Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

Visible lexical stress cues on the face do not influence audiovisual speech perception

Ronny Bujok[1], *Antje Meyer*[1], *Hans Rutger Bosker*[1,2]

[1]Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH Nijmegen, The Netherlands
[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

Ronny.Bujok@mpi.nl, Antje.Meyer@mpi.nl, HansRutger.Bosker@mpi.nl

Abstract

Producing lexical stress leads to visible changes on the face, such as longer duration and greater size of the opening of the mouth. Research suggests that these visual cues alone can inform participants about which syllable carries stress (i.e., lip-reading silent videos). This study aims to determine the influence of visual articulatory cues on lexical stress perception in more naturalistic *audiovisual settings*.

Participants were presented with seven disyllabic, Dutch minimal stress pairs (e.g., *VOORnaam* [first name] & *voorNAAM* [respectable]) in audio-only (phonetic lexical stress continua without video), video-only (lip-reading silent videos), and audiovisual trials (e.g., phonetic lexical stress continua with video of talker saying *VOORnaam* or *voorNAAM*).

Categorization data from video-only trials revealed that participants could distinguish the minimal pairs above chance from seeing the silent videos alone. However, responses in the audiovisual condition did not differ from the audio-only condition.

We thus conclude that visual lexical stress information on the face, while clearly perceivable, does not play a major role in audiovisual speech perception. This study demonstrates that clear unimodal effects do not always generalize to more naturalistic multimodal communication, advocating that speech prosody is best considered in multimodal settings.

Index Terms: prosody, lexical stress, audiovisual speech, articulatory cues, spoken-word recognition

1. Introduction

Spoken language is most commonly used face-to-face and is thus inherently multimodal. Beside the auditory signal, visual information contributes to speech perception as well [1, 2, 3]. The effect of visual information is well demonstrated by the McGurk effect [4], where participants who hear the sound /ba/, while seeing a video of a speaker saying /ga/, perceive an illusory "da". Moreover, visual information improves speech perception in noise [5], and specifically visual information on the face (e.g., articulatory movements of the lips, mouth and jaw) has been found to facilitate speech perception of speech segments (e.g., vowels and consonants) [6, 7].

However, speech consists of more than just segments. Prosody, as cued by suprasegmental information, is also an integral part of human language. For example, speech rate, lexical tone and lexical stress guide spoken word recognition [8, 9, 10]. For instance, lexical stress is lexically contrastive in many languages (e.g., English, Dutch, Spanish) and thus distinguishes segmentally identical words such as Dutch $VOORnaam$ [first name] vs. $voorNAAM$ [respectable]. Moreover, lexical stress drives online word recognition and disambiguation, even for non-minimal pairs, like $OCtopus$ and $okTOber$ [11, 12, 13]. Also, with the inclusion of lexical stress information, words reach the point of uniqueness much earlier. That is, in Dutch, without consideration of lexical stress, words become unique on average after 80% of the phonemes. With lexical stress, the uniqueness point is reached after 66% of the phonemes [14]. Furthermore, recent evidence suggests that lexical stress, like many segmental contrasts, is even represented at an abstract pre-lexical level [15].

Most studies on the influence of lexical stress on speech perception have only focused on the auditory modality, presumably because the suprasegmental correlates of lexical stress (e.g., fundamental frequency [F0], intensity and duration) are less visibly salient than visual correlates of segmental speech. Still, producing lexical stress leads to small but visible changes in articulation. Scarborough et al. [16] video-recorded native speakers of English producing words that differed in lexical stress (e.g., $SUBject$ vs. $subJECT$). They analyzed various measures of facial movement, such as maximum lip opening and jaw opening displacement, and found that they were generally larger in stressed syllables. They then presented the videos without any audio to participants in a 2-alternative forced choice task (2AFC) and observed that the participants could determine the position of primary lexical stress with 62.2% accuracy. Note that in English lexical stress is cued by both suprasegmental and segmental cues, such as vowel reduction. Hence, it remains unclear whether the visual articulatory cues to stress in English are driven by suprasegmental or segmental (i.e., vowel reduction) differences.

