Lexical and audiovisual bases of perceptual adaptation in speech

Lexical and audiovisual bases of perceptual adaptation in speech

Shruti Ullas

The projects described in this dissertation were funded by the NWO Gravitation consortium, Language in Interaction.

Printing: Ipskamp Printing, proefschriften.net
Cover images: shutterstock.com
ISBN: 978-94-6380-847-7

Lexical and audiovisual bases of perceptual adaptation in speech

Dissertation

to obtain the degree of doctor at Maastricht University,
on the authority of the Rector Magnificus Prof. dr. Rianne M. Letschert,
in accordance with the decision of the Board of Deans,
to be defended in public on Wednesday the 17th of June 2020 at 14.00 hours

by

Shruti Ullas

Supervisors

Prof. dr. Elia Formisano
Prof. dr. Anne Cutler (Western Sydney University)

Co-supervisor

Dr. Lars Hausfeld

Assessment Committee

Prof. dr. Bernadette Jansma (chair)
Prof. dr. Sonja Kotz
Prof. dr. Floris de Lange (Radboud University Nijmegen)
Prof. dr. James McQueen (Radboud University Nijmegen)

Contents

1

Introduction

Ullas, S., Bonte, M., Formisano, E., & Vroomen, J. (2020, forthcoming).
Adaptive Plasticity in Perceiving Speech Sounds.
Springer Handbook of Auditory Research: Auditory Cognitive Neuroscience.

Abstract

Listeners can rely on perceptual retuning and recalibration in order to make reliable interpretations during speech perception. Lexical and audiovisual (or speech-read) information can disambiguate the incoming auditory signal when it is unclear, due to speaker-related characteristics, such as an unfamiliar accent, or due to environmental factors, such as noise. With experience, listeners can learn to adjust boundaries between phoneme categories as a means of adaptation to such inconsistencies. Recalibration and perceptual retuning experiments use a targeted approach by embedding ambiguous phonemes into speech or speech-like items, and with continuous exposure, a learning effect can be induced in listeners, wherein disambiguating contextual information shifts the perceived identify of the same ambiguous sound. The following chapter will review recent and past literature regarding lexical and audiovisual influences on phoneme boundary recalibration, as well as theories and neuroimaging data that potentially reveal what facilitates this perceptual plasticity.

Key words: recalibration, perceptual learning, speech perception, phonetic processing, lexical processing, audiovisual speech, speech-reading

1 Introduction

Speech perception is seemingly easy and automatic to the listener, and it requires little to no effort to accomplish in most circumstances. While it may appear straightforward, a great deal of variability exists in the quality of the speech signal, which requires the listener to adapt to the novel characteristics of the encountered speech. The acoustic signal can differ significantly across speakers, often due to unfamiliar accents, the presence of noise, or speech rate. No two speakers will pronounce a phoneme in the exact same way, and even the same speaker may not produce a phoneme identically across multiple instances, yet listeners are effortlessly able to recognize what they are saying. Auditory quality can also vary within speakers, perhaps due to a cold or while speaking over the phone. Still, the listener is usually able to easily resolve these inconsistencies and understand what is spoken. In order to adapt to these irregularities, listeners can learn to reshape existing representations of speech sounds and categories to accommodate any possible variability.

Acoustics are not the only source of information capable of changing speech sound representations, as other contextual cues are also highly influential. Contextual features may be just as useful as auditory information, and possibly even more so. In a recent issue of *Acoustics Today*, Winn (2018) introduces some non-acoustic cues that impact what listeners perceive to hear, including visual cues, such as the lip movements of a speaker, as well as the listener's own lexical knowledge. These non-acoustic sources can also enable processes known as recalibration or lexically-guided perceptual learning. Contextual information can guide the retuning process of phoneme category boundaries, after continuous exposure to speech or videos of speech-like tokens, edited to contain ambiguous versions of a phoneme. Listeners can learn to incorporate these ambiguous sounds into the phoneme category itself, particularly when the sounds resemble already familiar phonemes.

