Visual articulatory gestures guide audiovisual speech perception of lexical stress but only in noise

Ivy Mok [1], *Ronny Bujok* [1], *and Hans Rutger Bosker* [1,2]

[1] Max Planck Institute for Psycholinguistics, Nijmegen

[2] Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen

hansrutger.bosker@donders.ru.nl

As an addressee in conversation, not only do we attune to the speech available to us, but we also attend to other aspects of the speaker. That is, we consider how the message is being conveyed auditorily (i.e., speech segments like consonants and vowels, and speech prosody) and visually (i.e., visual articulatory gestures like lip movements, hand gestures, body postures).

Visual articulatory cues can influence speech perception. This is shown for instance in the classic McGurk effect (McGurk and McDonald, 1976). In their study, the authors presented participants with nonsense consonant-vowel syllables (i.e., /ba/ and /ga/). When an auditory /ba/ was presented with a visual /ga/ many indicated that they perceived /da/. The McGurk effect clearly demonstrates that participants use both auditory speech cues and visual articulatory gestures from the face in audiovisual speech perception at the segmental level.

However, less is known about the audiovisual perception of spoken prosody. Visual prosody is understudied, arguably because suprasegmental information like pitch and intensity are less readily perceived from the face. Here we focus on lexical stress in the free-stress language Dutch. The meaning of a word can change depending on the position of the lexical stress, such as the minimal stress pair *CAnon/kaNON* "canon/cannon" (capitals indicate stress). Previous findings confirmed that humans can 'lip-read' lexical stress from facial articulatory gestures alone (Scarborough et al., 2009; Jesse & McQueen, 2014).

Using a similar paradigm, Bujok et al. (2022) video-recorded a talker producing Dutch disyllabic minimal stress pairs and created phonetic lexical stress continua, varying F0, ranging from a clear strong-weak pattern (SW; step 1) to clear weak-strong (WS; step 7). These materials were presented in audio-only (A-only), video-only (V-only), and audiovisual (AV) conditions, pairing the continua with videos of either word.

Participants were presented these materials in an intermixed fashion, each time indicating which of two responses (e.g., CAnon or kaNON) they thought the speaker was saying. The results showed that muted SW videos received a higher proportion of SW responses compared to the muted WS videos in the V-only condition suggesting that visual articulatory cues can be used to make decisions about lexical stress. However, when auditory speech was added in the AV condition, participants did not use visual articulatory cues any more to inform their perceptual decisions about lexical stress despite their availability showing identical word identification responses to the A-only condition.

The finding in Bujok et al. (2022) that participants ignore informative visual *prosody* cues in audiovisual stimuli stands in stark contrast with literature on audiovisual *segmental* perception. For instance, in the McGurk effect, visual articulatory gestures have a strong impact on the audiovisual perception of consonants. However, the literature on phonetic cue-weighting argues that humans can flexibly weigh multimodal signals, depending on their needs and the communicative context. For example, the McGurk effect is larger in speech-in-noise in comparison to clear speech (Stacey et al., 2020). Through an eye-tracking study, Stacey et al. showed that participants fixated on the mouth more on trials where the McGurk effect was perceived with added auditory noise (i.e., white noise). This indicates that presenting speech in noise leads to a greater uptake of visual articulatory cues when the auditory signal is less accessible.

In the present study, we investigated whether visual articulatory gestures on the face can contribute to the audiovisual perception of lexical stress after all by presenting the target speech inside mild to loud background babble. We hypothesized that facial articulatory cues can be weighed more heavily in lexical stress perception when the auditory target word is less accessible (i.e., masked by babble). We used the videos from the audiovisual condition in Bujok et al. (2022) in which a talker produced one of two words from minimal stress pairs, manipulating the F0 contour in 7-step auditory continua between strong-weak (SW; step 1) and weak-strong (WS; step 7). Critically, we mixed in 16-talker babble to mask the target speech, choosing four signal-to-noise ratios (SNRs) ranging from fully intelligible to fully unintelligible: 0, -6, -12, and -18 dB (see Figure 1). In total, 7 items (Dutch minimal pairs) x 7 steps x 2 videos x 4 SNRs were used. In the main experiment, run online on Gorilla, native Dutch participants (*N*=96; recruited through Prolific) performed a two-alternative forced choice (2-AFC) task in which they indicated which member of a minimal stress pair they thought the speaker said.

