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

Test-retest reliability of audiovisual lexical stress perception after >1.5 years

Floris Cos[1], *Ronny Bujok*[2], *Hans Rutger Bosker*[1,2]

[1]Donders Institute for Brain, Cognition and Behavior, Radboud University, Nijmegen, The Netherlands
[2]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
Floris.Cos@ru.nl, Ronny.Bujok@mpi.nl, HansRutger.Bosker@donders.ru.nl

Abstract

In natural communication, we typically both see and hear our conversation partner. Speech comprehension thus requires the integration of auditory and visual information from the speech signal. This is for instance evidenced by the Manual McGurk effect, where the perception of lexical stress is biased towards the syllable that has a beat gesture aligned to it. However, there is considerable individual variation in how heavily gestural timing is weighed as a cue to stress. To assess within-individual consistency, this study investigated the test-retest reliability of the Manual McGurk effect. We reran an earlier Manual McGurk experiment with the same participants, over 1.5 years later. At the group level, we successfully replicated the Manual McGurk effect with a similar effect size. However, a correlation of the by-participant effect sizes in the two identical experiments indicated that there was only a weak correlation between both tests, suggesting that the weighing of gestural information in the perception of lexical stress is stable at the group level, but less so in individuals. Findings are discussed in comparison to other measures of audiovisual integration in speech perception.
Index Terms: Audiovisual integration, beat gestures, lexical stress, test-retest reliability

1. Introduction

In everyday conversations, people typically communicate with others in a face-to-face context. Listeners thus receive visual information in addition to the acoustic signal, rendering natural speech multimodal. The successful perception of speech from these different modalities requires the online integration of both information streams; auditory and visual information complement each other to optimize perception. Indeed, the presence of the visual aspect of speech has a notable beneficial effect on speech intelligibility in noise when compared to the sound signal in isolation (e.g., [1, 2]).

On the other hand, visual information in the speech signal can also change what is being heard. For example, the classic McGurk effect [3], whereby lip movements change perception of auditory speech, is one of the most well-known effects in the field of audiovisual speech comprehension. While "some form of the effect" [4, p. 3] has frequently been found, different studies have reported susceptibility rates for the effect that vary wildly across studies and participants. For instance, Brown et al. [5] describe a McGurk study where some of the 175 participants never reported a McGurk response in the entire experiment, while others exclusively reported perceiving fused syllables. As such, McGurk susceptibilities ranged from 0% to 100%. Indeed, across studies, this is hardly an exception [6]. While failing to report a McGurk response does not necessarily mean that audiovisual integration does not occur [4], it is clear that there is a large variability between individuals in how cues from different modalities are weighed in the ultimate percept.

Nevertheless, there is some evidence that within a listener, susceptibility remains fairly constant. For the McGurk effect, strong test-retest correlations have been reported for a two month interval (r = .77, N = 58; [7]) and even based on sessions that were a whole year apart (r = .91, N = 40; [8]). Despite considerable interpersonal variability in audiovisual integration, these studies thus imply a certain reliability in multisensory cue weighting within a participant over extended time spans.

In addition to articulatory movements in the face, there are more visual cues that can facilitate and influence speech perception. Indeed, a considerable amount of information is conveyed through the use of manual gestures [9, 10]. For instance, beat gestures, brief up-and-down movements of the hand, occur frequently in natural spoken language [10]. They are often used to emphasize words and syllables through a close coupling to the acoustic signal, for example to highlight lexical stress [11]. As such, seeing visual beat gestures facilitates the perception of acoustic prominence [12, 13].

Moreover, beat gestures have been shown to influence the perception of lexical stress in Dutch; the same acoustic stimulus can be perceived as *CONtent* when combined with a beat gesture aligned to the first syllable, but as *conTENT* when the beat falls on the second syllable [14]. This effect is known as the Manual McGurk effect. In fact, Bujok et al. [15] recently used video editing techniques to fully cross articulatory cues to stress on the face (e.g., a head pronouncing either *CONtent* or *conTENT*) with various beat gesture alignments (e.g., a hand producing a beat gesture on the first or second syllable). The experiment entailed a 2AFC task with audiovisual stimuli of Dutch minimal stress pairs, where participants decided whether they perceived lexical stress on the first, or the second syllable. The results indicated that beat gestures indeed influenced the perception of lexical stress, whereas no evidence was found for any influence of different articulatory cues to stress on the face. The authors therefore concluded that beat gestures provide a much more informative cue to the placement of lexical stress than lip movements do.

