlocutors make use of both voice and body to convey meaning. Furthermore, both the auditory and visual modes of communication are temporally aligned so that in pitch-accent languages such as English or Spanish, gestures tend to co-occur with pitch-accented syllables (see [1] for a review). In laboratory settings where participants are asked to point or produce beat gestures (i.e., non-referential gestures that mark an underlying rhythmic pulse, see [2], [3]), this coupling in time has been shown to be quite precise: the 'apex' or point of maximum extension of the gesture most stably aligns with the peak of the pitch accent [4] and this gesture timing is affected by an upcoming prosodic boundary, showing similar patterns to pitch production at prosodic boundaries (namely showing a shift to earlier positions within the syllable to accommodate the additional boundary tone [5]).

The close link between gesture and prosody also has an impact on the listener. For example, the presence of a beat gesture coupled with a contrastive (L+H*) pitch accent leads to greater recall of target words than when presented with a pitch accent alone [6]. Such effects have also been found in EEG studies, where for example [7] found that words produced in contrastive focus with a beat gesture elicited a P300 component, which is said to reflect increased attention. This close link can even be used in language learning contexts, where beat gestures have been shown to facilitate word learning [8], the learning of lexical stress [9], and to improve pronunciation [10], [11].

In addition to areas of recall, attention, and learning, a number of studies have investigated the impact of manual gesture production on syntactic parsing and prominence perception. For example, regarding the former, [12] investigated how gesture can impact the perception of sentences that are syntactically ambiguous, yet can be disambiguated via prosody. They found that listeners showed no difference between modalities (audio-only, or AO vs. audiovisual, or AV conditions) when gestures were produced naturally to co-occur with the meaning conveyed with prosody. However, when seeing videos where the meaning of prosody and gesture did not match, listeners were significantly more likely to choose the meaning conveyed by gesture over that of prosody. Thus, the timing of gestures is used by listeners as an audiovisual prosody cue. Regarding the effects of beat gestures on the perception of speech prominence, [13] created matching and mismatching multimodal versions of Dutch sentences containing two target words, where a pitch accent could be placed on either the first or second word, and a beat gesture could also be placed on either the first or second word. The participants were then asked to rate the prominence of the target words in AO and AV conditions. They found that words produced with a beat gesture were perceived as more prominent in the AV condition than that same utterance in the AO condition. Additionally, when listeners saw a beat gesture on one target word, the perceived prominence on the other target word decreased.

The aforementioned studies show how the production of gestures facilitates the comprehension of semantic and pragmatic meanings at the level of the utterance (via syntactic parsing and focus marking). More recent work has targeted the effects of beat gestures on the perception of lexical stress (i.e., testing *word-internal* prominences), showing how in Dutch gestural timing can even impact spoken word recognition. In a 2-alternative forced choice (2-AFC) experiment, [14] used disyllabic lexical stress minimal pairs (e.g., *PLAto* vs. *plaTEAU*) to create lexical stress continua of manipulated audio which gradually went from a trochaic pattern (strong-weak; SW) to an iambic pattern (weak-strong; WS). The manipulated audios of the target words were then embedded in a carrier sentence and superimposed on a video of a speaker producing a beat gesture occurring with either the first syllable or the second syllable. The authors then presented these items to 48 native

10.21437/SpeechProsody.2024-142

speakers of Dutch, who were asked to choose the word they heard. Results showed that seeing a beat gesture on the first syllable (compared to seeing no gesture) prompted significantly more SW responses across all steps on the acoustic continuum. Follow-up experiments suggest that this effect of manual beat gestures on lexical stress perception is robust (e.g., also surfacing when data is collected online) and serves as an even stronger cue to lexical stress than articulatory cues to stress on the lips and face [15]. The authors thus termed this effect the *manual McGurk effect*, reflecting the classic finding by [16] that the visual modality has an impact on perception in the auditory modality, and subsequently, spoken word recognition.

