Beating stress: evidence for recalibration of word stress perception

Ronny Bujok[1,2], David Peeters[1,3], Antje S. Meyer[1,4] and Hans Rutger Bosker[1,4]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, NL

[2] International Max Planck Research School for Language Sciences, MPI for Psycholinguistics, Max Planck Society, Nijmegen, NL

[3] Department of Communication and Cognition, TiCC, Tilburg University, Tilburg, NL

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

Ronny Bujok: https://orcid.org/0000-0001-6557-9808

David Peeters: https://orcid.org/0000-0002-7974-9246

Antje Meyer: https://orcid.org/0000-0002-7735-9025

Hans Rutger Bosker: https://orcid.org/0000-0002-2628-7738

Corresponding Author: Ronny Bujok (Ronny.Bujok@mpi.nl)

Abstract

Speech is inherently variable, requiring listeners to apply adaptation mechanisms to deal with the variability. A proposed perceptual adaptation mechanism is recalibration, whereby listeners learn to adjust cognitive representations of speech sounds based on disambiguating contextual information. Most studies on the role of recalibration in speech perception have focused on variability in particular speech segments (e.g., consonants/vowels), and speech has mostly been studied in isolation. However, speech is often accompanied by visual bodily signals like hand gestures and is thus multimodal. Moreover, variability in speech extends beyond segmental aspects alone and also affects prosodic aspects, like lexical stress. We currently do not understand well how listeners adjust their representations of lexical stress patterns to different speakers. In three experiments, we investigated recalibration of lexical stress perception, driven by lexico-orthographical information (Experiment 1) and by manual beat gestures (Experiments 2-3). Across experiments, we observed that these types of disambiguating information (presented during an initial brief exposure phase) lead listeners to adjust their representations of lexical stress, with lasting consequences for subsequent spoken word recognition (in an audio-only test phase). However, evidence for generalization of this recalibration to segmentally different words was mixed as it was found only in the final experiment. These results highlight that recalibration is a plausible mechanism for suprasegmental speech adaption in everyday communication and show that even the timing of simple hand gestures can have a lasting effect on auditory speech perception.

Speech produced by different speakers can vary a lot in terms of actual realization. The same word can sound very different depending on who is producing it. This variation poses a problem for our speech perception system tasked with accurately determining what is being said. Listeners must adapt to different speakers and their specific ways of producing speech. One of the ways to achieve this, is by adjusting perceptual category boundaries (e.g., for individual speech sounds) to accommodate the speaker's way of speaking (for review see Ullas et al., 2022). This process is known as *recalibration* (e.g., Bertelson et al., 2003; Norris et al., 2003). Spoken utterances, however, typically combine segmental with suprasegmental information, such as lexical stress patterns and prosodic contours (for review see Cutler, 2008). In addition, many utterances are multimodal in nature, in that they combine spoken with visual information, for instance via co-speech hand gestures (Holler & Levinson, 2019; Kita & Özyürek, 2003; McNeill, 2008; Wagner et al., 2014). We do not yet fully understand whether and how suprasegmental information is recalibrated, and how in this process spoken and manual sources of information may jointly play a role. This study therefore aims to test whether listeners can recalibrate their perception of suprasegmental information, specifically lexical stress, when perceiving multimodal messages.

Recalibration is a domain-general perceptual mechanism (e.g., Noppeney, 2021) that serves to achieve perceptual constancy in the perceiver despite variability in the input. It has been observed in color perception (Mitterer & de Ruiter, 2008), auditory spatial localization (Radeau & Bertelson, 1974), audiovisual synchrony perception (Aller et al., 2022; Burg et al., 2013), and spoken word recognition (Bertelson et al., 2003; Norris et al., 2003). In the last-mentioned field, it has been proposed that listeners use recalibration to deal with variability in speech. If, for example, someone hears an ambiguous fricative, which lies somewhere between an /f/ and /s/, they can learn to interpret the ambiguous fricative as either an /f/ or /s/ depending on disambiguating information (Norris et al., 2003). For instance, when participants are repeatedly presented with the ambiguous fricative in a lexical context that disambiguates the sound as an /f/ (e.g., by hearing it in the word "gira?"), they learn to categorize the sound as /f/. In contrast, in an /s/-biasing context (e.g., hearing it in the word “platypu?”), they learn to categorize the same sound as /s/. Crucially, in a subsequent test phase in the absence of the disambiguating information they still categorize the ambiguous sound as either /f/ or /s/ depending on the biasing context they had been exposed to earlier (Norris et al., 2003). Recalibration is believed to involve changes in the perceptual boundaries of abstract phoneme representations, such that the initially ambiguous fricative is considered an acceptable token of /f/ (after exposure to /gira?/) or /s/ (after exposure to /platypu?/) (Kleinschmidt & Jaeger, 2015; Xie et al., 2023).

Studies have found recalibration effects in word recognition driven by various types of disambiguating information, including lexical (Norris et al., 2003), semantic (Jesse, 2021), lexico-orthographic (Bosker, 2022; Keetels et al., 2016), and visual articulatory cues (Bertelson et al., 2003). For instance, when repeatedly exposed to an ambiguous sound between a /b/ and /d/ together with a video of the speaker’s face producing either a visual /b/ or /d/ (i.e., lips touching each other vs. tongue touching alveolar ridge), participants recalibrated their perception of the ambiguous sound. That is, participants who saw a visual /b/ were more likely to perceive the ambiguous sound as a /b/ in a later audio-only test phase than participants who had seen a visual /d/. The participants’ perception of the same auditory stimulus changed based on the disambiguating visual information provided earlier (Bertelson et al., 2003). This observation has been taken as evidence for participants forming abstract representations of speech sounds, recalibrating their perception of different cues based on disambiguating context, and then using the recalibrated representations to comprehend a speaker, even in the subsequent absence of disambiguating cues.

However, speech can differ between speakers in many more ways than just the segments. Indeed, speech can also differ in its prosodic properties, such as speech rate (Maslowski et al., 2019) and lexical stress (Severijnen et al., 2021). Prosodic information can play a significant role in word recognition. In some languages including Dutch, which is studied here, lexical stress is contrastive, meaning that there are instances where lexical stress is the only cue differentiating two segmentally identical words (e.g., Dutch *VOORnaam* [first name] vs. *voorNAAM* [respectable]; /vo:r.na:m/). In these instances, lexical stress is crucial to understand which word the speaker means. Just like segmental variation, the production of lexical stress can vary significantly between speakers (e.g., Severijnen et al., 2021). It can be conveyed by different acoustic cues such as fundamental frequency (F0), duration and intensity (Rietveld & Heuven, 2009), and different speakers use these cues differently and in varying degrees (Severijnen et al., 2023). Therefore, the segmental variability problem introduced above extends to suprasegmental aspects of speech. Hence, people might benefit from adaptation to different prosodic realizations of speech.

Previous studies have found recalibration of prosodic aspects of speech including lexical tone in Mandarin (Mitterer et al., 2011) and sentence-level intonation in English (Kurumada et al., 2012). One study has found that recalibration of lexical stress perception is also possible (Bosker, 2022). Participants in that study listened to ambiguously stressed stimuli from a lexical stress continuum of a single Dutch minimal pair (*CAnon* [canon] *&* *kaNON* [cannon]; /kaː.nɔn/). One group of participants listened to the ambiguous stimuli with a concurrent orthographic word form presented on a computer screen indicating stress on the first syllable (strong-weak; SW, e.g., *CAnon*). Another group heard the same ambiguous speech while seeing an orthographic word form indicating stress on the second syllable (weak-strong; WS, e.g., *kaNON*). It was observed that participants learned to associate the ambiguous acoustic properties of the stimuli with either a strong-weak or weak-strong lexical stress pattern. That is, in a later test phase, they were instructed to categorize words taken from a lexical stress continuum of the same word pair (*CAnon* – *kaNON*) as either strong-weak or weak-strong. Crucially the test phase was audio-only; that is, no disambiguation by orthographic forms on screen was provided. The group that had listened to the stimuli while seeing the SW orthographic form on screen in exposure categorized the entire continuum as more SW-like (i.e., gave a higher proportion SW responses) in the test phase than the group that had listened to the same stimuli while seeing the WS orthographic form on screen in exposure (Bosker, 2022).

Moreover, the same study also provided preliminary evidence for generalization of the recalibration acquired during exposure, to novel word items at test. That is, when new participants were presented with a segmentally different stress continuum (*SERvisch* [Serbian] – *serVIES* [crockery]) in exposure, they too perceived the same *CAnon* – *kaNON* continuum at test as either more SW-like or WS-like depending on whether the disambiguating orthographic form in exposure indicated SW or WS stress, respectively. This could indicate that participants do not only learn stress patterns on a word-by-word basis, but that their changed perception of lexical stress can be generalized and applied to different words. In the literature this generalization is generally taken as evidence for abstraction of the acoustic signal (e.g., Cutler et al., 2010; Mitterer et al., 2011): at a prelexical level, the information in the speech signal is categorized in terms of sublexical units that can be adjusted on a speaker-specific basis (e.g., in such models like TRACE and Shortlist B (McClelland & Elman, 1986; Norris & McQueen, 2008)).