Norris et al. (2003) termed this effect lexically-guided perceptual learning, and observed that with the help of lexical knowledge, listeners could learn to expand a phoneme category by integrating an ambiguous phoneme. Similarly,

Bertelson et al. (2003) identified a comparable effect as recalibration, where listeners utilized visual or speech-reading information to adjust the phoneme category boundary. The two discoveries were made close in time, and while Norris et al. (2003) used recordings of words as stimuli, Bertelson et al. (2003) relied on video recordings of syllables. Still, while the types of available contextual information differed between the two studies, the experimental designs and stimuli constructions were remarkably similar. Since then, in the literature on lexical influences, the resulting after-effect is often referred to as perceptual retuning or phoneme adaptation, while the studies on visual/speech-reading influences refer to the analogous effect as audiovisual recalibration.

In laboratory settings, both recalibration and perceptual retuning are typically measured in two phases, starting with an exposure phase and followed by a test phase (Kraljic and Samuel 2009, for an overview). In the approach of lexically-guided perceptual learning, exposure stimuli are composed of audio recordings of words, whereas exposure stimuli in audiovisual speech-reading experiments comprise videos highlighting a speaker's lip movements while pronouncing a syllable. Both types of stimuli contain edited audio, where one particular phoneme is replaced with an ambiguous sound halfway between two clear phonemes. For instance, speech stimuli containing /f/-sounds are replaced with a token halfway between /f/ and /s/. Listeners are presented with many examples of such edited stimuli in the exposure phase. During subsequent test phases, listeners hear the ambiguous sounds again, but without any lexical or visual context available, and respond with the phoneme they perceive to be hearing. Consequently, listeners become more likely to respond hearing the same phoneme that was replaced in the previously presented words or videos. In the case of the aforementioned example, the listener would now report hearing the ambiguous token as /f/ as well. This response pattern is understood to reflect recalibration or perceptual retuning, and is a result of the listeners learning to include the ambiguous sound as a part of that particular phoneme category.

Listeners in such experiments can also learn to perceive the same ambiguous phoneme, with no change in acoustic features, in opposing ways, depending on the bias of the surrounding context. A 50-50 /f/-/s/ blend can be learned as either /f/

or /s/ depending on the type of exposure the listener has undergone. Again, in the same example, if listeners were instead presented with speech stimuli that replaced all /s/-sounds with the same ambiguous token (the 50-50 blend of /f/ and /s/), listeners would be more likely to perceive the ambiguous sound as /s/ as well. With this approach, the contributions of visual and lexical information on speech perception can be disentangled from the auditory signal itself, since the exact same ambiguous tokens can be learned as different phonemes depending on the contextual cues. Perceptual retuning and recalibration studies also reveal how flexible the units of speech are, and how they can be adapted depending on the surroundings or the input received. These experiments illuminate non-acoustic contributions to speech perception, and what listeners rely on in addition to the acoustic signal itself, which again, tends to fluctuate greatly both within and across speakers.

This chapter will present an overview of the current literature regarding lexical (sect. 2.1) and audiovisual influences (sect. 3.1) on phoneme boundary recalibration, as well as some related works on selective speech adaptation (sect. 3.2). Changes over time (sect. 2.2), generalization over speakers and sounds (sects. 2.3, 3.3), and other features (sect. 2.4) will also be discussed, as well a comparison between lexical and audiovisual perceptual learning (sect. 4). Theories and neuroimaging studies that may explain the underlying mechanisms of recalibration will also be reviewed (sect. 5), followed by a final conclusion and summary (sect. 6).