The results from a generalized linear mixed effect model with predictors Step Continuum (steps 1 through 7), Video Condition (stress on syllable1 [SW] or syllable 2 [WS]), and SNR (0, -6, -12, -18), and their interactions critically showed an interaction between Video Condition and SNR. This interaction indicated that participants did not use the visual articulatory cues when the speech was intelligible (SNR = 0 dB; two lines overlapping in bottom-right panel), replicating the clear speech findings from Bujok et al. (2022). Only as the babble intensity increased did people up-weigh the visual cues, gradually showing a separation of the two video conditions as SNR decreased in Figure 2.

To conclude, we find that listeners flexibly weigh visual and auditory cues to spoken prosody (here: lexical stress). When the speech is intelligible, visual prosody is ignored. However, when the auditory speech is less accessible in challenging listening conditions, listeners upweight the contribution of visual articulatory gestures to lexical stress. Thus, our results emphasize the remarkable flexibility in multimodal cue-weighting in audiovisual speech perception.

Index Terms: McGurk effect, lexical stress, visual prosody, visual articulatory gestures, speech in noise

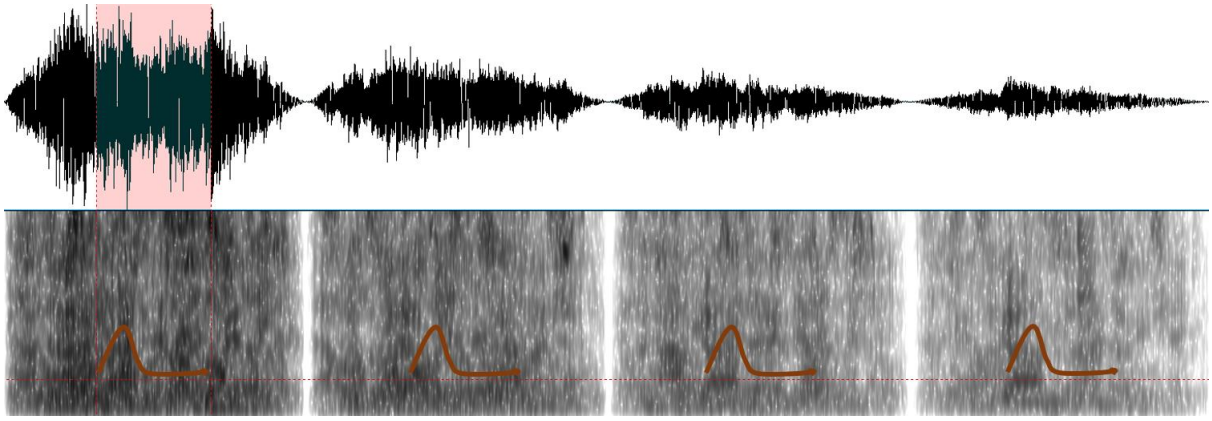

Figure 1. Four example speech-in-noise stimuli for the item *CAnon. From left to right, the SNR decreases from -18 to -12 to -6 to 0 dB. The target speech signal (CAnon) is in orange and highlighted in the waveform in pink. The noise in the spectrograms is seen to gradually decrease while the signal (in orange) remains the same and hence becomes more salient.*