The aim of the current study is to assess the generalizability of the manual McGurk effect. Specifically, we tested this effect in Spanish, where stress is a more informative lexical cue than in Dutch, as it is part of the regular verb conjugation system. Namely, lexical stress allows for the disambiguation between the verb inflections for first person present tense and third person preterit tense for many Spanish verbs (e.g., 'bailo', *I dance*, with stress on the initial syllable vs. 'bailó', *(s)he danced*, with stress on the final syllable). Therefore, one might expect the manual McGurk effect to have a larger effect size in Spanish than in Dutch (which only has a handful of, semantically unrelated, minimal pairs). In addition, we made several critical changes to the paradigm. First, whereas the continua in [14] were created by manipulating F0 only (keeping intensity and duration at ambiguous levels), the current study varies multiple cues to stress across the phonetic continua, as syllabic duration has been shown to be a strong cue to lexical stress in Spanish, with F0 being a more reliable cue to prominence at a phrasal level [17]. Second, we were able to test the effect in more items (given the larger repertoire of lexical stress minimal pairs in Spanish). Finally, the current study will make use of video recordings of a female speaker of Castilian Spanish producing a beat gesture with a different kinematic profile (i.e., a smaller, punctuating movement more in line with naturally produced gestures). Thus, we aim to assess the cross-linguistic generalizability of the manual McGurk effect in a more naturalistic setting with more variable stimuli properties. We expect to find the same pattern of results as reported in [14], namely that seeing a beat gesture on the first syllable will bias participants towards choosing more SW responses. Further, given the relative importance of stress as a lexical cue in Spanish, this effect size may be even larger than that reported for Dutch. Such findings would add to the growing body of literature that suggests speech perception is influenced by multimodal communication.

2. Methods

2.1. Participants

Twenty-four native speakers of Castilian Spanish (11 female, 11 male, 2 no response; Mean age: 28.26 ± 5.6) were recruited via Prolific and tested online using the Gorilla platform (http://gorilla.sc). Participants gave informed consent as approved by the Ethics Committee of the Social Sciences department of Radboud University (ECSW-LT-2023-8-31-15306). None of the participants reported any hearing or language deficits.

2.2. Materials

Materials consisted of 18 lexical stress minimal pairs (henceforth "items", see full list at https://osf.io/bmk2s/) that consisted entirely of segmentally identical verb conjugations, either in the first person singular in the present tense (e.g., 'bailo', *I dance*) or the third person singular in the preterit tense (e.g., 'bailó', *(s)he danced)*. A female native speaker of Castilian Spanish was recorded producing the single-word utterances twice with a beat gesture, and twice without. The speaker was instructed to produce the beat gesture in a way that was natural and comfortable for them. The video recordings were then manipulated in Praat [18] and Python [19] to create the final AV stimuli. All scripts for stimuli creation are publicly available at https://osf.io/bmk2s/.

2.2.1. Audio manipulation

The audio was extracted from the no-beat videos on a trial-by-trial basis. Syllable and segmental annotations were made manually in Praat. Adapting a script adopted from [20], we then created 11-step phonetic continua from clear SW ('bailo') to clear WS ('bailó') for each item. Specifically, the script took the WS recordings of each item and first linearly interpolated the two syllable durations in 11 steps (i.e., step 1 had a long first syllable, but step 11 had a short first syllable). Then the F0 contour was linearly interpolated in 10 ms time bins between the two original SW and WS contours, followed finally by interpolating the two syllables' mean intensity. This resulted in three-dimensional phonetic continua, where co-varying duration, intensity, and F0 cues together gradually changed from cueing SW to cueing WS. In an AO pretest, the manipulated audios (N = 198, 18 items x 11 steps per item) were presented to 12 native speakers of Castilian Spanish (not participating in the actual AV experiment), where they were asked to categorize the target words as either SW or WS in a 2-AFC task. Based on the results from the pretest, we selected five steps for each item that represented a perceptual continuum from SW to WS. The original recordings were then added at the extreme ends of the continua (i.e., steps 1 and 7 were completely unmanipulated) and thus the continua for each item contained 5 manipulated audio steps (in terms of duration, mean intensity, and F0).