Importantly, however, Bosker (2022) used highly artificial speech continua in the experiment. The original F0 contours of the recorded speech were removed and replaced by artificial linear downward slopes for each syllable with its mean F0 varying across the continuum. That is, the SW word had a relatively high mean F0 on the first syllable and a relatively low mean F0 on the second syllable. In contrast, the WS word had a relatively low mean F0 on the first syllable and a relatively high mean F0 on the second syllable. For the ambiguous steps, the mean F0 for the syllables was gradually lowered or raised to create the continuum. Most critically, this artificial F0 manipulation was then applied to *both word pairs* (i.e., same F0 values and contours in the *canon-kanon* continuum as in the S*ervisch-servies* continuum). Hence, participants in Bosker (2022) demonstrated evidence of generalizing their recalibration effect to a segmentally different, but suprasegmentally identical continuum. This means that the generalization effect found in Bosker (2022) does not necessarily reflect an adaptation of abstract representations of stress patterns but could also reflect an adaptation to specific F0 values. Hence, one goal of the present study was to assess whether recalibration and generalization are also possible with more naturalistic F0 contours and thus more acoustic distance between the words.

The second and most central goal of the current study was to assess whether listeners can use *visual* information to recalibrate perception of suprasegmental aspects of speech such as lexical stress. While the production of lexical stress is less clearly associated with visual articulatory cues than certain speech segments (e.g., salient mouth closing when producing a /b/), it nevertheless has visual correlates such as the typically wider and longer mouth opening on stressed syllables (Scarborough et al., 2009). These articulatory cues are visible and used by participants to categorize "talking faces" (i.e., muted videos) producing different stress patterns differently (Bujok et al., 2022; Jesse & McQueen, 2014; Scarborough et al., 2009). However, interestingly, when presented with audiovisual (AV) stimuli, the same visual articulatory information does not lead to different percepts (Bujok et al., 2022). That is, the same sound, paired with either a face articulating stress on the first or second syllable, is perceived similarly. Therefore, it is unlikely that the facial articulatory cues to stress, which do not even appear to be used in online audiovisual stress perception, could drive recalibration.

In contrast, other visual cues could be used to recalibrate the perception of lexical stress. Hand gestures are commonly produced in face-to-face conversations and have been shown to affect spoken word recognition, particularly in noisy settings (Drijvers & Özyürek, 2017). One particular kind of hand gestures, so-called beat gestures, mainly defined as simple bi-phasic up-and-down movements of the hands, tend to align with acoustically prominent parts of utterances (Krahmer & Swerts, 2007). Specifically, the point of maximum extension of the beat gesture, the so-called apex, is strongly temporally related with pitch accent (Leonard & Cummins, 2011) and affects the acoustic realization of the pitch accent as well (Krahmer & Swerts, 2007; Pouw et al., 2020; Swerts & Krahmer, 2007), making the accented utterance even more prominent (Krahmer & Swerts, 2007). Beat gestures have been found to help listeners focus their attention on important information (Biau &

Soto-Faraco, 2013, 2015) and to increase processing of the focused words (Dimitrova et al., 2016). Moreover, beat gestures boost memory recall of the words they are aligned with (Kushch & Prieto, 2016). However, we do not know whether listeners make use of the information provided by beat gestures for suprasegmental recalibration purposes.

On the word level, the apex of a beat gesture is usually temporally aligned to the F0 peak of a stressed syllable (Leonard & Cummins, 2011; Shattuck-Hufnagel & Ren, 2018). As such, listeners can take advantage of this close temporal link and use it in lexical stress perception. That is, people are more likely to perceive stress on a syllable when a beat gesture is aligned to it. For instance, when Dutch participants hear tokens from a lexical stress continuum of /kaː.nɔn/ (ranging from *CAnon* to *kaNON*) with a beat gesture on the first syllable, they are more likely to report perceiving *CAnon* rather than *kaNON* (Bosker & Peeters, 2021; Bujok et al., 2022). This effect is robust across lexical stress continua and was present for several different Dutch minimal word pairs.

Given this effect, we hypothesized that the temporal alignment of beat gestures could be a cue for recalibration of lexical stress, in analogy to how a talking face can recalibrate the perception of a /b/ - /d/ continuum (Bertelson et al., 2003). Essentially, we asked whether the effect of beat gestures goes beyond the immediate effect of disambiguating an ambiguously stressed word (Bosker & Peeters, 2021) and lead to lasting changes in speech perception. Finding evidence for a recalibration effect driven by beat gestures would extend current models of recalibration (Kleinschmidt & Jaeger, 2015; Xie et al., 2023) to include visual cues beyond articulation (Bertelson et al., 2003) and lexico-orthographical information (Bosker, 2022). Such evidence would be consistent with multimodal frameworks of spoken language comprehension (Holler & Levinson, 2019; Özyürek, 2014).

In three behavioral experiments, we tested participants' ability to recalibrate their perception of lexical stress in a specific word, as well as their ability to generalize recalibration to a novel item. We first targeted recalibration guided by disambiguating written word forms (Experiment

1) and then targeted – for the first time – recalibration guided by beat gestures (Experiments 2-3). Thus, in the first experiment, we adopted a paradigm similar to Bosker (2022), testing recalibration of the perception of lexical stress by written information. Critically, we used different stimuli, with more naturalistic phonetic stress continua based on the original F0 contours. Consequently, the F0 contours for the two minimal word pairs (used to test generalization of recalibration to new words) were distinct. This arguably makes finding evidence for generalization more difficult but it does better reflect naturalistic spoken communication, where every F0 contour is unique. In the second and third experiment, we used the same auditory stimuli and a similar experimental paradigm but tested whether the recalibration of lexical stress perception could be driven by the temporal alignment between spoken words and visual beat gestures.

Experiment 1 - recalibration driven by words on screen

The first experiment was a conceptual replication of Bosker (2022), but with different stimuli to test whether recalibration effects can be found with more naturalistic stimuli. We expected to replicate the original recalibration findings in the Segmental Overlap Condition (i.e., when participants are tested on the same word pair they were exposed to). However, given our more variable and naturalistic phonetic continua, we were not certain about the generalization of the effect to different words. If the generalization effect found by Bosker (2022) was at least in part driven by the artificial and identical F0 continua between the word pairs, we should not find a generalization effect with more naturalistic stimuli. On the other hand, finding a generalization effect here would provide strong evidence that listeners are in fact able to generalize recalibration to different words with similar, but not identical, stress cues.

Method

Participants

All participants tested in this study gave informed consent as approved by the Ethics Committee of the Social Sciences department of Radboud University (project code: ECSW-2019-019). Only participants who reported no hearing or language deficit and normal or corrected-to-normal vision participated. Participants were financially compensated for their participation. For Experiment 1 we tested 72 participants (59 female, 13 male), recruited from the Max Planck Institute for Psycholinguistics participant database. Their median age was 24 (SD = 3.67, range = 18 – 36).

Materials

Materials for this experiment were adopted from previous experiments (Bujok et al., 2022). Two disyllabic, segmentally identical minimal stress pairs of Dutch, which only differed in the position of lexical stress, were chosen (*CAnon* [canon] vs. *kaNON* [cannon]; *VOORnaam* [first name] vs.

voorNAAM [respectable]; capitals indicate lexical stress). We recorded high-definition videos of a male native speaker of Dutch (i.e., the last author), while he was sitting down, producing these words naturally without any manual gesture. The audio sampling rate was 48 kHz.

A lexical stress continuum was created by measuring the F0 contours of the original recordings and then linearly interpolating between the contours in 11 steps (ranging from the original SW recording to the original WS recording, see Figure 1). The duration and intensity of each syllable was held constant at an ambiguous, average value (i.e., midway between stressed/unstressed), determined for each pair based on the original recordings. The interpolated F0 contours were then applied to the SW token using PSOLA in Praat (Boersma, 2006). Note that this contrasts with the continuum manipulation in Bosker (2022), where F0 contours on either syllable always involved linear downward slopes, only varying in mean F0 height, removing any sign of the original contours. Another consequence of these artificial manipulations was that both continua were identical with regards to F0. In contrast, our more naturalistic contour interpolation method entailed that every manipulated stimulus had a unique F0 contour.

The manipulated 11-step continua were presented to 10 participants (who did not participate in any of the experiments) in a pretest in a two-alternative-forced-choice (2AFC) task, where they had to categorize the words as either SW or WS. Based on the categorization results, we selected 5 ambiguous steps for each pair, which ranged between ~80% and ~20% proportion SW responses, to create a perceptual continuum. Together with the original tokens, this resulted in a 7-step continuum. Step 1 thus refers to the original SW token, and step 7 to the original WS token. The five steps in between (steps 2 – 6), with varying degrees of ambiguity, will be referred to as ambiguous steps. The middle step (step 4) was the most ambiguous, lying closest to 50% SW categorization responses.

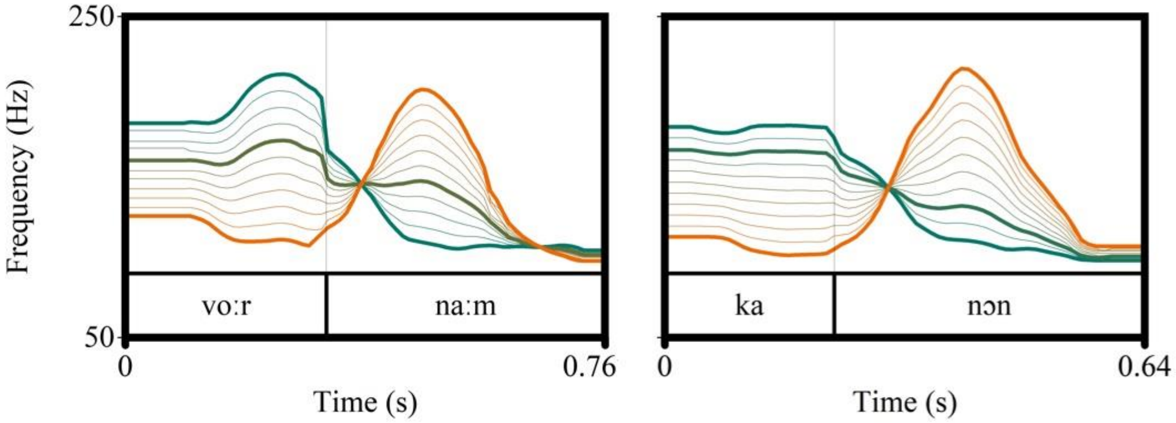

Figure 1. Visualization of the F0 stress manipulation for /voːr.naːm/ (left panel) and /kaː.nɔn/ (right panel): F0 contours were interpolated in 11 steps to go from SW (green) to WS (orange). Five manipulated steps were selected and presented with the original SW and WS recordings as a perceptual 7-step continuum. The extremes of the continua and the perceptually most ambiguous step are highlighted in bold.