In contrast, in Dutch, segmental changes only play a minimal role in the production of lexical stress. Therefore, Jesse and McQueen [17] tested visual perception of lexical stress in Dutch to determine the visibility of suprasegmental cues. They video-recorded a Dutch native speaker producing Dutch words that are segmentally identical in the first two syllables and differ in lexical stress (e.g., $OCtopus$ vs. $okTOber$). These two first syllables of the words were then presented to participants who could determine the position of lexical stress in a 2AFC task with approximately 70% accuracy. Taken together, these two studies demonstrate that the subtle visual articulatory cues of lexical stress are perceivable from video-only stimuli alone.

However, it remains unknown whether these visual articulatory cues play a role in more naturalistic $audiovisual$ perception. That is, Scarborough et al. [16], and Jesse and McQueen [17] only tested the perception of these cues under video-only conditions. However, whether participants actually use these perceptually visible cues when auditory cues are also present is not clear. Moreover, both studies used visual stimuli containing only the talker's face, presenting them at a scale that does

10.21437/SpeechProsody.2022-53

not reflect naturalistic conversations. Scarborough et al. [16] for example presented the talker's face at 90% life-size, with participants seated 50 cm away from the screen. As a result, the visual angle was presumably greater than encountered in actual face-to-face conversations [18]. From the measurements reported by Jesse and McQueen [17] this is also the case in their study. This could mean that the visual cues in their experiments were more salient compared to everyday conversations.

The present study assessed whether visual articulatory cues to lexical stress influence lexical stress perception in arguably more naturalistic *audiovisual* settings. Dutch participants were presented with phonetic continua of disyllabic minimal stress pairs (e.g., $VOORnaam$ and $voorNAAM$) combined with a video of a talker producing the word with stress on the first or second syllable. In a 2AFC task they had to determine the placement of lexical stress. If visual cues to stress influence audiovisual perception, participants should be more likely to report hearing stress on the first syllable if the talker in the video produced stress on the first syllable (and vice versa). However, if these visual cues are only used in unimodal video-only settings [16, 17] but not in audiovisual settings, we should find no difference between the two audiovisual conditions (visual stress on first vs. second syllable) and audio-only presentation. Finally, video-only trials were included to conceptually replicate previous video-only experiments [16, 17].

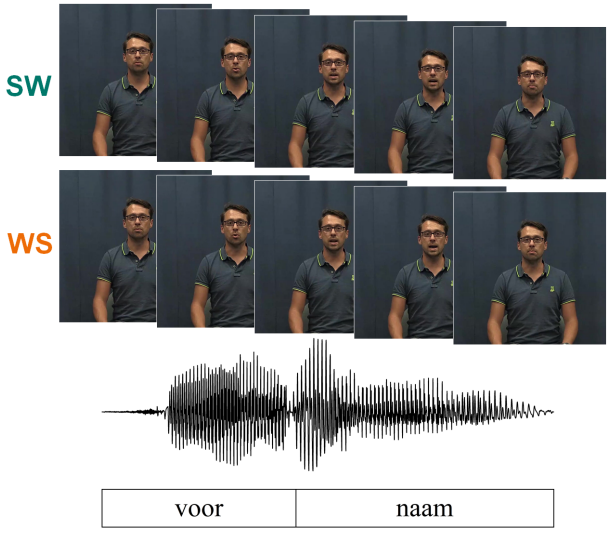

Figure 1: *Videos of the speaker producing either a 'strong-weak' word (SW, e.g., $VOORnaam$) or 'weak-strong' word (WS, e.g., $voorNAAM$) were presented with the audio from the lexical stress continuum. Note that differences in visual articulatory cues were subtle.*

2. Methods

2.1. Power analysis

We estimated statistical power by means of Monte Carlo simulations (N=1000) using Generalized Linear Mixed Models [19], setting the overall perceptual difference between videos with lexical stress on the first syllable ('strong-weak' [SW]) and videos with lexical stress on the second syllable ('weak-strong' [WS]) to 5%. With this effect size, we achieved a power of 0.81 with 48 participants.

