How to test gesture-speech integration in ten minutes

Matteo Maran, Hans Rutger Bosker

Donders Institute for Brain, Cognition and Behaviour, Radboud University
matteo.maran@donders.ru.nl, hansrutger.bosker@donders.ru.nl

Abstract

Human conversations are inherently multimodal, including auditory speech, visual articulatory cues, and hand gestures. Recent studies demonstrated that the timing of a simple up-and-down hand movement, known as a beat gesture, can affect speech perception. A beat gesture falling on the first syllable of a disyllabic word induces a bias to perceive a strong-weak stress pattern (i.e., "CONtent"), while a beat gesture falling on the second syllable combined with the same acoustics biases towards a weak-strong stress pattern ("conTENT"). This effect, termed the "manual McGurk effect", has been studied in both in-lab and online studies, employing standard experimental sessions lasting approximately forty minutes. The present work tests whether the manual McGurk effect can be observed in an online short version ("mini-test") of the original paradigm, lasting only ten minutes. Additionally, we employ two different response modalities, namely a two-alternative forced choice and a visual analog scale. A significant manual McGurk effect was observed with both response modalities. Overall, the present study demonstrates the feasibility of employing a ten-minute manual McGurk mini-test to obtain a measure of gesture-speech integration. As such, it may lend itself for inclusion in large-scale test batteries that aim to quantify individual variation in language processing.

Index Terms: beat gesture, lexical stress, audiovisual integration, gesture-speech integration, prosody

1. Introduction

In face-to-face communication, interlocutors provide each other with both acoustic (e.g., speech) and visual information [1]. The classic McGurk effect [2] is a well-known example of how visual information on lip movements affects speech perception. Gestures are an additional source of visual information produced in combination with speech during conversations [3]. Rapid up-and-down strokes of the hand, called beat gestures, are one of most frequently produced gestures [3], affecting comprehension already in early stages of development [4].

A defining functional property of beat gestures is their timing [5]. Beat gestures increase the prominence of the speech unit that is produced simultaneously with them [6]. This property is commonly exploited to emphasize a specific word in a sentence [7]. Interestingly, recent studies showed that beat gestures affect comprehension even at a lexical level, such as single word identification ([8], [9]). A beat gesture falling on the first syllable of a disyllabic word increases the syllable's prominence, inducing a bias towards perceiving a strong-weak (SW) stress pattern (e.g., "CONtent"). Conversely, a beat gesture falling on the second syllable of the same word biases towards perceiving a weak-strong (WS) stress pattern ("conTENT"). This bias has been termed the "manual McGurk effect" [8], and refers to the influence of gestural and prosodic temporal alignment on speech perception. That is, just like how lip movements shape speech perception in the classic McGurk effect, so can the timing of simple hand movements influence word recognition in the manual McGurk effect.

Previous manual McGurk studies ([8], [9]) investigated the influence of beat gesture on speech perception in seven steps across the phonetic continuum between clear SW and clear WS stress, resulting in a large number of trials that added up to sessions lasting approximately forty minutes. Such prolonged experimental sessions might limit the application of the manual McGurk paradigm in settings with high time constraints. A first goal of the present work is to test the possibility of employing an online and short version (hence "mini-test") of the manual McGurk paradigm, lasting approximately ten minutes, to test gesture-speech integration. An efficient manual McGurk mini-test would allow to easily investigate the most basic level of gesture-speech integration (i.e., single word), laying the ground for understanding multisensory effects arising at more complex levels (e.g., sentential level). Furthermore, the mini-test could be exploited both in basic research and clinical settings, for example including the mini-test in batteries characterizing different aspects of multimodal communication, of easy application also in populations with limited sustained attention.