Design and Procedure

Data for all experiments reported here were collected online using the Gorilla Experiment Builder (http://gorilla.sc) (Anwyl-Irvine et al., 2020). Participants had to complete a headphone screening prior to the experiments, to ensure usage of high quality headphones (based on Huggins Pitch, see Milne et al., 2021). We adopted the design used by Bosker (2022), consisting of two phases: an exposure phase and a test phase. In the exposure phase, participants were assigned to one of two conditions: The Segmental Overlap Condition or the Generalization Condition. These conditions were identical in their procedure, but differed in the items presented during exposure (see Table 1).

Table 1.

Overview of the Design of Experiment 1: Conditions and Stimuli presented

Condition	Group Bias	Exposure (AV)		Test (A-only)
		Audio	Video	
Segmental Overlap	SW Bias	amb. /ka:.nɔn/ (step 4)	*CAnon*	/ka:.nɔn/ - continuum (steps 2 – 6)
		WS /ka:.nɔn/ (step 7)	*kaNON*	
	WS Bias	amb. /ka:.nɔn/ (step 4)	*kaNON*	
		SW /ka:.nɔn/ (step 1)	*CAnon*	
Generalization	SW Bias	amb. /vo:r.na:m/ (step 4)	*VOORnaam*	
		WS /vo:r.na:m/ (step 7)	*voorNAAM*	
	WS Bias	amb. /vo:r.na:m/ (step 4)	*voorNAAM*	
		SW /vo:r.na:m/ (step 1)	*VOORnaam*	

In the Segmental Overlap Condition, participants were randomly divided into two counter-balanced groups (18 participants per group). In the SW-Bias group, on half of the trials, participants were presented with an audio recording of a speaker producing a clear token of the word *kaNON* (i.e., stress on second syllable; WS; step 7 from the 7-step continuum) while being presented with the orthographic form “kaNON” on screen (with capitalized letters indicating stress). Additionally, on other trials, they were presented with an acoustically ambiguous auditory token (step 4 from the 7-step continuum), which was disambiguated by the orthographic form “CAnon” with stress on the first syllable appearing on screen. Consequently, the SW-Bias group was predicted to learn that the talker produced ambiguous auditory stress cues associated with an SW prosodic pattern. In contrast, the other group (WS-Bias) was presented with audio recordings of the speaker producing a clear token of the word *CAnon* (i.e., stress on first syllable; SW) while seeing “CAnon” on screen. Moreover, they were also presented with the acoustically ambiguous auditory token, critically together with the written word “kaNON” on screen. This group was thus biased to associate the ambiguous stress cues with a WS prosodic pattern. The clear audio trials and the ambiguous audio trials were presented 24 times each, resulting in 48 exposure trials. Participants passively listened to all the stimuli (interstimulus interval: 600ms, static fixation cross). Then they moved on to the test phase described below.

In the Generalization Condition, the design and procedure of the exposure phase was similar to the Segmental Overlap Condition. However, during the exposure phase, participants in the Generalization Condition were presented with a different item pair. Specifically, the SW-Bias group received a clear auditory *voorNAAM* with the congruent orthographic form “voorNAAM” with stress on the second syllable, and an ambiguous auditory token from the *VOORnaam – voorNAAM* continuum (step 4) with a disambiguating orthographic form “VOORnaam”. Conversely, the WS-Bias group got a clear *VOORnaam* with congruent “VOORnaam” on screen, and the ambiguous token (step 4) with a disambiguating written word “voorNAAM” on screen.

All participants received the same test phase. That is, they were tested on the same manipulated F0 continuum made up of the 5 ambiguous steps from the *CAnon – kaNON* continuum (i.e., steps 2 - 6) in a two-alternative-forced-choice (2AFC) task. Hence participants from the Segmental Overlap Condition were tested on the word pair they had been exposed to, whereas participants from the Generalization Condition were tested on a different word pair than they had been exposed to. Each step was presented 15 times, equaling a total of 75 trials presented in random order. After stimulus offset, two response options were shown, one on either side of the screen. Participants were asked to categorize what they heard as corresponding either to *CAnon* (SW) or *kaNON* (WS) by pressing the left (“Z”) or right (“M”) button on their keyboard, corresponding to the left and right word on the screen respectively. The position of SW and WS words on screen was counter-balanced across participants. Participants were given a 4000 ms time limit to respond.

Results

We removed all trials where participants failed to give a response (*n* = 14, 0.25% of all observations). We analyzed our data with Generalized Linear Mixed Models using the lme4 library

(Bates et al., 2015) in R (R Core Team, 2021). The independent variable was the participants' Categorization Response (SW coded as 1, *CAnon*; WS coded as 0; *kaNON*). Fixed effects included Continuum Step (continuous, z-scored), Group Bias (categorical, deviance coded SW as 0.5 and WS as -0.5), Condition (categorical, deviance coded Segmental Overlap as -0.5 and Generalization as 0.5) and the interactions of Condition with the other fixed effects. Additionally, the model included random effects for Participants, as well as maximal random slopes for Continuum Step, Group Bias, and Condition (*Response ~ Condition*(Step + Group Bias) + (1+ Condition + Step + Group Bias|Participant_ID)*). All data and code are publicly available on https://osf.io/s3p6a/?view_only=e4e822e23a7440f2bd22a25bfb1dff95 .

The model showed a significant Intercept, demonstrating an overall bias to give slightly more SW than WS responses (mean proportion of SW responses = 0.57; β in logit space = 0.863, $SE = 0.113$, $z = 7.649$, $p < 0.001$). Continuum Step was also significant ($\beta = -3.13$, $SE = 0.278$, $z = -11.272$, $p < 0.001$), indicating that participants' proportions of SW responses decreased with increasing Continuum Steps. Condition ($\beta = 0.08$, $SE = 0.222$, $z = 0.36$, $p = .72$) and its interaction with Continuum Step ($\beta = -0.5$, $SE = 0.539$, $z = -0.927$, $p = .354$) were not significant suggesting similar response patterns across the Segmental Overlap and Generalization Conditions. Most critically, the predictor Group Bias did not have a main effect on the responses ($\beta = -0.18$, $SE = 0.156$, $z = 1.156$, $p = .248$), but showed an interaction with Condition ($\beta = -0.745$, $SE = 0.306$, $z = -2.436$, $p = .015$), meaning that the effect of Group Bias was stronger in the Segmental Overlap Condition than the Generalization Condition (see Figure 2). In fact, two follow-up models that were run on each condition separately confirmed that the Group Bias effect was significant in the Segmental Overlap Condition ($\beta = 0.566$, $SE = 0.247$, $z = 2.286$, $p = .022$) in the expected direction: the proportion of SW responses was higher in the SW-Bias group compared to the WS-Bias group. In contrast, no effect of Group Bias was found for the Generalization Condition ($\beta = -0.215$, $SE = 0.189$, $z = -1.14$, $p = .257$).