2.2. Participants

Forty-eight native speakers of Dutch (37 female, 11 male, median age = 25, range = 19 - 39) were recruited through the Max Planck Institute for Psycholinguistics participant pool. Participants gave informed consent as approved by the Ethics Committee of the Social Sciences department of Radboud University (project code: ECSW-2019-019). None of the participants reported any hearing or language deficit and all had normal or corrected-to-normal vision. Participants received €8 compensation for participation.

2.3. Materials

Materials consisted of seven disyllabic, segmentally identical minimal pairs of frequent Dutch words (see supplementary material at https://osf.io/um7ph). The pairs only differed in the position of lexical stress (e.g., $VOORnaam$ [first name] vs. $voorNAAM$ [respectable]). High-definition video recordings of a male native speaker of Dutch producing all 14 words were made. The speaker was recorded in front of a natural background in a sitting position with everything above the hip framed. He was instructed to produce the words naturally. Videos were cropped to 620 x 620 pixel squares showing the speaker's face and torso (see Figure 1) and exported as avi files. The audio sampling rate was 48 kHz and the video sampling rate was 50 Hz.

Lexical stress in Dutch is primarily cued by three suprasegmental cues: fundamental frequency (F0), duration, and intensity [20]. F0 is the biggest contributor in words that align with phrasal accent and in isolated words [20]. Therefore, we created a lexical stress continuum for each minimal pair (ranging from SW to WS) by manipulating F0, while keeping duration and intensity constant. The SW and WS audio were extracted from the video recordings and then manipulated. We determined the average duration of the first and second syllable within each item pair and set the values for the syllables in both words to these average values, making intensity and duration identical across words and thus ambiguous with regards to lexical stress. The F0 contours of both words were linearly interpolated in eleven steps (step 1 and 11 being the original SW and WS contours) and then applied to the SW recording (with ambiguous duration and intensity) using PSOLA in Praat [21](see Figure 2). For one item pair ($SERvisch$ vs. $serVIES$) the F0 contours were applied to the WS recording because it resulted in a more natural sounding word.

These manipulated speech tokens (N=77; 7 pairs x 11 steps) were presented to 10 participants in an audio-only pretest. Participants had to categorize the tokens as either SW or WS in a 2AFC task. Based on their categorization data, we selected five tokens for each pair that sampled a perceptually defined continuum from SW (>80% SW responses) to WS (<20% WS responses) with 3 more ambiguous steps in the middle. Additionally, the original recordings (i.e., with unmanipulated F0, duration, and intensity) were used as the extreme ends of the continua, resulting in a total 7-step perceptual lexical stress continuum for each pair.

2.4. Design and Procedure

The experiment had three conditions: audio-only (A), video-only (V), and audiovisual (AV). For the A condition, we presented the manipulated F0 continua with a still image of the speaker with a neutral facial expression and a closed mouth.

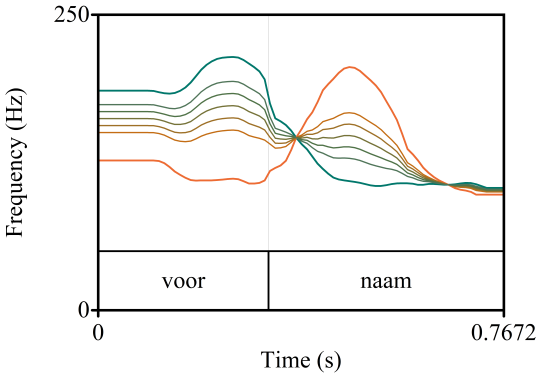

Figure 2: *Illustration of the F0 manipulation, ranging from clear SW (green) to clear WS (orange) with 5 ambiguous steps in between. These steps were selected after the pretest to reflect a perceptual stress continuum.*

In the V condition we presented muted videos of the speaker producing either the SW or WS word. Crucially, in the AV condition we combined each video with the entire lexical stress continuum, aligning the audio and video at the second syllable onset. This minimized synchrony issues on either syllable. The average asynchrony for our stimuli was 40 ms at word onset and 35 ms at word offset which were deemed acceptable since asynchronies in speech of up to 150 ms are perceived as synchronous [22]. All stimuli were cut such that there was approximately a 500 ms silent interval before word onset and after word offset. The average duration of the stimuli was 1875 ms. Taken together this resulted in 161 items, of which 49 A items (7 items x 7 steps), 14 V items (7 items x 2 videos), and 98 AV items (7 items x 7 steps x 2 videos).