2 Lexical Knowledge and Auditory Perception

2.1 Introduction to Lexically-Guided Perceptual Learning

As mentioned earlier in the introduction (sect.1), top-down lexical knowledge can assist listeners in interpreting unclear speech. To investigate this, some researchers have used noise-vocoded or degraded speech stimuli that systematically distort frequency and amplitude components of the speech (Davis et al. 2005). Others have studied how listeners adapt to accented speech (Clarke and Garrett 2004; Bradlow and Bent 2008), how listeners adapt to non-native speech in noise (Lecumberri et al. 2010), as well as how lexical knowledge supports

understanding accented speech (Maye et al. 2008). A review by Holt and Lotto (2008) describes the various ways in which listeners can build links between acoustic information and linguistic representations. Prior to many of these studies, the discovery of what is now known as the Ganong effect (Ganong 1980) established the specific influence of lexical information on speech sound perception. Ganong (1980) showed that listeners were likely to report hearing words even when exposed to auditory stimuli that were edited to begin with ambiguous sounds. Listeners who heard the word "?eep," where the /?/ sound was acoustically halfway between /d/ and /t/, were likely to interpret the stimulus in the form of a word, such as "deep," rather than "teep." The same held true in the opposite direction, when the same ambiguous token replaced /t/ in recordings of words beginning with /t/, such as "?each." Again, listeners were likely to report hearing a word, such as 'teach', rather than the non-word version, "deach." In essence, listeners were not hindered by the unclear auditory information and were still able to infer the intended words.

Similar to the Ganong effect, the findings of Norris et al. (2003) revealed how lexical information could not only affect perception of speech stimuli, but could also reshape speech sound representations. Native Dutch speakers performed a lexical decision task while listening to audio recordings of Dutch words, some of which typically ended in /f/, such as "witlo??" (*witlof,* meaning chicory), and "drui??" (*druif,* meaning grape), where all /f/-sounds were replaced with an ambiguous token halfway between /f/ and /s/. During the following test phase, where listeners responded to a continuum of sounds ranging from more /f/-like to more /s/-like, they were likely to report a significantly greater number of tokens as /f/-sounding. In contrast, another group of participants conducted the same lexical decision task while hearing words, but in contrast, these words typically contained /s/ (such as *radijs* and *relaas,* meaning radish and account) and were spliced with the same ambiguous token in the place of /s/, and the opposite pattern of results was found. These listeners responded to the same continuum of /f/ to /s/ sounds during the test phase, and were more likely to report hearing more of the sounds as /s/-like. A third control group heard pseudo-words containing the ambiguous phoneme to test whether the absence of any lexical information could impact

subsequent categorization, and this group showed no bias toward either phoneme during the test phase.

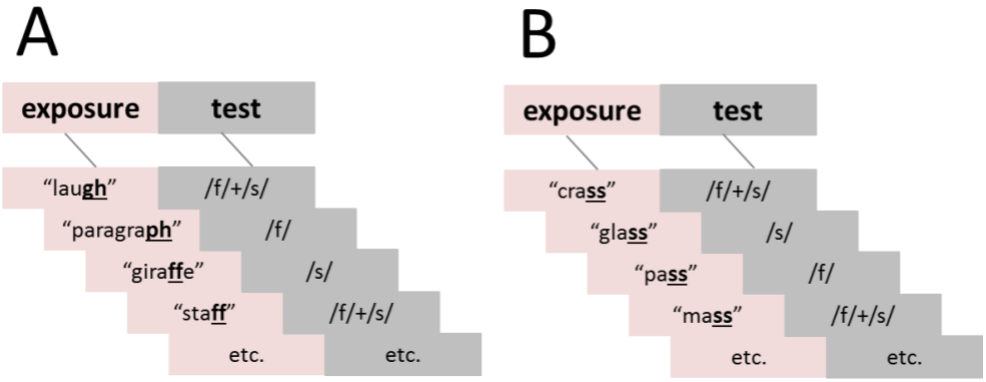


Figure 1. Schematic example of lexical retuning procedure. Exposure phases consist of recordings of words, ending with an ambiguous phoneme blends (such as a /f/-/s/ blend). One group may undergo ambiguous /f/-final exposure (in A) or ambiguous /s/-final exposure (in B), followed by a categorization task on the ambiguous blend along with other similar sounds, where listeners report what they perceived (/f/ or /s/).