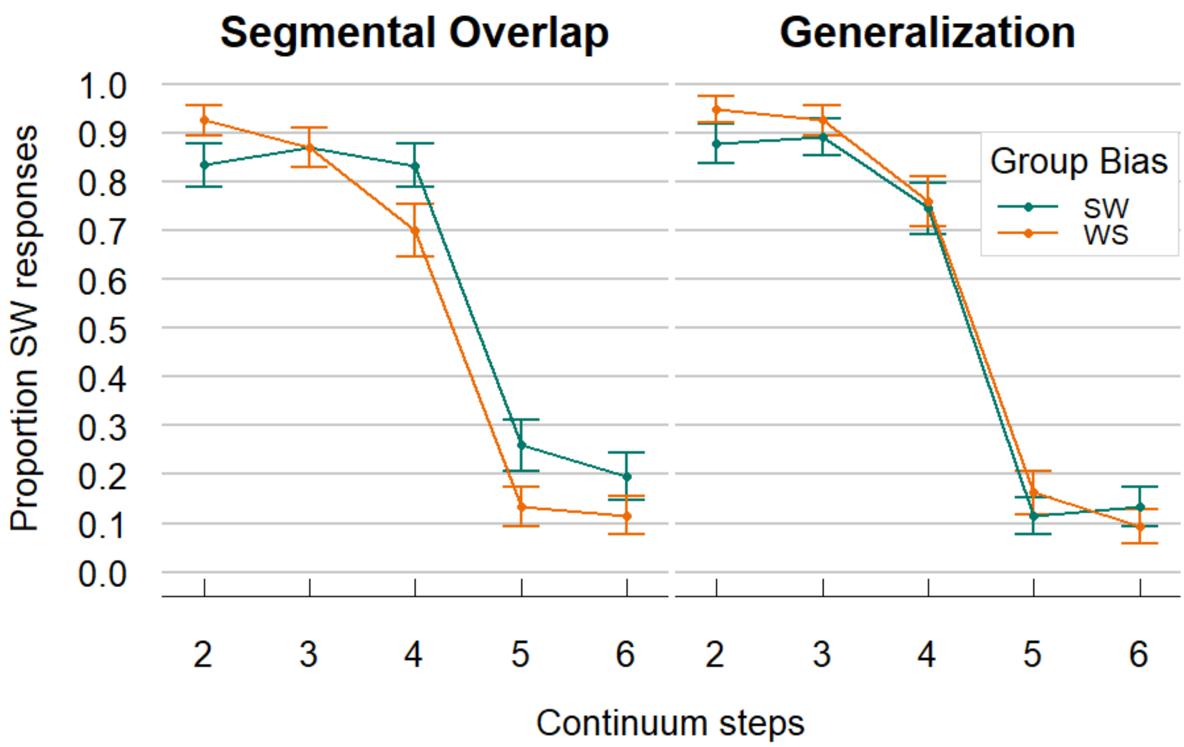

Figure 2. Results from Experiment 1: Comparison of the audio-only test results in the Segmental Overlap and Generalization Condition. The proportion of SW responses (i.e., stress on the first syllable) generally decreases as auditory Continuum Step increases (i.e., sounding more WS-like, stress on the second syllable). Different audiovisual (AV) exposure to disambiguating orthographic word forms in the two groups (SW-Bias vs. WS-Bias) changed responses to the audio only (A-only) test continuum in the Segmental Overlap Condition, as can be seen by the separation of the two lines in the left panel. That is, participants who were exposed to ambiguous /ka:.nɔn/ in exposure, paired with orthographic form "CAnon", generally perceived the continuum as more SW-like than participants who were exposed to the same ambiguous tokens of /ka:.nɔn/ paired with "kaNON". However, there was no main effect of Group Bias in the Generalization Condition. Note: Continuum steps go from 2 – 6, as participants were only tested on the 5 ambiguous tokens of the 7-step continuum. SW = strong-weak, stress on first syllable; WS = weak-strong, stress on second syllable. Error bars indicate a 95% confidence interval.

Interim Discussion

The current experiment aimed to conceptually replicate the findings from Bosker (2022) using more naturalistic stimuli. That is, we tested whether listeners were able to recalibrate their perception of lexical stress based on disambiguating lexico-orthographic information. In the Segmental Overlap Condition (i.e., when tested on the same word as heard during exposure), participants showed a group-dependent bias in their perception, which was shifted in the direction of the disambiguating information in exposure (e.g., ambiguous audio disambiguated by written "CAnon" leading to more "*CAnon*" responses). They did so not only for the specific ambiguous token (step 4), that was presented in the exposure phase but also for other tokens on the continuum, indicating a certain degree of generalization to novel acoustic tokens of the same word pair. However, participants did not generalize their recalibration to a segmentally different word, contrasting with the findings in Bosker (2022).

There were two major differences between our experiment and Bosker (2022), both related to the stimuli that were used. First, Bosker (2022) created artificial, linear slopes at different mean F0 heights, separately for each syllable, to generate the continua. Second, they applied identical F0-slopes to both items. This manipulation procedure likely facilitated generalization from one word in exposure to a novel item at test, carrying the exact same F0 contour as in exposure. In contrast, we tested for a recalibration effect using more naturalistic continua. By interpolating the original F0 contours of both members of a pair, we created more complex and arguably more natural continua. Not only the height but also the overall shape of the F0 contour cued stress in our stimuli. A consequence of this procedure was that the continua of the two item pairs were acoustically unique and distinct (see Figure 1). Despite these differences with Bosker (2022), we also found a recalibration effect. In fact, our data suggest that the perception of lexical stress can recalibrate even with more naturalistic continua, than used by Bosker (2022). However, we did not find a generalization effect. Thus, we caution that the generalization of the recalibration effect to novel words is sensitive to the stimuli used.

The recalibration effect might be sensitive to the source of the disambiguating information, too. From studies of segmental recalibration, we know that lexical information tends to result in smaller recalibration effects than audiovisual information (Ullas et al., 2020a, 2020b; van Linden & Vroomen, 2007). We do not know if this observation extends to recalibration of lexical stress. However, we rarely encounter speech-disambiguating orthography in our daily lives. In contrast, manual beat gestures are ubiquitous (McClave, 1994). It is thus possible that beat gestures, being a more ecologically valid cue, could be a stronger source of information and lead to larger effects. Beat gestures are temporally closely aligned to stressed syllables (e.g., Krahmer & Swerts, 2007; Pouw & Dixon, 2019; Shattuck-Hufnagel & Ren, 2018), which affects online perception of lexical stress (Bosker & Peeters, 2021; Bujok et al., 2022). That is, the alignment of a beat gesture with a syllable makes participants more likely to perceive stress on that syllable. Moreover, beat gestures have been found to redirect attention to a concurrently produced word (Biau & Soto-Faraco, 2013). There is also evidence that beat gestures are processed and integrated automatically with speech (Kelly et al., 2010), and that their uptake is not inhibited if they are only perceived in the visual periphery (Gullberg & Kita, 2009). This leads us to hypothesize that beat gestures could be strong inducers of recalibration in lexical stress perception. Hence Experiment 2 assessed the effect of beat gestures on the recalibration of lexical stress perception.

Experiment 2 – recalibration driven by beat gestures (between-subject)

Influential multimodal theories of language indicate that gestures are an inherent aspect of our everyday face-to-face communication (Holler & Levinson, 2019; Hübscher & Prieto, 2019; Kita & Özyürek, 2003) and are processed and integrated automatically with speech (Kelly et al., 2010). As such they could be a possible source of disambiguation for recalibration of lexical stress. We tested this hypothesis by running Experiment 2, which was similar in design to Experiment 1, but used temporally aligned beat gestures rather than orthographic words as disambiguating information in exposure. Participants in different groups were exposed to videos of a talker producing an ambiguously stressed word with a beat gesture on either the first (SW) or second syllable (WS). If beat gesture alignment can be used as a cue to recalibration, participants in the SW and WS groups should be biased to perceive the same stress continuum in the test phase differently.

Method

Participants

Seventy-two native speakers of Dutch (34 female, 37 male, 1 gender not reported) were recruited for this experiment through Prolific. Median age was 25 (SD = 4.9, ranging = 19 – 38).

Materials

For this study we again used stimuli from Bujok et al. (2022). For a detailed description of the audio and phonetic manipulations see the Materials section for Experiment 1 above. Additionally, for Experiment 2 we used video stimuli to test whether visual beat gestures could drive recalibration of lexical stress perception. The same talker from Experiment 1 had been video-recorded producing all four words (*CAnon, kaNON, VOORnaam, voorNAAM*) with a beat gesture. The beat gesture was an up-and-down, forward-rotating movement of the right hand, with the apex (the point of maximal extension) naturally aligned to the stressed syllable. The speaker was sitting in front of a neutral background and framed from the hip up. Videos were recorded at a sampling rate of 50 Hz and cropped to 620 x 620 pixel squares.

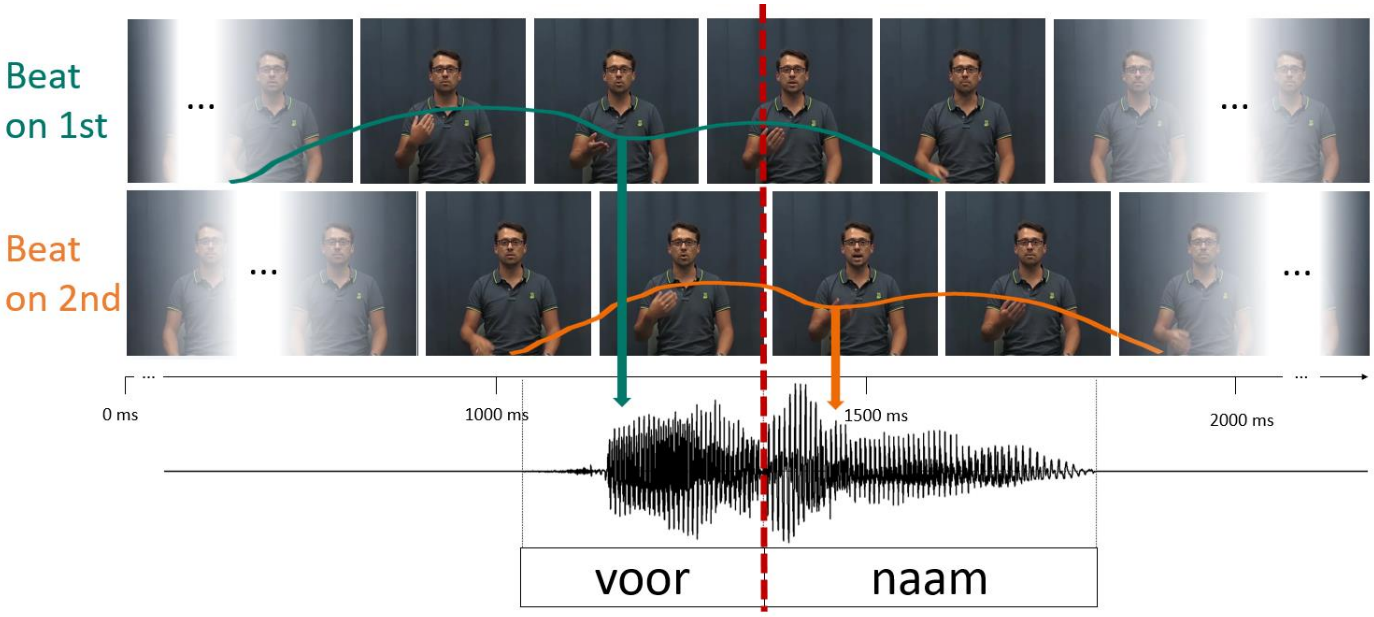

Figure 3. Audiovisual Stimuli from Experiment 2. Apex of the beat gesture was aligned to either the first (Beat on 1st, green) or second syllable (Beat on 2nd, orange). Colored lines show position of the hand and thus movement of the gesture over time. Arrows indicate approximate alignment of the gesture's apex with concurrent speech. Videos were combined with all steps of the auditory stress continuum, aligned at second syllable onset.