Participants were tested individually in a sound-attenuating booth. The experiment was run in Presentation® software (Version 18.0, Neurobehavioral Systems, Inc., Berkeley, CA) and presented on a 24" full HD screen with a refresh rate of 144 Hz. AV stimuli appeared in the center of the screen as 1080 x 1080 pixel displays on a white background. Audio was presented through high quality headphones (beyerdynamic DT 770 PRO 32 Ohm) at a comfortable volume. Participants were seated at a distance of approximately 60 cm from the screen. The videos were presented at full screen making the speaker's head 5.7 cm wide and 7.5 cm tall. From the distance to the screen (d) and size of the head (h) we could calculate a visual angle (θ) indicating how big the head appeared to the participants ($tan(\theta/2) = (h/2)/d$). The visual angle of of the head was 7.15°, which is equivalent to a conversation with someone at 1.93 m distance, assuming an average male head height of 24.1 cm from the chin to the top of the head [23]. This falls in the range of interpersonal interactions [18] and is considered a comfortable interaction distance [24].

All 161 unique items, from the three conditions (14 AV, 7 A and 2 V for each item), were presented once in a fully randomized order. This meant that A, V and AV were intermixed. Halfway through, participants had a chance to take a break. The task was to decide from two words presented on screen, what the speaker was saying (2AFC). Before the task participants received four practice trials to become familiar with the materials and the task. Four stimuli sampled from all three conditions were chosen as practice trials, using only original unmanipulated audio (A-SW, V-WS, AV-SW AV-WS).

Participants were instructed to look at the screen at all times. They were explicitly told beforehand that they would see videos with and without audio, and audio with a still image. A trial began with the two response options (e.g., $VOORnaam$ vs. $voorNAAM$) presented on either side of the screen (Arial, font size 16) for 1500 ms. Lexical stress was indicated by capital letters. The sides on which SW and WS response options were presented were counterbalanced across participants. Then a fixation cross was displayed for 500 ms, and then the stimulus. The fixation cross was positioned at the center of the speaker's mouth, which appeared 120 pixels above the center of the screen. After the stimulus, the response options appeared again for a maximum of 4000 ms. Participants responded by pressing the "Z" and "M" button on the keyboard, corresponding to the left and right word on screen. After a response, the selected word was highlighted by displaying it in a bigger font size (20) for 500 ms. After this, a 500 ms blank screen was presented before the next trial began automatically.

3. Results

Data were analyzed with Generalized Linear Mixed Models using the lme4 library [25] in R (R Core Team, 2021). Two different models were created, one comparing the AV condition to the A condition, and one for the V condition. In both models, participants' categorization responses, that is lexical stress on the first (SW coded as 1, e.g., $VOORnaam$) or second syllable (WS coded as 0, e.g., $voorNAAM$), were the dependent variable.

We ran the video-only model to assess whether participants could reliably use the visual cues to lexical stress in video-only trials, aiming to replicate findings by Jesse and McQueen [17]. This video-only model included Video (categorical predictor, deviance coded SW as 0.5 and WS as -0.5) as a predictor to test the perceptual differences between SW and WS video. This video-only model revealed a significant effect of Video (β= 1.143, SE= 0.217, z= 5.255, $p < 0.001$), indicating that articulatory cues to lexical stress were visibly different between SW and WS videos. This is illustrated in the right panel of Figure 3, which seems to suggest that this Video effect was primarily driven by visual stress cues in the WS videos.