A second goal of the present study is to explore the role of the response modality in the manual McGurk effect size. Previous studies employing this paradigm ([8], [9]) assessed participants' perception using a two-alternative forced choice (2AFC) task. We here test for the first time whether a significant manual McGurk effect can also be observed employing a visual analog scale (VAS), which allows to provide a response graded by confidence. With the VAS, participants are presented with a slider bar having at the two extremes two options (e.g., "CONtent" on the left, and "conTENT" on the right). Participants can provide a graded response by precisely moving the slider's handle more to the left or to the right: for example, by moving it only half way towards the left they can indicate that what they heard sounds like "CONtent", but that they are not totally sure. The VAS appears to limit the extent to which individuals transpose a graded perceptual experience into a discrete and categorical response [10], possibly revealing features of multisensory processing left unexamined by previous 2AFC studies. Moreover, the fine-grained property of VAS responses might prove particularly useful to model individual variability in beat gesture-speech integration when using the manual McGurk mini-test, considering the reduced number of trials presented to the participants. The present study assesses potential differences and similarities between the 2AFC and VAS response modalities in audiovisual processing, presenting the same group of participants with both.

Overall, we hypothesize that a significant manual McGurk effect will be observed in an online ten-minute mini-test, both with the 2AFC and VAS response modalities. Additionally, we

expect that the effect sizes obtained with the two response modalities will be correlated.

2. Method

The present hypotheses and methods were pre-registered as secondary analyses for a larger study, composed of two experimental sessions. The full pre-registration can be found at https://osf.io/6w348. Deviations from the pre-registration are explicitly mentioned. The analyses presented in this manuscript concern data from Session 1 only, in line with the scope of the present work. The pseudonymized experimental data and the code used in the analysis are available at https://osf.io/qbyfm.

2.1. Participants

We pre-registered that a sample of 32 participants would be included in the analysis. At the time of submission only a sample of 28 participants could be recruited (15 females, 13 males; mean age = 25 years, range = 18-39 years), recruited via Prolific and Radboud Research Participation System (SONA). Deviations from the pre-registration were implemented to attempt to reach the pre-registered sample size and are described in detail in https://osf.io/qbyfm. Participants recruited via Prolific were reimbursed with 9.75 British pounds for taking part in two experimental sessions, separated by at least a week (only data from Session 1 are relevant for the present study). Participants recruited via SONA received course credit for participation. All participants were native Dutch speakers, raised with native language only, who did not take part into previous studies employing similar paradigms from our lab, and who did not have any language disorder, Autism Spectrum Disorder, hearing or literacy difficulties. All participants in this study gave informed consent as approved by the Ethics Committee of the Social Sciences department of Radboud University (project codes: ECSW-LT-2023-7-6-20937 and ECSW-LT-2023-11-28-35780).

2.2. Materials and design

2.2.1. Experimental conditions and stimulus materials

In both tasks, the participants watched videos of a male native speaker of Dutch producing a single word out of four minimal pairs, whose members differ in stress pattern: *CONtent/conTENT* ("content/satisfied"), *SERvisch/serVIES* ("Serbian/tableware"), *VOORnaam/voorNAAM* ("first name/respectable"), and *VOORruit/voorUIT* ("windshield/forward"). The employed videos were taken from an adapted version of the manual McGurk paradigm [9], whose materials are publicly available. In the videos, a beat gesture could be absent (NoBeat condition), be aligned to the vowel onset of the first syllable (BeatOn1 condition), or be aligned to the vowel onset of the second syllable (BeatOn2 condition). The following trials were present for each minimal pair:

- Four NoBeat trials with an acoustically clear SW stress pattern, and four NoBeat trials with an acoustically clear WS stress pattern. These trials served for perceptual anchoring.
- Two NoBeat catch trials, one with an acoustically clear SW stress pattern and one with a clear WS one. In these trials, a white cross appears on top of the speaker's face. These trials served to motivate participants not to close their eyes during the task.
- Five NoBeat trials with a word with ambiguous acoustic information on the stress pattern.
- Five BeatOn1 trials with a word with ambiguous acoustic information on the stress pattern.
- Five BeatOn2 trials with a word with ambiguous acoustic information on the stress pattern.