2.2.2. Video manipulation

The original video recordings were imported into ELAN [21] for annotation of gesture production, focusing specifically on the gesture phase (that is, the stroke of the gesture as well as any preparation, hold, and recovery phases) and the apex (the point of maximum extension within the stroke). Based on this data, the average time-normalized position of the apex within the stressed syllable was calculated for each member of a lexical stress pair in R [22] (e.g., for a given word, the gesture's apex arrived at the 12%-point of a given stressed syllable's duration). This then informed a custom Python script that merged the manipulated audio and the original video file (taken from the WS recording for each word) so as to create two conditions: a version of the continua with the beat gesture apex occurring on the average time-normalized position within the first syllable ('beat-on-syllable-1'; Bo1) and a version of the continua with the beat gesture apex occurring on the average time-normalized position within the second syllable ('beat-on-syllable-2'; Bo2).

Figure 1 illustrates the two gesture conditions in experimental stimuli for the item "bailo" at step 4. Finally, the video was cropped to focus on the gesturing speaker and the speaker's face was masked to remove any articulatory cues to stress. In addition to these AV stimuli, videos were also made for 'catch trials' that were included to motivate participants to keep watching the AV stimuli (i.e., not close their eyes). A total of 18 videos (one for each lexical pair) was taken from the recordings without gesture, and a large red cross was superimposed on the video.

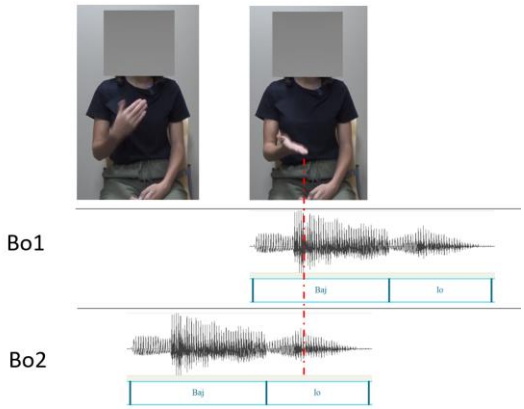

Figure 1: *Still images taken from the AV stimuli, showing the beginning and end (apex) of the gesture stroke. Lower panels show the temporal alignment with item "bailo" at ambiguous step 4 in Beat-on-syllable 1 (Bo1) and Beat-on-syllable 2 (Bo2) conditions (red dashed line indicates apex placement).*

2.3. Procedure

The experiment took place online on the Gorilla platform. After an initial screening to check for technical compatibility (i.e., checks for the use of headphones and media playback) and some pre-experimental individual differences tests (not reported here), the participants were given a short description of the 2-AFC task. They were specifically instructed to watch the video and decide which of the two target words they heard by pressing 'F' (stress on the first syllable, e.g., 'bailo') or 'J' (stress on the second syllable, e.g., 'bailó'). They were also told that if they saw a red cross in the video (i.e., a catch trial) to push the 'space bar', in an attempt to motivate participants not to close their eyes. They were given 6 practice trials and then moved on to the actual experiment. Items were presented in a randomized order (N = 252, 18 items x 7 steps x 2 conditions). Trials started with the presentation of a fixation cross along with the two potential target words on either side of the screen for 1000 ms. Then, the AV stimulus was played for the participant, after which, the target words reappeared on the screen and participants had to make their decision. They had 4 seconds to give a response before automatically moving on to the next trial. The entire experiment lasted approximately 1 hour and participants were renumerated for their participation at 6 GBP per hour.

3. Results

All scripts for statistical analysis are publicly available at https://osf.io/bmk2s/. Missing data due to time-out were excluded from analysis (14 trials; <0.1%). Figure 2 shows the percentage of SW responses that were recorded at each Step in the two Gesture conditions. Figure 3 illustrates some of the individual variation present in the data.