The present study aimed to replicate the Manual McGurk effect [15], focusing specifically on the reliability with which individuals weigh auditory and visual cues to stress. We assessed the test-retest reliability of the Manual McGurk effect by repeating the experiment performed by Bujok et al. [15], testing a subset of the same participants, which enabled us to make a direct comparison between sessions.

2. Methods

2.1. Participants

Of the original 99 participants in Bujok et al. [15] (henceforth: "test"), 43 (35 female, 8 male; M_{Age} = 25.5, SD_{Age} = 4.1,

10.21437/SpeechProsody.2024-176

range = 19-38) took part in the present experiment (henceforth: "retest"). Of these 43 retest participants, 24 originally participated in the online version of the test; the other 19 had been tested in the lab. Because Bujok et al. [15] did not find any qualitative differences between both versions, all participants completed the retest online. On average, the interval between participation in the original test and the retest was 597.8 days (*SD* = 68.9, range = 398-700). All participants gave informed consent before the experiment started, as approved by the Ethics Committee of the Social Sciences department of Radboud University (project code: ECSW-2019-019). Participants received a compensation of €8 for their participation.

2.2. Materials

We ran an identical replication study of Bujok et al. [15], thus reusing all materials and procedures. For additional details regarding the methods, please see the preprint by Bujok et al. [15]. In summary, the audiovisual stimuli were based on seven pairs of Dutch disyllabic words that differed only in meaning and the position of lexical stress (e.g., *VOORnaam*, "first name" vs. *voorNAAM*, "respectable"). A male native speaker was filmed producing all words twice: once with, and once without a beat gesture aligned to the word's stressed syllable. Both audio and video aspects of these recordings were manipulated to create the experimental conditions of the experiment.

In terms of acoustics, stimuli involved the two originally produced items of a pair (e.g., *VOORnaam* and *voorNAAM*), as well as a five-step F0 contour continuum ranging from stress on the first syllable (Strong-Weak, SW) to stress on the second syllable (Weak-Strong, WS). Duration and intensity were kept constant at the average of the stressed and unstressed variant of each syllable. These acoustic stimuli were combined with artificially edited videos, fully crossing facial articulatory cues and gestural cues. That is, from the videos without a beat gesture, the head and neck were cut out and pasted onto the video where the speaker did make a beat gesture, thereby ensuring that no traces of the beat movement were visible in the face whatsoever. Audio and video materials were combined, such that the apex of the beat was aligned with the onset of the vowel of the syllable in question. The average duration of a stimulus was 2375 ms. There were a total of 196 audiovisual stimuli (2 face conditions x 2 beat conditions x 7 audio steps x 7 word pairs).

2.3. Procedure and design

Participants received instructions and a personalized login code for the experiment, which was hosted in the Gorilla Experiment Builder [16]. The average screen size was ca. 1466 x 846 pixels, and participants reported an average seating distance of 49.5 cm. Lastly, it was required to complete the experiment using high quality headphones, which was checked using a headphone screening based on Huggins Pitch [17].

All 196 audiovisual stimuli were presented to every participant, and each trial consisted of the following components. First, the two words of the word pair of the upcoming stimulus were shown on either side of the screen for 1500 ms. Stress was marked by capital letters. Next, a fixation cross was presented at the center of the screen for 500 ms, followed by the stimulus. Then, the two words reappeared, and participants were asked to indicate which word they perceived, by pressing the Z or M button. The experiment also included 28 catch trials, to encourage participants to both watch and listen to the stimuli (i.e., not close their eyes). Participants had been instructed to press space on a catch trial, when a large white cross was presented over the speaker's face. Space was available as an answer option in all trials. If the participant failed to respond within 4000 ms, the experiment continued automatically. The main experiment was preceded by a practice block. Afterwards, participants completed a short questionnaire about the experiment.