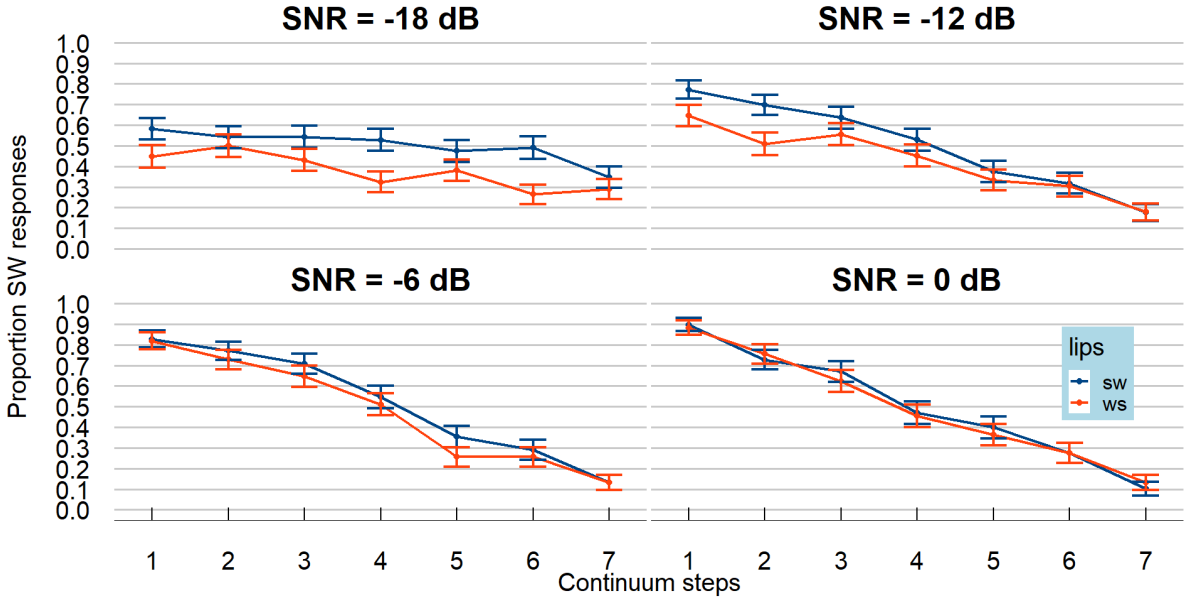

Figure 2. Results of the main experiment: *All four panels show the proportion of strong-weak (SW) responses by the seven-step audio continua (x-axis; step 1 is SW, step 7 is WS), separately for the two video conditions (visual articulatory cues; SW vs. WS) and grouped by SNR. At SNR = 0 dB (bottom right), when the audio signal is intelligible, the proportion of SW responses decreases when auditory steps become more WS-like. Moreover, the visual cues from the videos do not influence participants' responses, shown by the overlapping red and blue lines. As babble intensity increases (from SNR = 0 dB to SNR = -18 dB), participants rely more on the visual stimulus as shown by the red and blue lines separating and less on the auditory target speech as shown by shallower slopes of the lines.*

References

Bujok, R., Meyer, A., & Bosker, H. R. (2022). Audiovisual Perception of Lexical Stress: Beat Gestures are stronger Visual Cues for Lexical Stress than visible Articulatory Cues on the Face. https://doi.org/10.31234/osf.io/y9jck (pre-print).

Jesse, A., & McQueen, J. M. (2014). Suprasegmental lexical stress cues in visual speech can guide spoken-word recognition. *Q. J. Exp. Psychol, 67*, 793-808. DOI: 10.1080/17470218.2013.834371.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748. doi: 10.1038/264746a0

Scarborough, R., Keating, P., Mattys, S., Cho, T., & Alwan, A. (2009). Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English. *Language and speech, 52*, 135-75. 10.1177/0023830909103165.

Stacey, J. E., Howard, C. J., Mitra, S., & Stacey, P. C. (2020). Audio-visual integration in noise: Influence of auditory and visual stimulus degradation on eye movements and perception of the McGurk effect. *Attention, perception & psychophysics*, *82*(7), 3544–3557. https://doi.org/10.3758/s13414-020-02042-x.