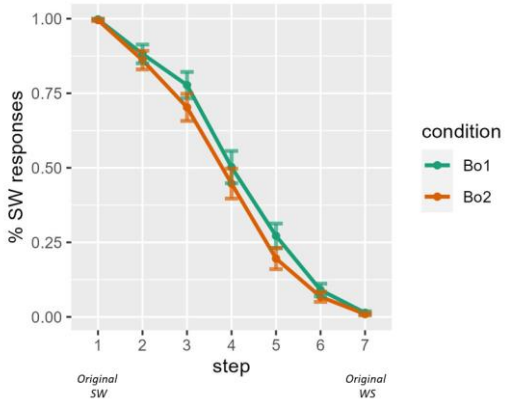

Figure 2: *Percentage of SW responses at each step by condition (green line = Beat-on-syllable 1, orange line = Beat-on-syllable 2), showing that the same acoustic stimulus was perceived as more trochaic (SW) if the beat occurred on the first syllable, but as more iambic (WS) when the beat occurred on the second syllable. Error bars represent standard error.*

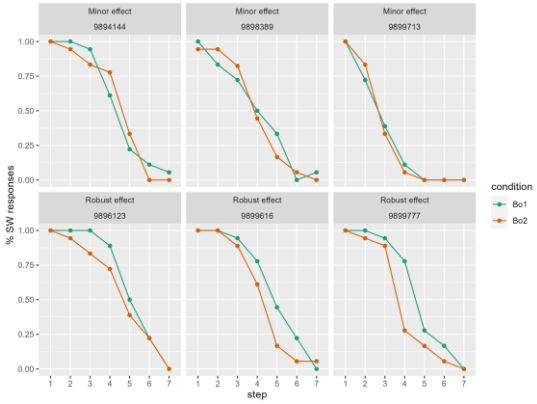

Figure 3: *Percentage of SW responses at each step by condition, sampled from 6 participants. The top row shows data from 3 participants with less clear effects, while the bottom row shows data from 3 participants with robust effects.*

Using the *lme4* package [23], a generalized linear mixed effects (GLM) model with logistic link function was built with the participants' responses as the binomial dependent variable (SW response coded as 1; WS response coded as 0) and fixed effects of Step (z-scored), Condition (categorical variable; Bo1 mapped onto the intercept), and their interaction. The random effects structure was determined using the *buildmer* package [24] which takes the most complex random effects structure as input, and returns the structure that best fits the data. Thus, the model included by-participant and by-item random intercepts. This model revealed a significant effect of Step ($\beta = -3.246$, s.e. = 0.106, $t = -30.582$, $p < .001$), as well as Condition ($\beta = -0.436$, s.e. = 0.089, $t = -4.906$, $p < .001$). Omnibus test results showed a significant main effect of Step ($\chi^2(1) = 1463.9$, $p < .001$) as well as a significant main effect of Condition ($\chi^2(1) = 23.87$, $p < .001$). A post-hoc Bonferroni-corrected pairwise test was carried out with the *emmeans* package [25] to further assess

the main effect of condition. It was found that when participants saw a beat gesture on the first syllable, they were significantly more likely to report hearing a SW pattern than a WS pattern ($z = 4.91$, $p < .001$) compared to trials where the beat gesture fell on the second syllable. There was no interaction between Step and Condition.

4. Discussion & Conclusions

The main objective of the current study was to corroborate the robustness of the manual McGurk effect by testing in a different language, namely Spanish. In this preliminary sample, 24 participants heard disyllabic single-word utterances sampled from 7-step continua from SW to WS (acoustically signaled by co-varying F0, duration, and mean intensity), while visually accompanied with a beat gesture either on the first or second syllable. The results showed that seeing a beat gesture on the first syllable led to significantly more reports of hearing the SW target word (e.g., 'bailo' with initial stress, *I dance*), while the same (possibly ambiguous) acoustic stimulus combined with a beat gesture on the second syllable biased perception towards the WS target word (e.g., 'bailó' with final stress, *(s)he danced*).