Next we compared the audiovisual conditions to the audio-only condition. In this model, we included Continuum Step (continuous; z-scored) and Condition (categorical predictor with three levels; SW, WS, A mapped on the intercept). The model only showed a significant effect of Continuum Step (β= -1.915, SE= 0.152, z= -12.621, $p < 0.001$) meaning that with increasing steps on the continuum the proportion of SW responses decreased. However, neither SW videos (β= 0.007, SE= 0.085, z= 0.079, p= 0.937) nor WS videos (β= -0.07, SE= 0.126, z= -0.555, p= 0.579) influenced the responses when compared to the A condition (intercept). The responses on AV trials were similar to A trials, which is demonstrated by the overlapping lines in the left panel in Figure 3, suggesting no effect of informative articulatory cues to lexical stress. Extending this model with an interaction between Continuum Step and Condition did not significantly improve the model fit to the data, as revealed by log-likelihood model comparison.

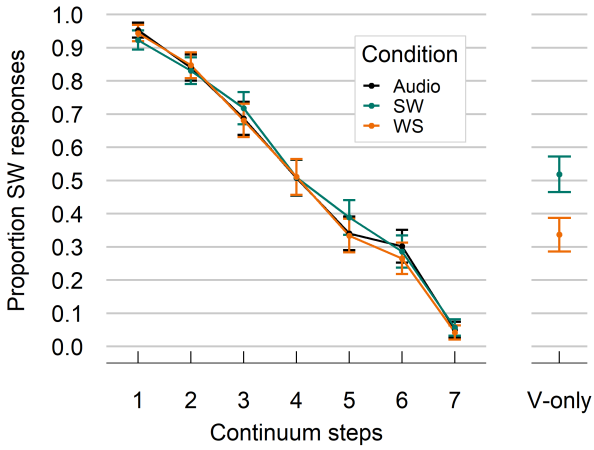

Figure 3: *Results as quantified by the proportion of SW responses to any given condition and step (error bars indicate 95% confidence intervals). Right panel: in video-only trials, participants could perceptually distinguish between SW and WS videos (mostly driven by visual stress cues in WS videos). Left panel: Nonetheless, these visual cues did not bias the categorization responses in audiovisual conditions, as illustrated by overlapping AV and A conditions. This suggests that visual articulatory cues do not play a strong role in audiovisual lexical stress perception.*

4. Discussion

Results from the V trials, conceptually replicated the findings by Jesse and McQueen [17] and extended them by demonstrating a similar video-only effect in an arguably more difficult task. That is, unlike Jesse and McQueen [17] who presented only V trials, we had V trials intermixed with AV and A trials. This intermixed design meant that participants had to switch between modalities during the task. Such modality-switching has been found to be very costly [26]. Nonetheless, participants could still differentiate between SW and WS stress patterns on average from silent videos. Moreover, they could do so even when the videos were presented at a more realistic size for face-to-face conversations. However, it appeared that this effect was largely driven by the categorization of WS videos. This could be indicative of clearer visual stress cues in our WS videos (compared to our SW videos).

Either way, we failed to find any video effect in the AV condition. Visual information on the face did not affect audiovisual perception of lexical stress. This null result cannot be explained by any visual properties of the articulatory cues in the videos themselves, such as low saliency, since we did find a video effect in the V condition. This indicates that the mere presence of an auditory signal reduced the perceptual weight of the visual information. We suggest two possible explanations. Perhaps the processing of auditory information causes an automatic downregulation of attentional demands necessary to notice the subtle visual cues. That is, participants perceive the visual information to a lesser degree in an audiovisual context. Alternatively, people might still be able to accurately perceive the visual information but intentionally weigh it less heavily in audiovisual intergration [27]. Some research indeed suggests that audiovisual integration is subject to attentional demands. For example, if visual attention is moved away from a speaker's face by instructing participants to attend to a visual distractor, the McGurk effect is reduced [28]. On the other hand, other studies have found that participants were unable to ignore the visual modality completely, even when they were instructed to do so [29]. This would suggest that it is impossible to ignore the visual modality.