The acoustically clear SW and WS patterns of stress involved original audio without any acoustic manipulation. The acoustically ambiguous stress patterns were sampled from 7-step phonetic stress continua involving gradual interpolation of the F0 contours from the SW and WS members of a pair (see [9] for more details). The item-specific steps chosen for the ambiguous condition (either step 3 or 4) were selected based on previously collected perceptual data [9], demonstrating close to 50% SW responses as well as good susceptibility to effects of gestural alignment. In NoBeat trials with clear acoustic information, the lip movements were congruent with the stress information. In BeatOn1 and BeatOn2 trials, the lip movements were congruent with an SW and a WS stress pattern, respectively. In NoBeat trials with ambiguous acoustic information, the lip movements were compatible three times with one stress pattern and two times with the other, according to lists (A and B) counter-balanced across participants. Converging on previous results [9], no significant effect of lip movement on audiovisual lexical stress perception in NoBeat trials with ambiguous acoustic information was observed in the 2AFC ($p = 0.538$) and VAS ($p = 0.723$) tasks. In total, 100 trials were included in both the 2AFC and VAS tasks and presented in a randomized order. The analyses focus on the 60 trials with ambiguous acoustic information.

2.2.2. General procedures

The experiment was created and hosted with the online experiment builder Gorilla [11]. Participants were asked to use headphones connected to their computer or laptop via cable and to ensure that no battery save mode was active. After providing basic demographic information, participants were presented with a headphone screening test based on dichotic pitch [12], designed to exclude participants not using headphones.

Participants passing the headphone screening test were presented with the two manual McGurk tasks (2AFC and VAS). In both the 2AFC and VAS manual McGurk tasks, each trial began with two members of a minimal pair presented in written form to the participant: the member with an SW stress pattern (e.g., *CONtent*) on the left side of the screen, the member with a WS stress pattern (e.g., *conTENT*) on the right side. After a fixation cross lasting for 500 milliseconds (ms), a video of a speaker producing a word with different audio conditions (SW, WS, or ambiguous stress pattern), with or without a beat gesture, was shown to the participants. In the 2AFC task, the videos were followed by the presentation of three response buttons: the member of the minimal pair with an SW stress pattern on the left, the member with a WS stress pattern on the right, and *Ik zag een WIT KRUIS!* ("I saw a WHITE CROSS!", to respond to catch trials) below in the center. After clicking on one of the options or 4 seconds passed, the next trial started after a 500 ms blank screen.

In the VAS task, the response was made by means of a sliding bar. The sliding bar was positioned between the SW option (on the left) and the WS option (on the right). One hundred and one values were organized progressively on the sliding bar, ranging from 0 to 100. These values were not

overtly shown to the participant, but served to quantify their response. The left half of the sliding bar indicated a preference for an SW stress pattern, with values ranging from 0 (strongest preference for SW) to 49 (minimal preference for SW). The center of the bar (50) indicated no preference. The right side of the bar indicated a preference for a WS stress pattern, with 51 indicating a minimal preference for WS and 100 indicating the strongest one. No time limit was present in the VAS task. On average, the 2AFC task lasted 9 minutes and 24 seconds, while the VAS task lasted 11 minutes and 42 seconds.

2.2.3. Statistical analysis

Data analysis was implemented using R software [13] and the R packages "lme4" [14] (2AFC task), "lmerTest" [15] (VAS task), and "car" [16] (calculation of logit values).

2.2.3.1 2AFC task

Data from the 2AFC task were analyzed using a generalized linear mixed model. The dependent variable (perceived stress) was coded as follows: SW = 1, WS = 0. The analysis focused on trials with ambiguous acoustic information. The fixed effects structure included only the factor Beat, with the three levels (NoBeat, BeatOn1, BeatOn2) treatment-coded with BeatOn1 as reference level. The manual McGurk effect was quantified by the contrast between BeatOn1 and BeatOn2, following [8]. The random effects structure included random slopes for Beat and random intercepts by participant, and random intercepts by word pair. More complex random effects structures, including the pre-registered one, failed to converge or led to singular fit issues. The effect size for the 2AFC task was calculated as the difference between the logit of the proportion of SW responses in BeatOn1 trials and the logit of the proportion of SW responses in BeatOn2 trials.