Together, these results built further upon the lexical effect first described by Ganong, and illustrated how lexical knowledge impacted the participants' perception in two ways. First, during the exposure phase, the words containing the ambiguous sounds were still perceived as words and nearly indistinguishable from unedited words. Then, in the test phase, listeners categorized ambiguous sounds of a continuum and were prone to hearing the continuum sounds resembling the phoneme replaced in the prior exposure phase. That is, listeners were likely to perceive the ambiguous token as /f/ after exposure to f-final words containing said token. Thus, phoneme categories boundaries were found to be flexible, as listeners adjusted the boundary between two phonemes using their lexical knowledge. The authors proposed that the results mirrored what listeners may be doing in response to an unfamiliar accent, by shifting a category boundary to make room for the pronunciation of the newly encountered speaker (this will be discussed more in sect. 2.3).

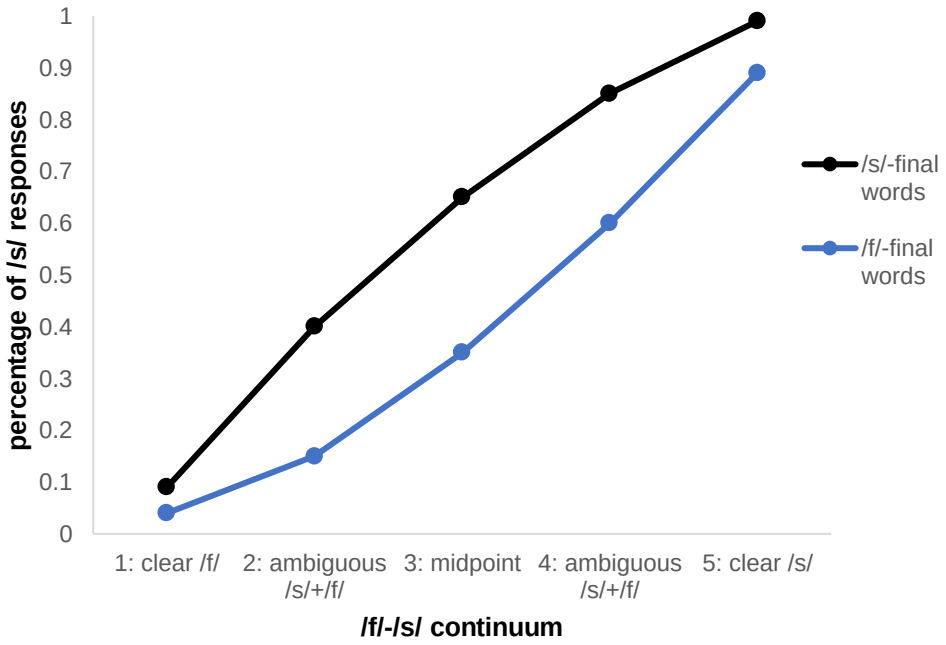


Figure 2. Example results of lexical retuning study**.** As in Norris, McQueen, Cutler (2003), a group exposed to ambiguous /s/-final words would be likely to perceive more /s/ on a /f/-/s/ continuum (black line) than a group exposed to ambiguous /f/-final words (blue line).

2.2 Perceptual Retuning over Time

Since Norris et al. (2003), later studies of perceptual learning explored the other attributes of this effect, such as the duration of time for which the retuning effects could last in the listener, as well as if these changes were permanent or if the categories returned to their previous state. Kraljic and Samuel (2005) used nearly the same approach as Norris et al. (2003), testing native English speakers using English words ending in either /s/ or /ʃ/ (the "sh"-sound in shoe) instead. After a 25-minute delay, participants were tested on a continuum from /s/ to /ʃ/, and their responses reflected the shift induced by the preceding exposure phase (i.e. more /s/ responses after /s/-final words, or more /ʃ/ after /ʃ/-final words). Despite the delay, the listeners could still retain the newly learned phoneme boundary position.