The manipulated auditory stress continua, as described in Experiment 1, were combined with the original video recordings such that every auditory step was combined with both SW and WS videos. This created our final audiovisual stimuli for use in the exposure phase (see Figure 3). Because the duration of the audio was manipulated to make the duration cues ambiguous with regards to lexical stress, the audio was slightly misaligned with the original, unchanged video (mean = 40ms). We aligned audio and video at second syllable onset precisely by shifting the second syllable onset of the manipulated audio to the time of the original second syllable onset. We decided to align at second syllable onset to minimize misalignment at word onset and offset. This led to slight variation of the beat gesture alignment within each syllable, but, because of the alignment at the syllable boundary, all beat gestures were still aligned with the correct syllable.

Design and Procedure

The experimental design was similar to Bosker (2022) and Experiment 1, consisting of an exposure phase and a test phase. It used the same auditory stimuli as Experiment 1. However, now in exposure, participants were exposed to the audio together with video, with the disambiguating information being beat gestures (no orthographic forms; see Table 2). Participants were again assigned to one of two conditions: The Segmental Overlap Condition or the Generalization Condition. Within each condition participants were assigned to either the SW or WS-Bias group.

Table 2.

Overview of the Design of Experiment 2: Conditions and Stimuli presented

Condition	Group Bias	Exposure (AV)		Test (A-only)
		Audio	Video	
Segmental Overlap	SW Bias	amb. /ka:.nɔn/ (step 4)	*Beat on 1st*	/ka:.nɔn/ - continuum (steps 2 – 6)
		WS /ka:.nɔn/ (step 7)	*Beat on 2nd*	
	WS Bias	amb. /ka:.nɔn/ (step 4)	*Beat on 2nd*	
		SW /ka:.nɔn/ (step 1)	*Beat on 1st*	
Generalization	SW Bias	amb. /vo:r.na:m/ (step 4)	*Beat on 1st*	
		WS /vo:r.na:m/ (step 7)	*Beat on 2nd*	
	WS Bias	amb. /vo:r.na:m/ (step 4)	*Beat on 2nd*	
		SW /vo:r.na:m/ (step 1)	*Beat on 1st*	

In the exposure phase of the Segmental Overlap Condition, the participants in the SW-Bias group were presented with an original video of the speaker producing the word *kaNON* (i.e., stress on second syllable; WS) while making a beat gesture on the second syllable. Additionally, they were presented with an acoustically ambiguous auditory token (step 4 from the 7-step continuum), which was disambiguated by the talker making a beat gesture on the first syllable. In contrast, participants in the WS-Bias group were presented with a video of the speaker producing clear *CAnon* (i.e., stress on the first syllable; SW) with a congruent beat gesture on the first syllable. The ambiguous auditory token was presented with a video of the talker producing a beat gesture on the second syllable. Participants in the SW-Bias group were expected to learn that the acoustic properties of the ambiguously stressed word were intended to express an SW prosodic pattern. Conversely, the WS-Bias group was expected to learn that the same ambiguous acoustic cues were intended to express a WS prosodic pattern. The exposure phase in the Generalization Condition was identical in design, but participants were presented with videos of the speaker producing a different word pair: *VOORnaam – voorNAAM.*

In total, the exposure phase had 48 trials of which 24 involved original videos and 24 involved ambiguous audio disambiguated by the temporal alignment of the beat gesture. Participants had no task in exposure and were only asked to passively watch the videos, although we emphasized that they had to pay attention to both audio and video. Participants proceeded from one trial to the next by pressing the spacebar. Each video was preceded by a fixation cross for 500ms and then the video was played and disappeared once it stopped playing, asking participants to press the spacebar to continue.

In the test phase, all participants were tested only on the ambiguous auditory tokens from the auditory *CAnon – kaNON* continuum (steps 2 - 6), without any video. The test phase was identical to the test phase of Experiment 1 in terms of stimuli and procedure; hence it exclusively included audio-only trials (no videos).

Results

We removed all timeout trials before analysis (*n*= 32, 0.5% of all observations). The remaining data were analyzed with Generalized Linear Mixed Models with logistic linking function using the lme4 library (Bates et al., 2015) in R (R Core Team, 2021). We used the same model from Experiment 1 to analyze these data. The model revealed a significant Intercept, indicating an overall bias to give more SW than WS responses (mean proportion of SW responses = 0.57, $\beta = 0.736$, $SE = 0.086$, $z = 8.609$, $p < 0.001$). As expected, Step was highly significant ($\beta = -2.291$, $SE = 0.2$, $z = -11.484$, $p < 0.001$) reflecting our auditory stimulus manipulation. As in Experiment 1, with increasing steps (i.e., continuum becoming more WS-like), the proportion of SW responses decreased. Crucially, we found an effect of Group Bias ($\beta = 0.676$, $SE = 0.144$, $z = 4.691$, $p < 0.001$), which meant that participants generally gave a higher proportion of SW responses when they were in the SW-Bias group and a lower proportion of SW responses when they were in the WS-Bias group (see Figure 4). This effect suggests successful recalibration of lexical stress perception driven by beat gesture alignment. A model with a Step*Group Bias interaction term did not improve model fit. This demonstrates an overall recalibration effect, indicating that participants generalized their Group Bias to varying degrees of ambiguity. However, there was also a significant interaction between Group Bias and Condition ($\beta = -0.939$, $SE = 0.283$, $z = -3.32$, $p < 0.001$). That is, the size of the Group Bias effect (i.e., the recalibration effect) was reduced in the Generalization Condition relative to the Segmental Overlap Condition. To confirm this, the model was releveled to map the Segmental Overlap Condition onto the intercept, and then once more with the Generalization Condition on the intercept. The releveled models confirmed that the Group Bias effect was present in the Segmental Overlap Condition ($\beta = 1.146$, $SE = 0.245$, $z = 4.673$, $p < 0.001$), but was statistically not significant in the Generalization Condition ($\beta = 0.206$, $SE = 0.146$, $z = 1.409$, $p = 0.159$).

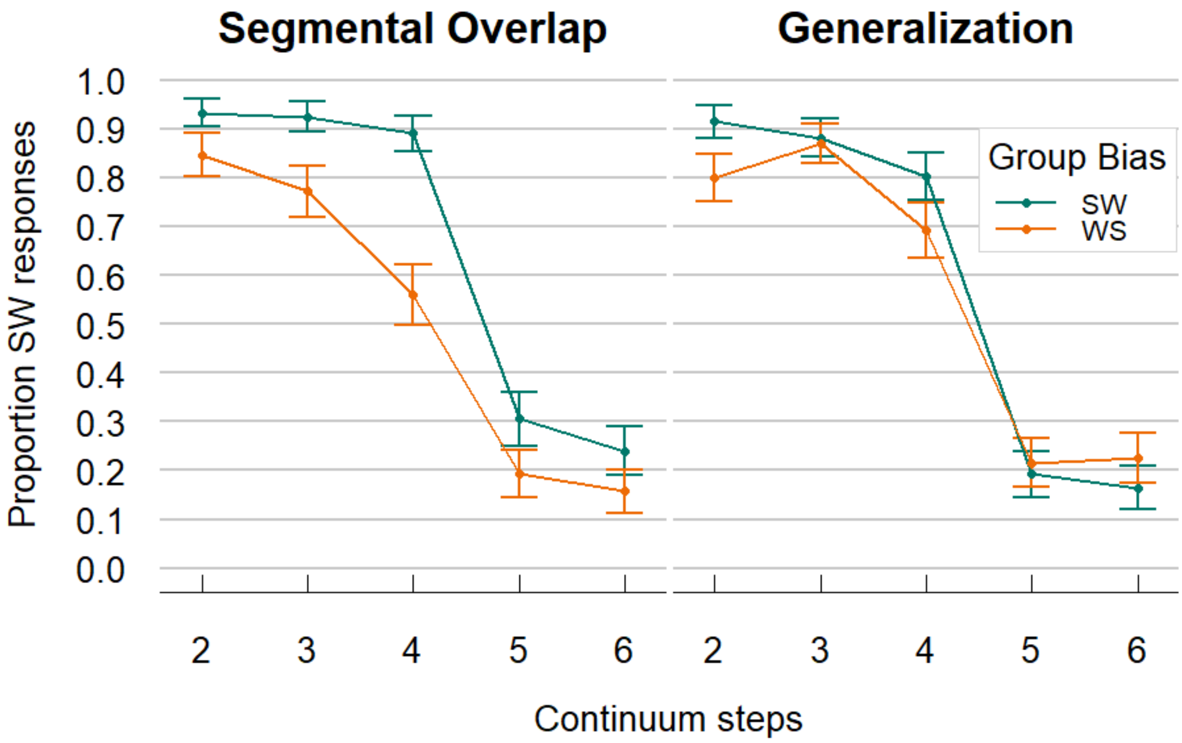

Figure 4. Results from Experiment 2: Recalibration driven by Beat Gestures. Comparison of the results of the Segmental Overlap (left) and the Generalization Condition (right). The proportion of SW responses (i.e., stress on first syllable) is generally highest at step 2 (most SW-like) and lowest at step 6 (most WS-like) in both conditions. The differently colored lines show if participants were in the SW (green) or WS (orange) Bias group in the exposure phase. Only in the Segmental Overlap Condition did the Group Bias from exposure reliably affect the responses in the test phase equally across all steps. That is, participants from the SW Group Bias consistently responded more SW-like, and participants from the WS Group Bias responded more WS-like. The same effect could not reliably be found in the Generalization Condition. SW = strong-weak, stress on first syllable; WS = weak-strong, stress on second syllable. Error bars indicate a 95% confidence interval.