These results thus lend further support to the manual McGurk Effect by extending the effect to Spanish, where gestural timing could potentially be a more important cue in speech perception as we demonstrate here that it can distinguish verb inflections for the first person present tense and the third person preterit tense in many verbs. However, the effect size in the current study does not seem to be larger than previously reported for Dutch (where only few minimal stress pairs exist). If anything, it appears to be smaller (maximally 7% shift at ambiguous step 4 vs. an average of 20% in [14]; for similar results regarding how non-manual cues affect the perception of contrastive focus in Spanish, see [26]). One potential explanation is that Spanish listeners are used to assessing lexical stress contrasts, and thus may be more in tune with acoustic cues, relying less on visual cues. Nevertheless, strong conclusions about this comparison are not warranted given methodological differences. That is, the present smaller effect size could have been driven by the larger number of items (increasing item variability), greater acoustic variability, the precise gestural timing on a millisecond timescale, sample size, or variable gesture kinematics. Future studies may want to test bilingually-raised participants in both their native languages to be able to draw more direct cross-linguistic comparisons.

In any case, the current study takes an important step in extending the effect with more naturalistic stimuli, particularly in terms of (a) the kinematic profile of the gesture, and (b) the acoustic cues to prominence. Regarding the kinematic gesture profile, the present outcomes emphasize the impact of relatively simple hand gestures on spoken word recognition. Previous studies presented participants with a male speaker producing a rather forceful beat gesture, while the current speaker presented a female using more naturally produced subtle beat gestures. Studies have shown that biomechanical forces from hand movements and gestures have an impact on speech acoustics ([22], [23]). Indeed, it has been argued that the manual McGurk Effect may not necessarily reflect the perception of the auditory signal alone, but rather that participants are perceiving a "limb-vocalic speech act, and varying information about physical impulses of gesture interacts with audition in the perception of a more global array of multimodal information" [24, p. 1]. Though the gesture used in the current study does not directly address this issue (as biomechanical effects may still be present in the gestures that were produced), it is interesting to see a more subtle gesture produce a manual McGurk effect. Future studies may assess whether the effect would be present with much smaller gestures, as "beat" gestures have been described to be as small as simple flicks of the finger [2].

The current study confirmed the effect when varying multiple cues to prominence (particularly syllabic duration as a major cue to word-level stress prominence in Castilian Spanish [17]). However, further cross-linguistic comparisons of how gestural timing can impact speech perception are needed. For example, less is known about how this effect would play out in fixed-stress languages which do not have lexical stress minimal pairs (e.g., French). Beat gestures could then potentially impact word recognition by means of facilitating speech segmentation. Alternatively, assessing the effect in tonal languages where specific F0 contours (i.e., lexical tones) act as a major cue for word recognition would further our cross-linguistic knowledge of how gesture timing and speech acoustics interact, impacting speech perception.

Finally, little is known about the ways in which individual variation plays a role in the manual McGurk Effect. As shown in Figure 3, some participants showed no clear patterns of the manual McGurk effect (top row) while others showed quite consistent patterns (bottom row). While the manual McGurk effect is rather robust at the population level, it would be interesting to see what different cognitive factors may drive these individual differences, such as phonological/visuospatial working memory, musical abilities, or other cognitive measures. Some of these individual differences may even lead to explanations as to why some individuals differ in their timing patterns of multimodal speech production. Indeed, most studies on the temporal integration of speech and gesture focus on group averages, and much less is known about individual differences in multimodal speech production, even though it has been shown that considerable individual variation does indeed exist (See [30] for a short discussion). All in all, the current study adds further evidence to the idea that the temporal gesture-speech alignment patterns impact the sounds we hear, and opens up multiple future lines of research to better understand multimodal communication, both in perception and 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 the Centre for Language Studies at Radboud University for a Small Research Grant (#023) awarded to Lieke van Maastricht. Finally, we would also like to thank the members of the SPEAC research group for their helpful feedback on the study, as well as our speaker for volunteering their time.

6. References

[1] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," *Speech Commun.*, vol. 57, pp. 209–232, Feb. 2014, doi: 10.1016/j.specom.2013.09.008.

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

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

[4] 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.

[5] 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).

[6] O. Kushch and P. Prieto, "The effects of pitch accentuation and beat gestures on information recall in contrastive discourse," in *Speech Prosody 2016*, ISCA, May 2016, pp. 922–925. doi: 10.21437/SpeechProsody.2016-189.