2.4. Data preparation and analyses

A regression analysis was used at the group level, to check whether the current study would replicate the findings of the original test [15]. Generalized Linear Mixed Models were fitted on a dataset containing the data of the current set of participants, from both the test and the retest, using the *lme4* [18] and *lmerTest* packages [19] in R [20]. We used a logistic linking function to predict the perceived stress location as a categorical numerical dependent variable (i.e., SW, like *VOORnaam*, coded as 1; or WS, like *voorNAAM*, coded as 0). As independent variables, the model contained the factors Audio Step (continuous, z-scored), Face (categorical, deviance coded with SW as 0.5 and WS as -0.5), Beat (categorical, deviance coded with BeatOn1 as 0.5 and BeatOn2 as -0.5) and the interaction of Face and Beat. Furthermore, we included the factor Session (categorical, deviance coded with Test as -0.5 and Retest as 0.5), and interactions between Session and Audio Step, Face and Beat. Random effects included participant and item, and the model also included by-participant and by-item random slopes for the factors Audio Step and Beat.

In addition to a group-level replication of the Manual McGurk effect, we aimed to assess the effect of the Beat cue on the responses, and more importantly, its stability over time. Therefore, we calculated participant-specific effect sizes of the beat predictor, both for the original test [15] as well as the retest, allowing us to compare participant behavior on an individual basis. Effect sizes were calculated by taking the proportions of SW responses for either beat condition (BeatOn1 and BeatOn2), aggregating across Face conditions and Audio Steps. The resulting proportions were transformed into logit space. Finally, the effect size was computed as the difference between these logit-transformed SW proportions of BeatOn1 and BeatOn2 responses. In this case, a positive effect size denotes a higher proportion of SW responses for stimuli with a BeatOn1 cue compared to stimuli with a BeatOn2 cue; conversely, a negative effect size corresponds to an (unexpected) larger proportion of SW responses in the BeatOn2 condition. Effect sizes of both sessions were correlated to assess test-retest reliability.

3. Results

One participant exhibited a particularly large effect size (5.053 in the original test), falling outside 3 SD from the mean effect size (M = 0.524, SD = 0.628). This suggests that they completely ignored any auditory stress cues in the stimuli and strategically focused on beat alignment. Therefore, this outlier was removed from the dataset; all analyses below therefore report tests on the remaining 42 participants. Finally, there were no timeout trials with reaction times exceeding 4000 ms.

3.1. Group-level analysis

Results from the regression analysis replicated the outcome of the original test [15]. The model revealed a significant effect of Audio Step (β = -1.748, SE = 0.245, z = -7.128, $p < .001$), indicating that with the F0 contour becoming more WS-like, the proportion of SW responses decreased. The effect of Beat was significant as well (β = 0.865, SE = 0.134, z = 6.464, $p < .001$),

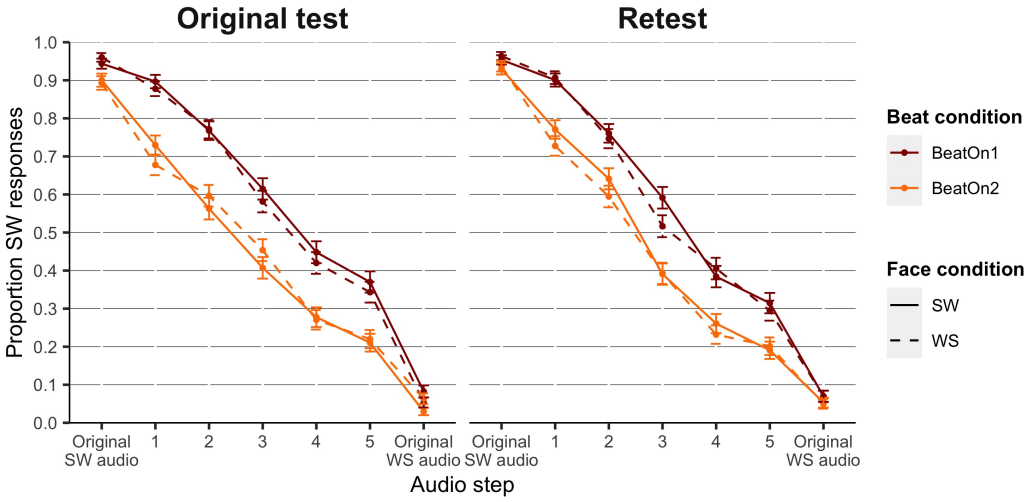

Figure 1: *The proportion of SW responses for each Audio Step. Color denotes the alignment of the beat (first or second syllable) and the line type denotes facial movements (the mouth producing the SW or WS item of the pair). Finally, the two panels show results for the original test on the left (but only those participants who also participated in the retest) and the retest on the right, highlighting the similarity between the two studies.*