2.2.3.2 VAS task

The VAS data were scaled between 0 and 1, re-oriented to have 0 corresponding to the strongest confidence in WS and 1 to the strongest confidence in SW, and transformed with a logit function. Since logit(1) and logit(0) correspond to plus and minus infinity, these values were adjusted respectively to logit(0.999) and logit(0.001). Data were then analyzed with a linear mixed model. The fixed effects structure included only the factor Beat. The random effects structure included random intercepts by participant, word pair, and randomization group. More complex random effects structures, including the pre-registered one, failed to converge or led to singular fit issues. The effect size for the VAS task was calculated as the difference between the logit of the average of scaled re-oriented responses in BeatOn1 trials and the logit of the average of scaled re-oriented responses in BeatOn2 trials.

2.2.3.3 2AFC and VAS relationship

The relationship between the 2AFC and VAS individual effect sizes was analyzed with a Pearson's correlation.

3. Results

Participants only rarely failed to report a catch trial (10 and 11 misses in total in the 2AFC and VAS tasks, respectively), or wrongly reported a catch trial when it was not present (0 and 2 false positives in total in the 2AFC and VAS tasks, respectively). This suggests that participants performed both tasks with their eyes open and under good attentive conditions.

3.1. Manual McGurk effect in the 2AFC task

The contrast between BeatOn1 and BeatOn2 was significant for the 2AFC task ($p < 0.001$), indicating that the same ambiguous acoustic recordings were categorized more frequently as SW when the beat fell on the first syllable compared to when it fell on the second one (i.e., manual McGurk effect, Figure 1). The contrast between BeatOn1 and NoBeat was also significant ($p < 0.001$), indicating fewer SW responses when a beat was absent compared to when it fell on the first syllable.

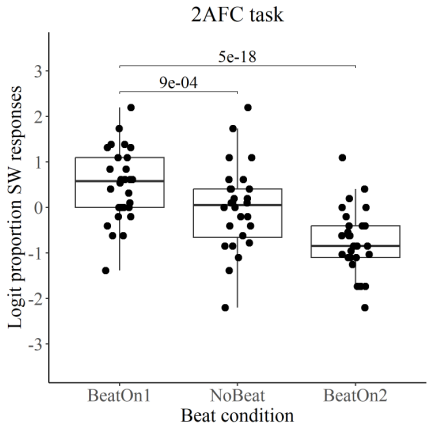

Figure 1: *The same acoustic recordings are perceived differently depending on the beat presence and its timing in the 2AFC task. The y axis indicates the logit of the proportion of SW responses, with the value of 0 corresponding to 50% proportion. The x axis indicates the different beat conditions. The black dots indicate subject-specific averages.*

3.2. Manual McGurk effect in the VAS task

The contrast between BeatOn1 and BeatOn2 was significant for the VAS task ($p < 0.001$), indicating a bias towards perceiving the same ambiguous acoustics as more SW-like when the beat fell on the first syllable compared to when it fell on the second one (i.e., manual McGurk effect, Figure 2). The contrast between BeatOn1 and NoBeat was not significant ($p > 0.05$), despite the numerically smaller preference for SW when no beat was present compared to when it fell on the first syllable.

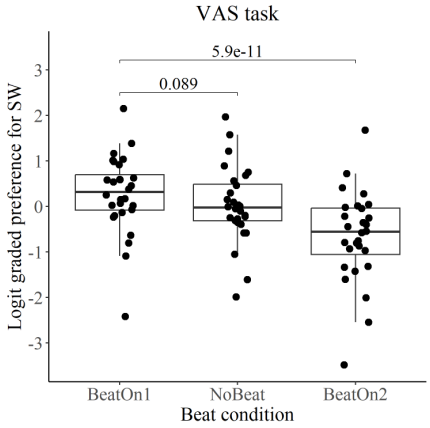

Figure 2: *The same acoustic recordings are perceived differently depending on the beat presence and its timing in the VAS task. The y axis indicates the average logit-transformed graded preference for SW, with the value of 0 corresponding to no preference. The x axis indicates the different beat conditions. The black dots indicate subject-specific averages.*

3.3. Relationship between the 2AFC and VAS effect sizes

The correlation between the individual effect sizes of the 2AFC and VAS tasks failed to reach significance ($r(26) = 0.37$, $p = 0.051$, Figure 3). Only a trend towards a significant relationship between the two tasks' effect sizes can be appreciated.