[7] D. Dimitrova, M. Chu, L. Wang, A. Özyürek, and P. Hagoort, "Beat that Word: How Listeners Integrate Beat Gesture and Focus in Multimodal Speech Discourse," *J. Cogn. Neurosci.*, vol. 28, no. 9, pp. 1255–1269, Sep. 2016, doi: 10.1162/jocn_a_00963.

[8] O. Kushch, A. Igualada, and P. Prieto, "Prominence in speech and gesture favour second language novel word learning," *Lang. Cogn. Neurosci.*, vol. 33, no. 8, pp. 992–1004, Sep. 2018, doi: 10.1080/23273798.2018.1435894.

[9] L. van Maastricht, M. Hoetjes, and E. van Drie, "Do gestures during training facilitate L2 lexical stress acquisition by Dutch learners of Spanish?," in *The 15th International Conference on Auditory-Visual Speech Processing*, ISCA, Aug. 2019, pp. 6–10. doi: 10.21437/AVSP.2019-2.

[10] D. Gluhareva and P. Prieto, "Training with rhythmic beat gestures benefits L2 pronunciation in discourse-demanding situations," *Lang. Teach. Res.*, vol. 21, no. 5, pp. 609–631, Sep. 2017, doi: 10.1177/1362168816651463.

[11] J. Llanes-Coromina, P. Prieto, and P. Rohrer, "Brief training with rhythmic beat gestures helps L2 pronunciation in a reading aloud task," in *Speech Prosody 2018*, ISCA, Jun. 2018, pp. 498–502. doi: 10.21437/SpeechProsody.2018-101.

[12] B. Guellaï, A. Langus, and M. Nespor, "Prosody in the hands of the speaker," *Front. Psychol.*, vol. 5, Jul. 2014, doi: 10.3389/fpsyg.2014.00700.

[13] 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.

[14] H. R. Bosker and D. Peeters, "Beat gestures influence which speech sounds you hear," *Proc. R. Soc. B Biol. Sci.*, vol. 288, no. 1943, p. 20202419, Jan. 2021, doi: 10.1098/rspb.2020.2419.

[15] R. Bujok, A. S. Meyer, and H. R. Bosker, "Audiovisual Perception of Lexical Stress: Beat Gestures are stronger Visual Cues for Lexical Stress than visible Articulatory Cues on the Face." 2022. doi: 10.31234/osf.io/y9jck.

[16] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, Dec. 1976, doi: 10.1038/264746a0.

[17] M. Ortega-Llebaria and P. Prieto, "Acoustic Correlates of Stress in Central Catalan and Castilian Spanish," *Lang. Speech*, vol. 54, no. 1, pp. 73–97, Mar. 2011, doi: 10.1177/0023830910388014.

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

[19] G. Van Rossum and F. L. Drake Jr., *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[20] R. Bujok and H. R. Bosker, "F0 stress continuum interpolation." 2023. [Online]. Available: https://hrbosker.github.io/resources/scripts/interpolate-f0/

[21] 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

[22] 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/

[23] 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.

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

[25] R. V. Lenth *et al.*, "emmeans: Estimated Marginal Means, aka Least-Squares Means." Oct. 17, 2023. Accessed: Dec. 06, 2023. [Online]. Available: https://cran.r-project.org/web/packages/emmeans/index.html

[26] P. Prieto, C. Puglesi, J. Borràs-Comes, E. Arroyo, and J. Blat, "Exploring the contribution of prosody and gesture to the perception of focus using an animated agent," *J. Phon.*, vol. 49, pp. 41–54, Mar. 2015, doi: 10.1016/j.wocn.2014.10.005.

[27] 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.

[28] 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.

[29] W. Pouw and J. A. Dixon, "What you hear and see specifies the perception of a limb-respiratory-vocal act," *Proc. R. Soc. B Biol. Sci.*, vol. 289, no. 1979, p. 20221026, Jul. 2022, doi: 10.1098/rspb.2022.1026.

[30] 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.