revealing that the probability of an SW response increased significantly when the beat gesture was aligned to the first syllable. The Face predictor was not significant (β = 0.075, SE = 0.044, z = 1.734, p = .083), nor was the interaction between the Beat and Face predictors (β = 0.061, SE = 0.087, z = 0.701, p = .484). Furthermore, the Session predictor did not have a significant effect (β = 0.072, SE = 0.044, z = 1.663, p = .096), nor did its interactions with Face (β = -0.071, SE = 0.087, z = -0.815, p = .415) and, importantly, with Beat (β = 0.120, SE = 0.088, z = 1.369, p = .171), suggesting that the effect size of Beat was similar across experiments. Finally, we observed a small yet reliable interaction between Session and Audio Step (β = 0.235, SE = 0.072, z = 3.257, p = .001), indicating a slightly weaker effect of Audio Step in the Retest compared with the original Test. The results of the current participant group's test and retest are visualized in Figure 1.

3.2. Correlation of individual differences

The correlation analysis on the Beat effect sizes between the original test and the retest, more than 1.5 years later, did not show a statistically reliable correlation (r(40) = .279, p = .074); see Figure 2A. However, a stronger correlation may have been obscured by the way the effect sizes were computed. Namely, in calculating the SW-response proportions for the BeatOn1 and BeatOn2 conditions, all Audio Steps and Face conditions were aggregated together. In the Audio Step levels with original (unmanipulated) audio, however, the proportions of SW responses remained close to either 1 or 0 respectively, regardless of the Beat condition (see Figure 1). Since this held for all participants, individual variance was lost by including those Audio Steps with original audio recordings. In an exploratory analysis, we therefore omitted the Audio Step levels with unmanipulated audio from the calculation of our Beat effect sizes, which increased individual variability. The logit-transformed effect sizes from this subset of Audio Steps were then once again correlated for the original test and the retest. This resulted in a small but significant correlation (r(40) = .323, p = .035). The correlation coefficient thus implies that the by-participant effect of Beat is slightly more consistent over time in conditions where the audio cues are more ambiguous. The correlation plot of the subset of Audio Step levels was included in Figure 2B.

4. Discussion

This study aimed to replicate the Manual McGurk effect [14, 15]. More importantly, we set out to assess the test-retest reliability of the Manual McGurk effect. To this end, we reran a 2AFC experiment contrasting the effects of pitch, lip movements and beat gestures on the perception of lexical stress [15]. On the group level, the results of the current study successfully corroborate Bujok et al.'s [15] findings that the alignment of the beat can bias the perception of lexical stress, thereby replicating the Manual McGurk effect. Compared to the original test, over one and a half year prior, the participant group on average showed a similar effect size of susceptibility to the visual cue of beat in determining where lexical stress lies. In addition, like Bujok et al. [15], we found no evidence of an effect of facial movements on the perception of lexical stress. This suggests that, in the context of the complete audiovisual speech signal, facial cues are weighed relatively lightly in the perception of lexical stress – even though facial movements in isolation have been shown to provide enough information to distinguish the location of lexical stress [21, 22].

Retesting a subset of Bujok et al.'s [15] participants enabled us to correlate beat effects from the original and the current study on an individual level. This correlation was not significant when all audio steps were taken together, suggesting that, despite a comparable effect size between sessions at the group level, there exists considerable within-participant variability in the influence of beats on lexical stress perception. On the other hand, we did find a small but significant test-retest correlation when analyzing trials where the audio was manipulated to be more ambiguous. Indeed, Bujok et al. [15] also report that the effect of Beat is larger in acoustically ambiguous contexts, implying that the beat cue is weighed more heavily if acoustic cues