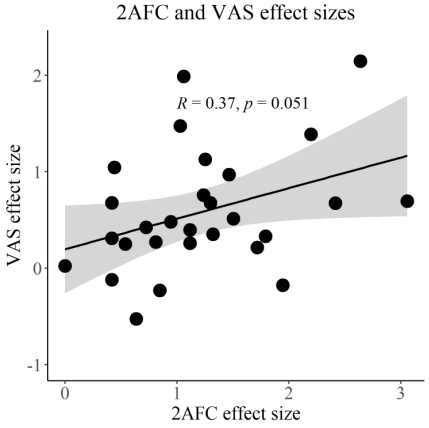

Figure 3: *Correlation between individual manual McGurk effect sizes in the 2AFC (on the x axis) and VAS (on the y axis) tasks. The black dots indicate subject-specific effect sizes.*

4. Discussion

The present study tested the feasibility of employing a ten-minute mini-test to quantify the manual McGurk effect, testing for the first time two response modalities. A significant manual McGurk effect was observed both in the 2AFC and the VAS tasks. Notably, these significant effects at the group level were achieved despite the minimal number of trials present in each of the three experimental conditions tested (i.e., 20 per participant) and a moderate sample size analyzed. These results speak in favor of a reliable influence of beat gestures on speech perception, which can be quantified in a short ten-minute experimental session. The sensitivity and short duration of the mini-test can be exploited by test batteries focusing on multisensory processing in communication, of easy use especially in clinical settings that are arguably more affected by time-constraints compared to in-lab studies. In this case, the graded information provided by the VAS might prove particularly useful to quantify any gradual alterations and improvements in gesture-speech integration. Another avenue for the application of the VAS scale is research on multimodal communication in a second language ([17], [18]), possibly capturing graded changes in gesture-speech integration at different levels of proficiency.

Beyond validating the mini-test, the present study examined similarities and differences between two response modalities commonly employed in the speech perception literature, namely 2AFC and VAS [10]. Contrary to our predictions, the individual effect sizes in the 2AFC and VAS tasks did not significantly correlate. It is possible that larger samples are needed to highlight a significant correlation at the group level between the two effect sizes, for which we could observe only a moderate trend towards significance. A related issue concerns whether the VAS might be an appropriate response modality in experimental designs like the one here employed, in which the manipulation of interest (i.e., beat timing) involves only a minimal number of levels (i.e., two: aligned to either the vowel onset of the first or second syllable) rather than multiple ones differing more gradually. Further studies are needed to shed light on these open questions. An additional aspect which deserves further examination concerns the contrast between BeatOn1 and NoBeat, which was significant in the 2AFC task but not in the VAS, where only a trend towards significance was observed. In the 2AFC, participants are pushed to guess when they are unsure, therefore a reliable but small and implicit bias in the BeatOn1 condition might lead to a significant difference compared to a condition where no bias is present (NoBeat). In the VAS, participants might moderate the implicit bias induced by the beat, choosing values that are closer to the center of the sliding bar. Possibly, increased power is needed to observe a statistical difference between BeatOn1 and NoBeat in the VAS.

An interesting research question for future studies concerns whether the temporal alignment between prosodic and gestural prominence affects lexical processing or only earlier perceptual analysis. Evidence from investigations employing pseudo-words [8] suggests that the manual McGurk effect can be observed as long as the phonotactic rules of a given language are followed.