Eisner and McQueen (2006) also measured perceptual learning effects in subjects after a longer delay, where participants completed one test immediately after exposure, and also returned 12 hours after the exposure to complete the test

phase again. The exposure phase was slightly altered from the original version by Norris et al. (2003) and consisted of words with ambiguous segments, all embedded into a short story. The potential confound of sleep was also accounted for, as one group waited 12 hours during the day to be retested, while another group waited 12 hours overnight, and returned for the second test phase after they had slept. Both groups still maintained retuning effects after the 12 hour delay, with or without sleeping. As such, perceptual learning is seemingly unaffected by long gaps between exposure and test, which suggests that lexically-guided perceptual learning is considerably stable over time.

2.3 Does Perceptual Retuning Generalize?

Although lexically-driven perceptual learning appears to be quite robust, research has also identified the limitations of such learning. For example, perceptual learning tends to be restricted by the stimuli, particularly by the speakers of the tokens. Kraljic et al. (2008a) compared effects of speaker characteristics on perceptual learning in American participants, with an idiosyncratic pronunciation versus an accent commonly known to the participants. The idiosyncrasy, or speaker-specific version, was generated by placing an ambiguous /s/-/ʃ/ sound before all consonants in the word stimuli, whereas the accented version only placed the ambiguous sound before an occurrence of /tr/ (such as /s/ in *string*), as is typical of many regional American accents. Boundary retuning was successful in the latter group that was exposed to the accented speech, but was not detected in the former, idiosyncratic group. In other words, knowledge of reasonable and unrealistic deviations, which may be implicit or explicit, also seems to impact perceptual learning. Similarly, native English participants who heard exposure stimuli in English by a speaker with a Mandarin accent were more likely to generalize retuning to another acoustically-similar Mandarin-accented speaker (Xie and Myers 2017), even if exposure tokens were from multiple speakers with the same accent.

Notably, not only does acoustic similarity affect speaker-specificity of perceptual retuning, but it can be also affected by the phoneme pair used in the

experiment. Eisner and McQueen (2005) had two groups of participants undergo exposure to Dutch words containing either an ambiguous /f/ or /s/ spoken by one speaker, but were tested on a continuum of /f/-/s/ sounds by a different speaker. Participants did not show the retuning effect when tested with the continuum by the novel speaker, so responses to the items on the continuum did not show a shift towards any particular phoneme. Thus, the authors concluded that the participants treated the sounds contained in the exposure stimuli as an idiosyncrasy, so it was tied specifically to the speaker of the ambiguous sounds and did not generalize to ambiguous sounds by a different speaker.

Kraljic and Samuel (2007) also addressed a possible discrepancy in generalization to new speakers based on phoneme types. Listeners who were exposed to words containing ambiguous /d/ or /t/ (plosives or stop consonants) sounds could generalize retuning to the same tokens of a new speaker during the test phase, translating to a shift in categorization responses towards the phoneme replaced in the prior exposure phase (i.e. more /d/ responses after exposure to /d/-words replaced with /d/-/t/ blend). However, those who were exposed to words spliced with ambiguous /s/ or /ʃ/ (fricatives) could not generalize any retuning to a new speaker, so no shift was found in categorization responses during the test phase. Evidently, perceptual learning may not always be constrained by the speaker, and depending on the type of phoneme pair used, it may also be token-specific.

Just as there are mixed results regarding generalization of retuning between speakers, it also not straightforward as to whether perceptual learning can transfer across phonemes. Kraljic and Samuel (2006) saw that recalibration could generalize between pairs of plosives or stop consonants, particularly between /d/-/t/ and /b/-/p/. During the exposure phase, listeners heard words containing either an ambiguous /d/ or /t/, but during the test phase, they responded to both a /d/-/t/ continuum as well as a /b/-/p/ continuum. Participants were able to generalize recalibration to the /b/-/p/ continuum in the same direction of voicing, such that participants who heard words with an ambiguous /b/, were more likely to report hearing a greater amount of /b/ along the /b/-/p/ continuum, as well as more /d/ responses during an additional test phase on a continuum of /d/-/t/. Responses to both continua reflected a shift in the same acoustic direction as the exposure phase.