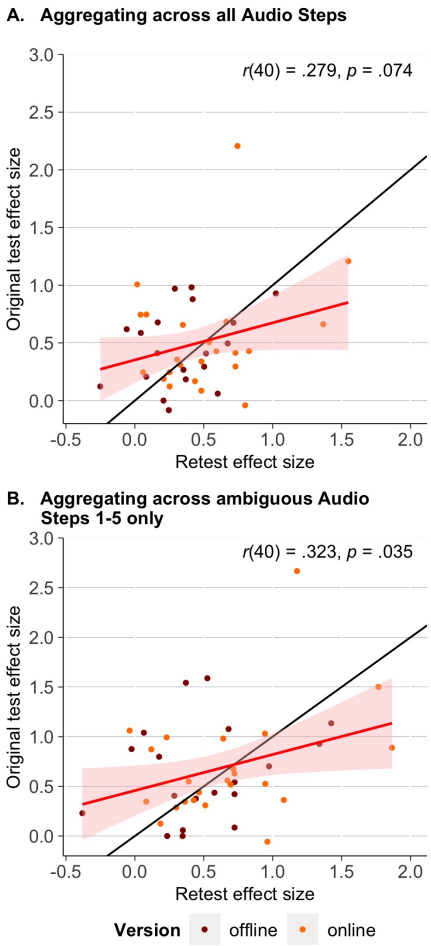

Figure 2: *Effect size correlations. The original test effect sizes are on the y-axis, the retest effect sizes are on the x-axis. Point color indicates whether the participant originally participated in the online or in-house (offline) version; note that in this study, all participants were tested online. The black line is an ab-line, or a perfect correlation. The red lines are linear regression lines detailing the relationship between the test and retest effect sizes.*

are less reliable. Taken together, the results suggest that there is variability in how a listener weighs pitch and gestural cues in the perception of lexical stress, but that the beat cue is processed with greater consistency in contexts where auditory cues are less reliable.

Compared to the classical McGurk effect, the test-retest reliability of the Manual McGurk effect appears to be quite low; we found a correlation coefficient of $r = .32$, with $N = 42$ and a test interval of 1.5 years, compared to $r = .77$ ($N = 58$, with an average test interval of 61 days [7]) or $r = .91$ ($N = 40$ with a test interval of one year [8]).

The discrepancy between both paradigms may be explained by the differences in the signal on the one hand, and the ultimate perceptual goal on the other: in the classic McGurk effect, segmental information and lip movements are perceived and integrated in order to discern (often nonsense) syllables. In addition to these cues, the signal used to elicit the Manual McGurk effect also contains suprasegmental acoustic cues and gestural movements. Furthermore, the ultimate perceptual goal of words, including associated semantics, is arguably more complex and linguistically abstract. Therefore, it is likely that the processes underlying the audiovisual perception of lexical stress in words include more mechanisms than is required for the audiovisual perception of mere syllables. As such, the low correlation observed here may be due to variability in any combination of perceptual or cognitive mechanisms that are recruited during the Manual, but not the classical McGurk effect. Another difference between both paradigms is that the classic McGurk paradigm typically makes use of incongruent stimuli, whereas the Manual McGurk paradigm uses perceptually ambiguous audio (i.e., phonetic continua). Hence, the mechanisms that process incongruent and conflicting multisensory stimuli may be distinct from the ones processing (arguably more naturalistic) ambiguous videos [4, 6, 23].

Several studies support this view of variability in recruited mechanisms. For example, no correlations were found between classical McGurk susceptibility and audiovisual benefit [24, 25]. Indeed, even for non-linguistic audiovisual integration tasks, Wilbiks et al. [25] report no reliable between-test correlations for tasks using audiovisually congruent or incongruent stimuli. The authors therefore conclude that it is likely that different audiovisual tasks engage with different constructs underlying the perception and integration processes, and recommend caution with respect to comparing different tasks that all claim to measure audiovisual integration [25] (see also [26]).

Finally, in many studies of individual variation in audiovisual speech perception, it is unknown to what extent the differences in audiovisual effects genuinely have their roots in cognitive integration processes, rather than unimodal perception skills. For the classical McGurk effect, for example, it has been found that susceptibility to the effect is influenced to some extent by unimodal lipreading skill [5, 7]. Future research into the Manual McGurk effect may therefore benefit from indexing both visual and auditory perceptual acuities of participants, in addition to cognitive capabilities such as visual working memory. Together, these measures may provide an additional insight in the factors that are crucial for weighing and integrating multimodal information to ultimately perceive lexical stress.

5. Conclusions

In this rerun of the experiment by Bujok et al. [15], testing a subset of the same participants, we successfully replicated the Manual McGurk effect, where the perception of lexical stress is biased towards a syllable that has a beat gesture aligned to it. At the group level, we observed a similar effect size compared to more than 1.5 years earlier. We found a small yet reliable correlation between participants' test and retest effect sizes but only when specifically analyzing trials with ambiguous audio, where individual differences were the most pronounced. Future research may therefore focus on identifying sources of this variability, be it based in unimodal perception or cognitive integration mechanisms.