This is in stark contrast to our findings where participants were specifically instructed to look at the screen and yet seemingly did not use the visual information in audiovisual perception. We did not find a reduced effect but rather no effect of visual information in the audiovisual trials. However, it is possible that findings from studies on segmental speech cannot be generalized onto suprasegmental speech perception. Our study suggests that visual information might be less relevant in the perception of suprasegmental properties of speech, which might be related to the subtlety of the visual cues.

Our results thus caution against generalizing results from unimodal studies to multimodal processing. Although, lexical stress is visible on the face, as evidenced by performance on silent videos, it appears to have little influence in a more naturalistic audiovisual setting. Nevertheless, our study does not claim that visual information to lexical stress is never used in audiovisual perception. Under different circumstances, for example in noise, the auditory cues could be less reliable and thus the visual cues could be of more relevance. Moreover, the saliency of articulatory cues could vary depending on the speaker or on the setting the speech is produced in (e.g., being more exaggerated in Lombard Speech) [30]. Lastly, other (i.e., non-articulatory) visual cues could affect audiovisual processing differently. In fact, manual beat gestures have been found to influence audiovisual lexical stress perception [31].

To conclude, at this point our findings suggest that articulatory cues to lexical stress are not used in audiovisual word recognition despite being visible in the visual modality alone. They demonstrate that unimodal effects do not always hold up in multimodal settings. This should be kept in mind when considering the implication of unimodal effects seen in the lab for language use in everyday life, namely in multimodal communication.

5. 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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364661319301299

[2] P. Perniss, "Why We Should Study Multimodal Language," *Frontiers in Psychology*, vol. 9, 2018, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01109/full

[3] L. D. Rosenblum, "Speech Perception as a Multimodal Phenomenon," *Current Directions in Psychological Science*, vol. 17, no. 6, pp. 405–409, Dec. 2008, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1111/j.1467-8721.2008.00615.x

[4] H. Mcgurk and J. Macdonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, Dec. 1976, number: 5588 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/264746a0

[5] 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, Mar. 1954, publisher: Acoustical Society of America. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.1907309

[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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0749596X07000708

[7] D. W. Massaro, *Speech Perception By Ear and Eye: A Paradigm for Psychological Inquiry*. Psychology Press, 1987, google-Books-ID: U0nrAgAAQBAJ.

[8] M. Maslowski, A. S. Meyer, and H. R. Bosker, "How the tracking of habitual rate influences speech perception," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 45, no. 1, pp. 128–138, 2019, place: US Publisher: American Psychological Association.

[9] A. Cutler, D. Dahan, and W. van Donselaar, "Prosody in the Comprehension of Spoken Language: A Literature Review," *Language and Speech*, vol. 40, no. 2, pp. 141–201, Apr. 1997, publisher: SAGE Publications Ltd. [Online]. Available: https://doi.org/10.1177/002383099704000203

[10] A. Cutler, "Lexical Stress," in *The Handbook of Speech Perception*. John Wiley & Sons, Apr. 2008, pp. 264 – 289.

[11] A. Cutler and W. Van Donselaar, "Voornaam is not (really) a Homophone: Lexical Prosody and Lexical Access in Dutch," *Language and Speech*, vol. 44, no. 2, pp. 171–195, Jun. 2001, publisher: SAGE Publications Ltd. [Online]. Available: https://doi.org/10.1177/00238309010440020301

[12] D. Pisoni and R. Remez, *The Handbook of Speech Perception*. John Wiley & Sons, Apr. 2008, google-Books-ID: EwY15naRiFgC.

[13] E. Reinisch, A. Jesse, and J. M. McQueen, "Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately," *Quarterly Journal of Experimental Psychology*, vol. 63, no. 4, pp. 772–783, Apr. 2010, publisher: SAGE Publications. [Online]. Available: https://doi.org/10.1080/17470210903104412

[14] V. van Heuven and P. Hagman, "Lexical statistics and spoken word recognition in Dutch," *Linguistics in the Netherlands*, pp. 59–68, 1988.