Mitterer et al. (2013) also explored phoneme specific retuning by creating exposure stimuli using Dutch words ending in an approximant /r/ (the /r/ in *red*) or a dark /l/ (the /l/ in *pool*). Participants showed retuning effects during a test phase with a continuum of the versions of /r/ or /l/ they previously heard during exposure, but could not generalize to other allophones, or phonetic neighbors of /r/ and /l/, such as a trill /r/ (which is not part of the phonology in American English, but is closest to the /r/ in *better*) or a light /l/ (the /l/ in *leaf*). Once again, the specificity of retuning seems to be partially dependent on the acoustic features of the phoneme pair being learned.

Overall, results regarding generalization of lexically-driven perceptual retuning are complex. It appears that retuning is often phoneme- and speaker-specific, but this is not always the case, as it is also contingent on the specific phoneme pair used. Generalization to a new speaker is more likely to occur if the phoneme boundary is adjusted between two plosives and not between fricatives. Perceptual retuning effects upon plosives or stop consonants are also more likely to extend to other plosives, but again, are unlikely to do so for fricatives or approximants. Acoustic similarity also plays an important role as to whether retuning effects can be applied to new sounds.

2.4 Other Attributes of Perceptual Retuning

Most studies of the lexically-guided perceptual learning studies described throughout sect. 2 are two-fold. They typically start with an exposure phase, with words containing one particular ambiguous phoneme, presented along with other filler words and pseudo-words. Listeners are also often asked to perform a lexical decision task during this exposure phase, in order to maintain their attention. This is followed by a categorization task, or the test phase, on a continuum between two clear phonemes with the aforementioned ambiguous phoneme in between. However, this design is not always used, and other similar designs can still lead to measureable retuning effects. McQueen et al. (2006b) concluded that perceptual learning is not dependent on a lexical decision task during the exposure phase. Instead, the lexical decision task was replaced with a simple counting task, and

learning effects remained intact. However, a more recent study by Samuel (2016) suggested that targeted distractions during exposure that can prevent access to the lexicon are detrimental to perceptual retuning. In this study, listeners heard two voices only separated by 200 ms during exposure, of words containing an ambiguous /s/-/ʃ/ phoneme by a male speaker, and irrelevant words by a female speaker, and were asked to perform a lexical decision task on the male speaker, or to count the number of syllables spoken by the female speaker. Listeners who attended to the female speaker showed no recalibration during subsequent testing, however, when the voices were separated by 1200 ms, recalibration effects were reinstated. Similarly, listeners were also unable to undergo learning in the presence of background noise (Zhang and Samuel 2014), suggesting that recalibration cannot be performed automatically and requires attentional resources. But attention alone is also not enough to induce retuning, as listeners can still account for potentially transient characteristics of a speaker. In a creative design by Kraljic et al. (2008b), listeners viewed stimuli of a speaker with a pen in their mouth while pronouncing words dubbed with an ambiguous phoneme. These listeners did not show retuning during the subsequent test phase, implying that listeners also acknowledge temporary atypical pronunciations of a speaker before adjusting phoneme representations.

Attention aside, the prototypical test phase, most often a continuum of sounds between two phonemes, is also not a requisite to detect perceptual retuning effects. Effects were still preserved when test phase items were replaced with minimal word pairs ending in an ambiguous phoneme (McQueen et al. 2006a). Participants were then more likely to hear one of the two words of the pair, predicated by the prior exposure phase. For instance, after exposure to words with an ambiguous /f/ (such as *paragraph*, ending with an /f/-/s/ blend) participants were likely to hear "knife" rather than "nice" when presented with "kni-", ending in the same /f/-/s/ blend. The effect was observed in the opposite direction when listeners were presented with /s/-words ending in the ambiguous token during the exposure, In the same example, listeners were more likely to hear "nice".