Interim Discussion

Experiment 1 had demonstrated that lexical stress could be recalibrated through lexico-orthographic information (i.e., orthographic form) (Bosker, 2022). Our findings from Experiment 2 are the first to demonstrate that such recalibration can be driven by the alignment of beat gestures to speech as well. We found a Group Bias effect, indicating that participants were biased to perceive the audio-only stress continuum at test differently, depending on the disambiguating gestural information they had been presented with in exposure. Participants who had been presented with the ambiguous stimuli in exposure, disambiguated with a beat gesture aligned to the first syllable (SW), gave more SW responses at test than participants from the WS-Bias group, who had been presented with the same ambiguous auditory stimuli in exposure, but then disambiguated with a gesture aligned to the second syllable.

In contrast, despite a numerical difference, we did not find a significant recalibration effect in the Generalization Condition, where participants were required to categorize the same words after having been exposed to segmentally different words in exposure. We could thus not replicate the generalization findings from Bosker (2022). Note that Experiment 2 tested the two conditions (Segmental Overlap vs. Generalization) between participants, which adds more noise to the data compared to a within-participants design, potentially explaining why the numerical group difference observed in the Generalization Condition in Experiment 2 was not statistically reliable. Because of this consideration, Experiment 3 used an adjusted design, where Condition was tested within participants. That is, all participants were exposed to the same word (i.e., the /ka:.nɔn/ item) and tested both in the Segmental Overlap Condition (/ka:.nɔn/) and in the Generalization Condition (/vo:r.na:m/). As an additional benefit of this design, the number of observations was doubled.

Experiment 3 - recalibration driven by beat gestures (within-subject)

Method

Participants

At our target of seventy-two participants, the groups were not counter-balanced properly, due to a scripting error, so we decided to test an additional eight participants. This left us with the final sample of eighty native speakers of Dutch (38 female, 42 male) with a median age of 25 (SD = 5.32, ranging from 18 - 39). Participants were recruited through Prolific.

Materials, Design and Procedure

We used the same stimuli as in Experiment 2, with the main difference that at test in Generalization we now additionally used steps 2 – 6 from the /vo:r.na:m/ continuum. The design of this experiment was also similar to Experiment 2 with the critical difference that the Segmental Overlap and Generalization Condition were now tested within participants (see Table 3). Group Bias remained a between-participants variable, as is common in recalibration studies.

Table 3.

Overview of the Design of Experiment 3: Conditions and Stimuli presented

Group Bias	Exposure (AV)		Test (A-only)	
	Audio	Video	Segmental Overlap	Generalization
SW Bias	amb. /ka:.nɔn/ (step 4)	*Beat on 1st*	/ka:.nɔn/ continuum (steps 2 -6)	/vo:r.na:m/ continuum (steps 2 -6)
	WS /ka:.nɔn/ (step 7)	*Beat on 2nd*		
WS Bias	amb. /ka:.nɔn/ (step 4)	*Beat on 1st*		
	SW /ka:.nɔn/ (step 1)	*Beat on 2nd*		

The exposure phase, with two groups (SW bias vs. WS bias), was identical to the exposure phase in the Segmental Overlap Condition of Experiment 2. Specifically, participants were only presented with *CAnon – kaNON* videos. The test phase was different from the previous experiments, as all participants were now tested in both conditions. This means the test phase consisted of two different Blocks: the Segmental Overlap and Generalization Block. In the Segmental Overlap Block participants were tested on the middle five steps from the *CAnon – kaNON* continuum (steps 2 - 6), which was the same word they had been exposed to in the exposure phase. In the Generalization Block they were tested on the middle five steps of the *VOORnaam – voorNAAM* continuum (steps 2 - 6), to test their ability to generalize the recalibration effect to segmentally different words. All participants were exposed to the Segmental Overlap and Generalization Condition in separate blocks, with the block order counterbalanced across participants. Each condition presented each of the five steps 15 times, resulting in a total of 75 trials per block, and 150 trials in the entire test phase.

Results and Interim Discussion

We ran the same model as in the previous two experiments, but with the crucial difference that Condition was now a within-participant variable. Results showed a significant Intercept, reflecting a general bias to give more SW than WS responses (mean proportion SW responses = 0.52; $\beta = 0.101$, $SE = 0.047$, $z = 2.15$, $p = 0.032$). Again, the significant effect of Step confirmed lower proportions of SW responses for higher (i.e., more WS-like) steps ($\beta = -2$, $SE = 0.17$, $z = -11.778$, $p < 0.001$). Most importantly, we found a significant effect of Group Bias (see Figure 5; $\beta = 0.436$, $SE = 0.1$, $z = 4.4$, $p < 0.001$). This means that participants gave more SW responses when they were in the SW-bias exposure group than when they were in the WS-bias exposure group. Another model with Step*Bias interaction did not improve model fit, suggesting that the Group Bias effect is thus largely driven by the main effect. Crucially, the interaction with Condition was not significant ($\beta = 0.045$, $SE = 0.131$, $z = 0.345$, $p = 0.73$), suggesting that the Group Bias effect was similarly present in both conditions. Models run on either Condition as subset confirmed that the Group Bias effect was significant in the Generalization Condition ($\beta = 0.393$, $SE = 0.11$, $z = 3.584$, $p < 0.001$) and the Segmental Overlap Condition ($\beta = 0.456$, $SE = 0.12$, $z = 3.659$, $p < 0.001$). Finally, we also observed a main effect of Condition, suggesting an overall SW-bias on the *kanon* continuum compared to the *voornaam* continuum ($\beta = -0.794$, $SE = 0.1$, $z = -7.923$, $p < 0.001$). An additional model with Block Order (i.e., whether participants received the Segmental Overlap or Generalization first) did not improve model fit.

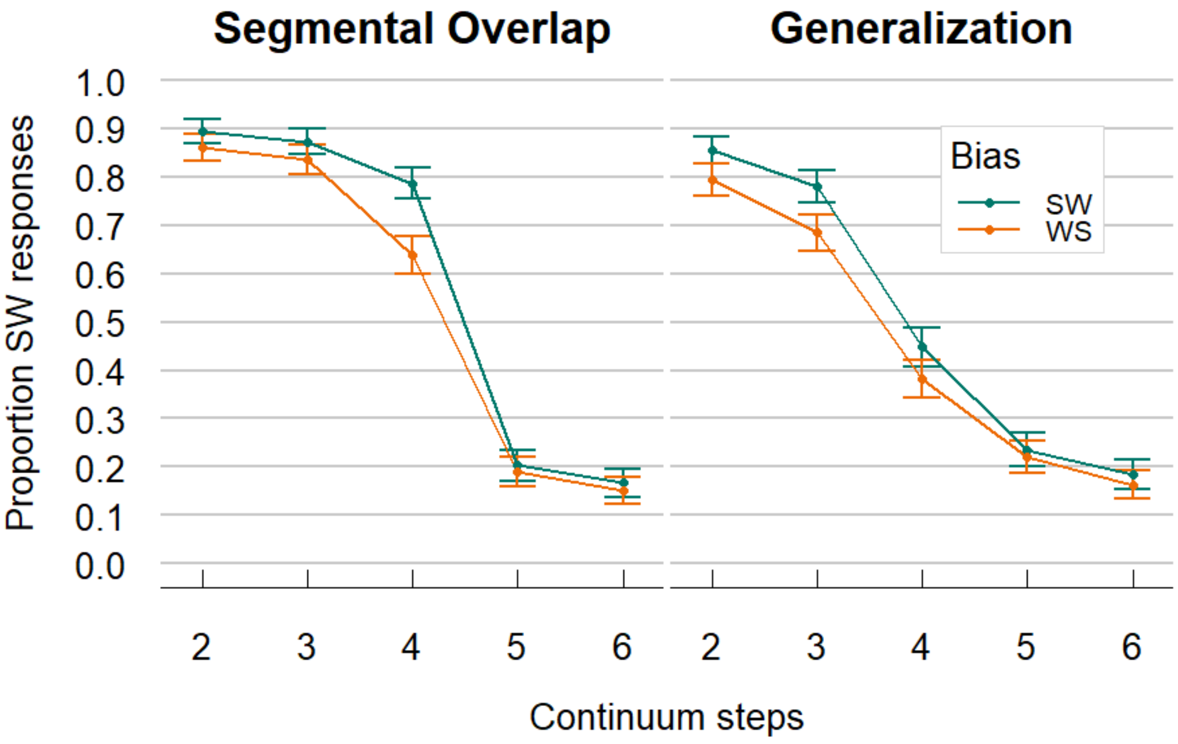

Figure 5. Results from Experiment 3: Recalibration driven by Beat Gestures. Comparison of the results from Segmental Overlap (left) and the Generalization Condition (right). The proportion of SW responses (i.e., stress on first syllable) is generally highest at step 2 (most SW-like) and lowest at step 6 (most WS-like) in both conditions. The differently colored lines show if participants were in the SW (green) or WS (orange) Bias group in the exposure phase. Bias group in Exposure affected the responses in the test phase in both conditions, giving evidence for recalibration. That is, people from

the SW Bias group consistently responded more SW-like, and participants from the WS Bias group responded more WS-like. This recalibration effect was present in both the Segmental Overlap Condition as well as the Generalization Condition. SW = strong-weak, stress on first syllable; WS = weak-strong, stress on second syllable. Error bars indicate a 95% confidence interval.