On a final note, the present study converges on the feasibility of employing online testing to address research questions on audiovisual integration in speech processing [9]. Building on these observations, future online studies might be designed to test gesture-speech processing in populations with limited mobility (e.g., patients) or whose geographical location might prevent them from joining in-lab sessions (e.g., students studying a second language abroad). An additional line of research that could be addressed via online studies concerns cross-linguistic variation in the manual McGurk effect, expanding the available findings reported in Dutch.

5. Conclusions

The present work showed that it is possible to test beat gesture and speech integration in approximately ten minutes. A reliable manual McGurk effect was observed both using a 2AFC and a VAS task as response modality, speaking in favor of employing either of them according to specific needs of the researcher or clinician. Surprisingly, the individual effect sizes obtained with the two response modalities did not significantly correlate, albeit a trend towards significance was observed. Further studies are required to understand the relationship between the 2AFC and VAS effect sizes.

6. Acknowledgements

Funded by an ERC Starting Grant (HearingHands, 101040276) from the European Union awarded to Hans Rutger Bosker. 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.

7. References

[1] J. Holler and S. C. Levinson, "Multimodal Language Processing in Human Communication," *Trends in Cognitive Sciences*, vol. 23, no. 8, pp. 639–652, Aug. 2019, doi: 10.1016/j.tics.2019.05.006.

[2] H. Mcgurk and J. Macdonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, Art. no. 5588, Dec. 1976, doi: 10.1038/264746a0.

[3] D. McNeill, *Hand and mind: What gestures reveal about thought.* in Hand and mind: What gestures reveal about thought. Chicago, IL, US: University of Chicago Press, 1992, pp. xi, 416.

[4] J. Llanes-Coromina, I. Vilà-Giménez, O. Kushch, J. Borràs-Comes, and P. Prieto, "Beat gestures help preschoolers recall and comprehend discourse information," *Journal of Experimental Child Psychology*, vol. 172, pp. 168–188, Aug. 2018, doi: 10.1016/j.jecp.2018.02.004.

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

[6] 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*, vol. 57, no. 3, pp. 396–414, Oct. 2007, doi: 10.1016/j.jml.2007.06.005.

[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," *Journal of Cognitive Neuroscience*, vol. 28, no. 9, pp. 1255–1269, Sep. 2016, doi: 10.1162/jocn_a_00963.

[8] H. R. Bosker and D. Peeters, "Beat gestures influence which speech sounds you hear," *Proceedings of the Royal Society B: Biological Sciences*, vol. 288, no. 1943, p. 20202419, Jan. 2021, doi: 10.1098/rspb.2020.2419.

[9] R. Bujok, A. 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," PsyArXiv, preprint, May 2022. doi: 10.31234/osf.io/y9jck.

[10] K. S. Apfelbaum, E. Kutlu, B. McMurray, and E. C. Kapnoula, "Don't force it! Gradient speech categorization calls for continuous categorization tasks," *The Journal of the Acoustical Society of America*, vol. 152, no. 6, pp. 3728–3745, Dec. 2022, doi: 10.1121/10.0015201.

[11] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder," *Behav Res*, vol. 52, no. 1, pp. 388–407, Feb. 2020, doi: 10.3758/s13428-019-01237-x.

[12] A. E. Milne *et al.*, "An online headphone screening test based on dichotic pitch," *Behav Res*, vol. 53, no. 4, pp. 1551–1562, Aug. 2021, doi: 10.3758/s13428-020-01514-0.

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

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

[15] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *J. Stat. Soft.*, vol. 82, no. 13, 2017, doi: 10.18637/jss.v082.i13.

[16] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, Third. Thousand Oaks CA: Sage, 2019. [Online]. Available: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[17] P. L. Rohrer, E. Delais-Roussarie, and P. Prieto, "Beat Gestures for Comprehension and Recall: Differential Effects of Language Learners and Native Listeners," *Front. Psychol.*, vol. 11, p. 575929, Oct. 2020, doi: 10.3389/fpsyg.2020.575929.

[18] E. I. Levantinou and C. Navarretta, "An investigation of the effect of beat and iconic gestures on memory recall in L2 speakers," 2015.