Even fully intact lexical information is not a necessity for retuning to occur, and implicit knowledge of phonotactic information, or the rules within a language

regarding allowable phoneme combinations, can be sufficient (Cutler et al. 2008). Here, exposure stimuli were phonotactically-valid pseudo-words containing an ambiguous phoneme. Perceptual retuning can also be observed with other known phonemes that are acoustically related, such as /θ/ (represented as theta, or the "th"-sound in thing) in place of /s/ or /f/, instead of the oft-mentioned ambiguous phoneme (Sjerps and McQueen 2010). Again, the plausibility of the acoustic shift can determine whether retuning is induced or not.

Thus, the exposure and test phases do not necessarily have to follow one particular procedure for phoneme boundary retuning, but all of the studies discussed within section 2, as well as most of the classical studies of lexically-driven perceptual retuning have focused on native listeners. More recent works have also studied non-native listeners, and retuning can take place in non-native listeners as well. Native Dutch speakers with high proficiency in English also showed perceptual learning effects in response to English stimuli spoken by a British English speaker (Drozdova et al. 2016). Native German speakers of Dutch were also observed to undergo retuning effects in response to Dutch stimuli, at levels comparable to native Dutch speakers (Reinisch et al. 2013). However, proficiency in the second language can also determine whether recalibration can occur, as a group of native Arabic speakers with lower English proficiency than another group of native Hebrew speakers showed no retuning effects with English phonemes, while the latter group did (Samuel and Frost 2015).

Section 2 summarized the seminal studies as well as some more recent findings about lexically-guided perceptual learning. These effects are potentially long-lasting but may not generalize to new speakers. Non-native speakers are also capable of demonstrating learning effects, but this may be mitigated by the listener's proficiency in the second language. Generalization to new speakers and to other phonemes is mitigated by the type of phoneme category being adjusted. Retuning effects may be applied from stop consonants or plosives to other phonemes within this classification, but this is less likely for fricatives or approximants. While lexical knowledge is primarily driving the subsequent learning, acoustic features still place constraints on what can and cannot be extended to other speech sounds.

3 Audiovisual Information and Speech

3.1 Introduction to Audiovisual Recalibration

Visual or speech-read information, much like lexical information, can also provide clarity when the available acoustics are unclear. Speech-reading can be relied upon if noise is present (Sumby and Pollack 1954), and also significantly alter what listeners perceive to hear. McGurk and MacDonald (1976) made the groundbreaking discovery that participants who viewed videos of a speaker pronouncing the syllable /gaga/, dubbed with audio of the syllable /baba/, perceived an entirely new percept, and reported hearing /dada/. Bertelson et al. (2003) extended this finding, and detected aftereffects on categorization responses following exposure to McGurk-like stimuli. Again, not only did speech-reading influence the perception of incongruent audiovisual tokens, but continuous exposure led to responses biased by the visual/speech-reading information. Much like the approach used by Norris et al. (2003) described in sect. 2, participants first underwent an exposure phase, where they viewed audiovisual stimuli of a speaker's lip movements while pronouncing /aba/, dubbed with audio of an ambiguous phoneme halfway between /aba/ and /ada/. During a subsequent test phase, participants only heard the audio token of the ambiguous phoneme and its two neighbors from a continuum, and were more likely to report them as /aba/-sounding. Unlike Norris et al. (2003), a within-subjects design was used, and the same group of participants also viewed videos of the speaker pronouncing /ada/, but dubbed with the same ambiguous token. In this case, participants were more likely to report hearing the token as ada/ during the test phase.