Thus, Experiment 3 demonstrates generalization of the recalibration effect in the perception of lexical stress. Participants who were exposed to tokens from the *CAnon – kaNON* continuum, with a beat gesture either on the first or the second syllable, responded differently in the test phase when asked to categorize a different audio-only *VOORnaam – voorNAAM* continuum. Disambiguating beat gestures on the first syllable in exposure led to more SW responses in test, while disambiguating beat gestures on the second syllable in exposure led to fewer SW responses in test, even when participants were tested on a different continuum. They thus demonstrated generalization of their recalibration to a segmentally different word.

General Discussion

The current study investigated recalibration of lexical stress perception driven by orthography (Experiment 1; as in Bosker (2022)) and by manual beat gestures the speaker produced while talking (Experiments 2-3). Across all three experiments, we found reliable evidence for recalibration of lexical stress perception. These recalibration effects emerged after mere passive exposure to the multimodal stimuli, in an online testing setup without control over participants' looking behavior or attention, and using as few as 24 acoustically ambiguous trials during exposure. Thus, we were able to replicate previous recalibration findings, driven by lexico-orthographic information (Bosker, 2022), with more variable and arguably more naturalistic phonetic continua. And, most strikingly, we demonstrate recalibration of lexical stress perception driven by simple up-and-down manual beat gestures.

These results highlight the importance of beat gestures and specifically their temporal alignment to the speech signal in audiovisual speech perception. Not only articulatory visual cues (i.e., lip movements) that are causally linked to the speech signal can affect speech perception (Bertelson et al., 2003; McGurk & MacDonald, 1976), but also other visual signals such as simple up-and-down gestures produced by the hands shape speech perception. Here we show that these effects of gestural timing go beyond immediate perception (Bosker & Peeters, 2021) and can lead to lasting changes to perceptual representations. This finding is consistent with recent models of speech-gesture integration, which highlight the multimodal nature of spoken communication (Holler & Levinson, 2019; Kita & Özyürek, 2003). Based on these models, we speculate that the timing of other types of gestures could also serve as recalibrators. Other types of co-speech gestures, such as iconic (Pouw & Dixon, 2019) and pointing gestures (Peeters, 2015; Pouw & Dixon, 2019), are also produced in synchrony with the spoken signal (see McNeill, 1992). As such, the present recalibration findings are unlikely to be restricted to beat gestures alone. However, these types of gestures may offer additional cues to disambiguate the acoustic signal. Iconic gestures convey a specific meaning, and pointing gestures can direct attention, referring to a specific object. Therefore, perceptual recalibration through other types of gestures may be driven by temporal alignment to the speech, as shown here, but possibly through other disambiguating cues as well.

Our results show how the use of beat gestures to recalibrate lexical stress perception could be a plausible explanation for suprasegmental adaptation in day-to-day communication, where beat gestures and speech co-occur frequently (in contrast to speech-disambiguating orthography). Moreover, our findings support the notion that gestures are processed and integrated automatically with speech (Kelly et al., 2010). We found recalibration effects even when participants were not instructed to pay attention to the beat gestures presented to them in exposure. More active tasks, for instance requiring comprehension of the presented words in a communicative context, might potentially even lead to larger recalibration effects. However, as our experiments have shown, explicit tasks are not necessary for beat gesture integration with speech. This is in line with previous research showing effects of beat gesture alignment in more implicit tasks, like shadowing and vowel length perception (Bosker & Peeters, 2021).

Lexical stress production is quite variable between people (Severijnen et al., 2021), so it is important to map out the mechanisms that listeners have at their disposal to adapt to variable lexical stress production. Kleinschmidt and Jaeger (2015) suggest at least two possible mechanisms to underlie recalibration. First, listeners may shift their perceptual category boundaries, such that one category is shifted to include the originally perceptually ambiguous acoustic space. Alternatively, participants could relax their decision-making criteria in general, accepting any non-standard cues to fit a given category. Our study was not designed to discriminate between these mechanisms and hence our results cannot disentangle them. However, prior evidence from Bosker (2022) speaks to this issue. In that study, some participants were presented with pseudoword stimuli in the exposure phase that were acoustically very similar to words. Yet, despite this acoustic similarity, listeners did not recalibrate their perception of lexical stress on these pseudowords (Bosker, 2022). This may be viewed as an argument against the involvement of general decision-making mechanisms as the sole basis of recalibration, which should apply equally to word- and pseudoword stimuli. Therefore, presumably a specific shift in perceptual boundaries is thus more likely than a general relaxation of decision-making criteria.

Current models of audiovisual recalibration (e.g., Kleinschmidt & Jaeger, 2015; Xie et al., 2023) do not address suprasegmental recalibration and/or visual gesture information. Still they might be able to explain our findings. Kleinschmidt and Jaeger's (2015) Ideal Adaptor Framework proposes that the perceptual system resolves ambiguity by statistical inference and thus drives adaptation. For example, when presented with an ambiguous sound (e.g., midway between /ba/ and /da/), together with a clear visual articulation (e.g., closing lips), one can infer the likely sound from prior experience (i.e., informed by the statistics about which sounds these mouth shapes usually co-occur with). In theory, the same process of inference could be used to recalibrate the perception of lexical stress.

When presented with acoustically ambiguous stress cues, together with a certain temporally aligned visual beat gesture, one could infer that these acoustic cues convey a certain stress pattern based on the prior experience that beat gestures usually accompany stressed syllables. Still, it is unclear to what extent the underlying processes responsible for segmental and suprasegmental recalibration are the same. Moreover, it is also unclear whether these models make different predictions for beat gestures, which are only relevant in their temporal alignment to speech and do not convey any phonetic information, unlike articulatory cues that are both time-aligned as well as phonologically informative. As our results that show that recalibration can be driven by beat gestures, models should address different sources of visual information and suprasegmental aspects of speech more specifically.

Another goal of the present study was to test for *generalization* of recalibration to novel and segmentally distinct words, not encountered during exposure. Our study provides mixed evidence for generalization of the recalibration effect. We did not find a generalization effect in Experiment 1 and only numerical evidence in Experiment 2. However, we found a statistically reliable generalization effect in Experiment 3, using a within-participant design with arguably less noise and increased statistical power. Note that this finding replicates the generalization outcomes reported in Bosker (2022), but this time with more naturalistic auditory phonetic continua. All in all, we interpret these findings as indicating that the generalization effect is more fragile than the within-item recalibration effect and possibly particularly sensitive to the stimuli used. Thus, we argue that recalibration of lexical stress can be generalized to novel words but presumably only under ideal circumstances (i.e., within-participant design and/or acoustic overlap between the words).

Generalization is usually taken as evidence of phonological abstraction. According to metrical phonology, in the case of lexical stress the abstraction is a metrical structure of a whole phrase that cues relative prominence (Ladd & Arvaniti, 2023; Pierrehumbert, 1980). When perceiving speech, phonetic information cues this metrical structure. In our study, the orthography (Experiment 1) and beat gestures (Experiment 2 & 3) provided additional information cueing a specific metrical structure for this speaker (i.e., acoustically ambiguous cues referring to either SW or WS). Experiment 3 showed that this cued structure can also be applied to a segmentally different disyllabic word. An interesting avenue for future research would be to investigate different and more complex metrical structures (e.g., polysyllabic words) and thus test the limits of generalization.

Another potential future topic of interest is to assess to what extent the recalibration effect, induced by beat gestures, that we consistently observed in the present study is speaker-specific (Eisner & McQueen, 2005). For instance, if a listener recalibrates their perception of speaker A, will the perception of speaker B change as well or remain unchanged? There is mixed evidence for speaker-specificity in recalibration of segmental speech. Some studies testing the recalibration of certain phonemes have found such speaker-specificity (e.g., Eisner & McQueen, 2005). These studies argue that generalization to different speakers is not beneficial unless there are indications that the pronunciation variation is driven by group-level factors (e.g., demographics). Other studies, testing different phonemes, found generalization across speakers (e.g., Kraljic & Samuel, 2006), which could facilitate processing of acoustic patterns that multiple talkers have in common. It is thus unclear to what extent recalibration of suprasegmental aspects of speech, such as lexical stress is speaker-specific. Further research could explore the limits of recalibration of lexical stress and investigate the influence of beat gestures on speech comprehension more broadly.

In sum, the results of all three experiments reported here consistently show evidence for recalibration of lexical stress. Simple flicks of the hand appear to have a lasting impact on speech perception. The mere alignment of beat gestures with speech can shape our perception of lexical stress and remain effective even when beat gestures are no longer present. The temporal alignment of gestures and speech conveys important information to a listener even in passive-viewing tasks. This highlights the importance of gesture-speech integration in face-to-face communication.

Data Availability

All experimental data, including scripts and stimuli are publicly available on OSF (https://osf.io/s3p6a/?view_only=e4e822e23a7440f2bd22a25bfb1dff95) under a CC-By Attribution 4.0 International license.

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 authors 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. The authors would also like to express gratitude to the members of the SPEAC research group, as well as the Sound Learning research group, in particular James McQueen, for their valuable feedback and comments on the study.