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] W. H. Sumby and I. Pollack, "Visual Contribution to Speech Intelligibility in Noise," *The Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[2] L. Drijvers and A. Özyürek, "Visual Context Enhanced: The Joint Contribution of Iconic Gestures and Visible Speech to Degraded Speech Comprehension," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 1, pp. 212–222, 2017.

[3] H. Mcgurk and J. Macdonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[4] L. D. Rosenblum, "Audiovisual Speech Perception and the McGurk Effect," in *Oxford Research Encyclopedia of Linguistics*, 2019.

[5] V. A. Brown, M. Hedayati, A. Zanger, S. Mayn, L. Ray, N. Dillman-Hasso, and J. F. Strand, "What accounts for individual differences in susceptibility to the McGurk effect?" *PLOS ONE*, vol. 13, no. 11, pp. 1–20, 2018.

[6] K. J. Van Engen, A. Dey, M. S. Sommers, and J. E. Peelle, "Audiovisual speech perception: Moving beyond McGurka)," *The Journal of the Acoustical Society of America*, vol. 152, no. 6, pp. 3216–3225, 2022.

[7] J. Strand, A. Cooperman, J. Rowe, and A. Simenstad, "Individual Differences in Susceptibility to the McGurk Effect: Links With Lipreading and Detecting Audiovisual Incongruity," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 6, pp. 2322–2331, 2014.

[8] D. Basu Mallick, J. F. Magnotti, and M. S. Beauchamp, "Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type," *Psychonomic Bulletin & Review*, vol. 22, no. 5, pp. 1299–1307, 2015.

[9] J. Holler and S. C. Levinson, "Multimodal Language Processing in Human Communication," *Trends in Cognitive Sciences*, vol. 23, no. 8, pp. 639–652, 2019.

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

[11] T. Leonard and F. Cummins, "The temporal relation between beat gestures and speech," *Language and Cognitive Processes*, vol. 26, no. 10, pp. 1457–1471, 2011.

[12] 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, 2007.

[13] E. Biau and S. Soto-Faraco, "Beat gestures modulate auditory integration in speech perception," *Brain and Language*, vol. 124, no. 2, pp. 143–152, 2013.

[14] H. R. Bosker and D. Peeters, "Beat gestures influence which speechsounds you hear," *Proceedings of the Royal Society B*, vol. 288, no. 1943, p. 20202419, 2021.

[15] 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," May 2022. [Online]. Available: https://doi.org/10.31234/osf.io/y9jck

[16] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder," *Behavior Research Methods*, vol. 52, no. 1, pp. 388–407, 2020.

[17] A. E. Milne, R. Bianco, K. C. Poole, S. Zhao, A. J. Oxenham, A. J. Billig, and M. Chait, "An online headphone screening test based on dichotic pitch," *Behavior Research Methods*, vol. 53, no. 4, pp. 1551–1562, 2021.

[18] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[19] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[21] R. Scarborough, P. Keating, S. L. Mattys, T. Cho, and A. Alwan, "Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English," *Language and Speech*, vol. 52, no. 2-3, pp. 135–175, 2009.

[22] A. Jesse and J. M. McQueen, "Suprasegmental Lexical Stress Cues in Visual Speech can Guide Spoken-Word Recognition," *Quarterly Journal of Experimental Psychology*, vol. 67, no. 4, pp. 793–808, 2014.

[23] A. Alsius, M. Paré, and K. G. Munhall, "Forty Years After Hearing Lips and Seeing Voices: the McGurk Effect Revisited," *Multisensory Research*, vol. 31, no. 1-2, pp. 111–144, 2018.

[24] K. J. Van Engen, Z. Xie, and B. Chandrasekaran, "Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect," *Attention, Perception, & Psychophysics*, vol. 79, no. 2, pp. 396–403, 2017.

[25] J. M. P. Wilbiks, V. A. Brown, and J. F. Strand, "Speech and non-speech measures of audiovisual integration are not correlated," *Attention, Perception, & Psychophysics*, vol. 84, no. 6, pp. 1809–1819, 2022.

[26] J. F. Strand, L. Ray, N. H. Dillman-Hasso, J. Villanueva, and V. A. Brown, "Understanding Speech amid the Jingle and Jangle: Recommendations for Improving Measurement Practices in Listening Effort Research," *Auditory Perception & Cognition*, vol. 3, no. 4, pp. 169–188, 2020.