[15] G. G. A. Severijnen, H. R. Bosker, V. Piai, and J. M. McQueen, "Listeners track talker-specific prosody to deal with talker-variability," *Brain Research*, vol. 1769, p. 147605, Oct. 2021.

[16] 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, Jun. 2009, publisher: SAGE Publications Ltd. [Online]. Available: https://doi.org/10.1177/0023830909103165

[17] 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, Apr. 2014, publisher: SAGE Publications. [Online]. Available: https://doi.org/10.1080/17470218.2013.834371

[18] P. A. M. Ruijten and R. H. Cuijpers, "Do Not Let the Robot Get too Close: Investigating the Shape and Size of Shared Interaction Space for Two People in a Conversation," *Information*, vol. 11, no. 3, p. 147, Mar. 2020, number: 3 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2078-2489/11/3/147

[19] L. Kumle, M. L.-H. Võ, and D. Draschkow, "Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R," *Behavior Research Methods*, May 2021. [Online]. Available: https://doi.org/10.3758/s13428-021-01546-0

[20] T. Rietveld and V. J. v. Heuven, *Algemene Fonetiek (3e geheel herziene druk)*. Bussum : Coutinho, 2009, accepted: 2010-06-24T13:27:54Z. [Online]. Available: https://repository.ubn.ru.nl/handle/2066/79395

[21] P. Boersma, "Praat : doing phonetics by computer," *http://www.praat.org/*, 2006. [Online]. Available: https://ci.nii.ac.jp/naid/10017594077/

[22] N. F. Dixon and L. Spitz, "The Detection of Auditory Visual Desynchrony:," *Perception*, Jun. 2016, publisher: SAGE PublicationsSage UK: London, England. [Online]. Available: https://journals.sagepub.com/doi/10.1068/p090719

[23] J.-H. Lee, S.-J. Shin, and C. Istook, "Analysis of Human Head Shapes in the United States," *International journal of human ecology*, vol. 7, Jan. 2006.

[24] E. Sundstrom and I. Altman, "Interpersonal relationships and personal space: Research review and theoretical model," *Human Ecology*, vol. 4, no. 1, pp. 47–67, Jan. 1976, company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 1 Publisher: Kluwer Academic Publishers-Plenum Publishers. [Online]. Available: https://link.springer.com/article/10.1007/BF01531456

[25] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using **lme4**," *Journal of Statistical Software*, vol. 67, no. 1, 2015. [Online]. Available: http://www.jstatsoft.org/v67/i01/

[26] R. Sandhu and B. J. Dyson, "Re-evaluating visual and auditory dominance through modality switching costs and congruency analyses," *Acta Psychologica*, vol. 140, no. 2, pp. 111–118, Jun. 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0001691812000571

[27] J. L. Mozolic, C. E. Hugenschmidt, A. M. Peiffer, and P. J. Laurienti, "Modality-specific selective attention attenuates multisensory integration," *Experimental Brain Research*, vol. 184, no. 1, pp. 39–52, Jan. 2008. [Online]. Available: https://doi.org/10.1007/s00221-007-1080-3

[28] K. Tiippana, T. S. Andersen, and M. Sams, "Visual attention modulates audiovisual speech perception," *European Journal of Cognitive Psychology*, vol. 16, no. 3, pp. 457–472, May 2004. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/09541440340000268

[29] J. N. Buchan and K. G. Munhall, "The Influence of Selective Attention to Auditory and Visual Speech on the Integration of Audiovisual Speech Information," *Perception*, vol. 40, no. 10, pp. 1164–1182, Oct. 2011, publisher: SAGE Publications Ltd STM. [Online]. Available: https://doi.org/10.1068/p6939

[30] M. Garnier, L. Ménard, and B. Alexandre, "Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues?" *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. 1059–1074, Aug. 2018, publisher: Acoustical Society of America. [Online]. Available: https://asa.scitation.org/doi/full/10.1121/1.5051321

[31] 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, publisher: Royal Society. [Online]. Available: https://royalsocietypublishing.org/doi/full/10.1098/rspb.2020.2419