References

Aller, M., Mihalik, A., & Noppeney, U. (2022). Audiovisual adaptation is expressed in spatial and decisional codes. *Nature Communications*, *13*(1), 1. https://doi.org/10.1038/s41467-022-31549-0

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual Recalibration of Auditory Speech Identification: A McGurk Aftereffect. *Psychological Science*, *14*(6), 592–597. https://doi.org/10.1046/j.0956-7976.2003.psci_1470.x

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, *124*(2), 143–152. https://doi.org/10.1016/j.bandl.2012.10.008

Biau, E., & Soto-Faraco, S. (2015). Synchronization by the hand: The sight of gestures modulates low-frequency activity in brain responses to continuous speech. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00527

Bosker, H. R. (2022). Evidence For Selective Adaptation and Recalibration in the Perception of Lexical Stress. *Language and Speech*, *65*(2), 472–490. https://doi.org/10.1177/00238309211030307

Bosker, H. R., & Peeters, D. (2021). Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1943), 20202419. https://doi.org/10.1098/rspb.2020.2419

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* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/y9jck

Burg, E. V. der, Alais, D., & Cass, J. (2013). Rapid Recalibration to Audiovisual Asynchrony. *Journal of Neuroscience*, *33*(37), 14633–14637. https://doi.org/10.1523/JNEUROSCI.1182-13.2013

Cutler, A. (2008). Lexical Stress. In *The Handbook of Speech Perception* (pp. 264–289). John Wiley & Sons.

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In *Laboratory Phonology 10* (pp. 91–111). De Gruyter Mouton.

Dimitrova, D., Chu, M., Wang, L., Özyürek, A., & Hagoort, P. (2016). Beat that Word: How Listeners Integrate Beat Gesture and Focus in Multimodal Speech Discourse. *Journal of Cognitive Neuroscience*, *28*(9), 1255–1269. https://doi.org/10.1162/jocn_a_00963

Drijvers, L., & Özyürek, A. (2017). Visual Context Enhanced: The Joint Contribution of Iconic Gestures and Visible Speech to Degraded Speech Comprehension. *Journal of Speech, Language, and Hearing Research*, *60*(1), 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238. https://doi.org/10.3758/BF03206487

Gullberg, M., & Kita, S. (2009). Attention to Speech-Accompanying Gestures: Eye Movements and Information Uptake. *Journal of Nonverbal Behavior*, *33*(4), 251–277. https://doi.org/10.1007/s10919-009-0073-2

Holler, J., & Levinson, S. C. (2019). Multimodal Language Processing in Human Communication. *Trends in Cognitive Sciences*, *23*(8), 639–652. https://doi.org/10.1016/j.tics.2019.05.006

Hübscher, I., & Prieto, P. (2019). Gestural and Prosodic Development Act as Sister Systems and Jointly Pave the Way for Children’s Sociopragmatic Development. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01259

Jesse, A. (2021). Sentence context guides phonetic retuning to speaker idiosyncrasies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(1), 184–194. https://doi.org/10.1037/xlm0000805

Jesse, A., & McQueen, J. M. (2014). Suprasegmental Lexical Stress Cues in Visual Speech can Guide Spoken-Word Recognition. *Quarterly Journal of Experimental Psychology*, *67*(4), 793–808. https://doi.org/10.1080/17470218.2013.834371

Keetels, M., Schakel, L., Bonte, M., & Vroomen, J. (2016). Phonetic recalibration of speech by text. *Attention, Perception, & Psychophysics*, *78*(3), 938–945. https://doi.org/10.3758/s13414-015-1034-y

Kelly, S. D., Creigh, P., & Bartolotti, J. (2010). Integrating Speech and Iconic Gestures in a Stroop-like Task: Evidence for Automatic Processing. *Journal of Cognitive Neuroscience*, *22*(4), 683–694. https://doi.org/10.1162/jocn.2009.21254

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*(1), 16–32. https://doi.org/10.1016/S0749-596X(02)00505-3

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203. https://doi.org/10.1037/a0038695

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, *57*(3), 396–414. https://doi.org/10.1016/j.jml.2007.06.005

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–268. https://doi.org/10.3758/BF03193841

Kurumada, C., Brown, M., & Tanenhaus, M. (2012). Pragmatic interpretation of contrastive prosody: It looks like speech adaptation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*(34). https://escholarship.org/uc/item/6jw49594

Kushch, O., & Prieto, P. (2016). *The Effects of pitch accentuation and beat gestures on information recall in contrastive discourse*. https://doi.org/10.21437/SpeechProsody.2016-189

Ladd, D. R., & Arvaniti, A. (2023). Prosodic Prominence Across Languages. *Annual Review of Linguistics*, *9*(1), 171–193. https://doi.org/10.1146/annurev-linguistics-031120-101954

Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, *26*(10), 1457–1471. https://doi.org/10.1080/01690965.2010.500218

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2019). Listeners normalize speech for contextual speech rate even without an explicit recognition task. *The Journal of the Acoustical Society of America*, *146*(1), 179–188. https://doi.org/10.1121/1.5116004

McClave, E. (1994). Gestural beats: The rhythm hypothesis. *Journal of Psycholinguistic Research*, *23*(1), 45–66. https://doi.org/10.1007/BF02143175

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 5588. https://doi.org/10.1038/264746a0

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

McNeill, D. (2008). *Gesture and Thought*. University of Chicago Press.

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, *53*(4), 1551–1562. https://doi.org/10.3758/s13428-020-01514-0

Mitterer, H., Chen, Y., & Zhou, X. (2011). Phonological Abstraction in Processing Lexical-Tone Variation: Evidence From a Learning Paradigm. *Cognitive Science*, *35*(1), 184–197. https://doi.org/10.1111/j.1551-6709.2010.01140.x

Mitterer, H., & de Ruiter, J. P. (2008). Recalibrating Color Categories Using World Knowledge. *Psychological Science*, *19*(7), 629–634. https://doi.org/10.1111/j.1467-9280.2008.02133.x

Noppeney, U. (2021). Perceptual Inference, Learning, and Attention in a Multisensory World. *Annual Review of Neuroscience*, *44*(1), 449–473. https://doi.org/10.1146/annurev-neuro-100120-085519

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. https://doi.org/10.1037/0033-295X.115.2.357

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130296. https://doi.org/10.1098/rstb.2013.0296

Peeters, D. (2015). *A Social and Neurobiological Approach to Pointing in Speech and Gesture*. https://doi.org/10.13140/RG.2.1.3346.2883

Pierrehumbert, JB. (1980). The Phonology and Phonetics of English Intonation. *Doctoral Dissertation, MIT*. https://cir.nii.ac.jp/crid/1571698600473298432

Pouw, W., & Dixon, J. A. (2019). Quantifying gesture-speech synchrony. *Proceedings of the 6th Gesture and Speech in Interaction Conference*. https://doi.org/10.17619/UNIPB/1-815

Pouw, W., Harrison, S.J., & Dixon, J.A. (2020). Gesture-speech physics: The biomechanical basis for the emergence of gesture-speech synchrony. *Journal of Experimental Psychology - General*, *149*, 391–404. https://doi.org/10.1037/xge0000646

Radeau, M., & Bertelson, P. (1974). The After-Effects of Ventriloquism. *Quarterly Journal of Experimental Psychology*, *26*(1), 63–71. https://doi.org/10.1080/14640747408400388

Rietveld, T., & Heuven, V. J. van. (2009). *Algemene Fonetiek (3e geheel herziene druk)*. Bussum : Coutinho. https://repository.ubn.ru.nl/handle/2066/79395

Scarborough, R., Keating, P., Mattys, S. L., Cho, T., & Alwan, A. (2009). Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English. *Language and Speech*, *52*(2–3), 135–175. https://doi.org/10.1177/0023830909103165

Severijnen, G. G. A., Bosker, H. R., & McQueen, J. M. (2023). *Individual differences in lexical stress in Dutch: An examination of cue weighting in production*. the 5th Phonetics and Phonology in Europe Conference (PaPE 2023). https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_3530281

Severijnen, G. G. A., Bosker, H. R., Piai, V., & McQueen, J. M. (2021). Listeners track talker-specific prosody to deal with talker-variability. *Brain Research*, *1769*, 147605. https://doi.org/10.1016/j.brainres.2021.147605

Shattuck-Hufnagel, S., & Ren, A. (2018). The Prosodic Characteristics of Non-referential Co-speech Gestures in a Sample of Academic-Lecture-Style Speech. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01514

Swerts, M., & Krahmer, E. (2007). Acoustic effects of visual beats. *Proceedings of the International Conference on Auditory Visual Speech Processing (AVSP 2007)*, 252–257.

Ullas, S., Bonte, M., Formisano, E., & Vroomen, J. (2022). Adaptive Plasticity in Perceiving Speech Sounds. In L. L. Holt, J. E. Peelle, A. B. Coffin, A. N. Popper, & R. R. Fay (Eds.), *Speech Perception* (pp. 173–199). Springer International Publishing. https://doi.org/10.1007/978-3-030-81542-4_7

Ullas, S., Formisano, E., Eisner, F., & Cutler, A. (2020a). Audiovisual and lexical cues do not additively enhance perceptual adaptation. *Psychonomic Bulletin & Review*, *27*(4), 707–715. https://doi.org/10.3758/s13423-020-01728-5

Ullas, S., Formisano, E., Eisner, F., & Cutler, A. (2020b). Interleaved lexical and audiovisual information can retune phoneme boundaries. *Attention, Perception, & Psychophysics*, *82*(4), 2018–2026. https://doi.org/10.3758/s13414-019-01961-8

van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1483–1494. https://doi.org/10.1037/0096-1523.33.6.1483

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209–232. https://doi.org/10.1016/j.specom.2013.09.008

Xie, X., Jaeger, T. F., & Kurumada, C. (2023). What we do (not) know about the mechanisms underlying adaptive speech perception: A computational framework and review. *Cortex*. https://doi.org/10.1016/j.cortex.2023.05.003