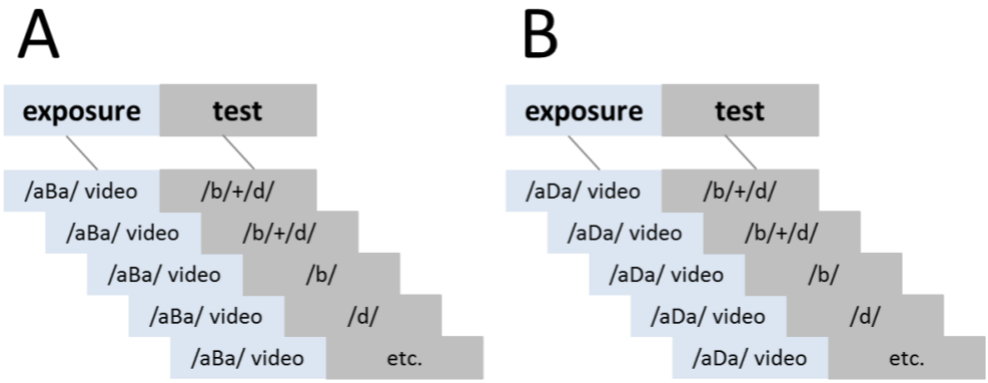


Figure 3. Example schematic of audiovisual recalibration procedure. Audiovisual recalibration studies have typically tested effects in both directions within participants. Participants are presented with exposure blocks, containing videos of a speaker pronouncing a syllable, such as /aba/ (in A) or /ada/ (in B) containing an ambiguous phoneme (/b/-/d/ blend). Following the exposure blocks, participants are presented with the ambiguous token (/b/-/d/ blend plus other similar sounds) and asked to respond with what they hear.

In a follow-up experiment, listeners were exposed to congruent stimuli, or clear audio of /aba/ combined with lip-movements of /aba/, and the same for an audio and video combination of /ada/. These unambiguous stimuli showed the reverse effect of the recalibration experiment and led to selective speech adaptation (Eimas and Corbit 1973). As a result of said selective speech adaptation, participants made fewer /aba/ responses to the ambiguous sounds if exposed to clear /aba/ tokens, and similarly gave fewer /ada/ responses after exposure to clear /ada/ tokens. This response is unlike recalibration, where participants who listen to ambiguous sounds during the exposure phase then become more likely to report hearing the phoneme being biased for by the lip-movements of the speakers (i.e. ambiguous audio coupled with video of /aba/ leading to more /aba/ responses during the test phase). Selective speech adaptation will be discussed in more detail in the next section (3.2).

3.2 Audiovisual Recalibration and Selective Speech Adaptation

Prior to studies of audiovisual recalibration, a perceptual learning effect known as selective speech adaptation was discovered (Eimas and Corbit 1973) and has also been helpful for understanding the building blocks of speech perception.

Recalibration and selective speech adaptation share considerable overlap, especially in terms of their experimental design, but are also distinct in their interpretations. Both styles of experiments use a similar two-part procedure with an exposure and test phase. Unlike recalibration, which typically uses ambiguous sounds, selective speech adaptation relies on exposure to clear sounds. While recalibration experiments lead to an increase in responses of the phoneme indicated by the videos during exposure, selective adaptation results in a reduction. For example, listeners repeatedly exposed to tokens of a clear /ba/ become less likely to perceiving /ba/ when given a categorization task on a /ba/-/da/ continuum. Selective speech adaptation is thought to reflect a fatigue effect, where listeners become desensitized to the auditory token during the exposure phase. The listener then becomes more sensitive to the acoustic differences in other similar sounds, and thereby reports hearing the ambiguous tokens as the phoneme opposing the preceding exposure phase. The original study of selective speech adaptation (Eimas and Corbitt 1973) relied on solely auditory stimuli, but later studies measured the same effects when exposure stimuli were coupled with videos of a speaker's lip movements, as Bertelson et al. (2003) reported. These unambiguous, or congruent audiovisual stimuli, also led to fewer responses of the phoneme presented in the test phase, as described in sect. 3.1.

Selective speech adaptation and recalibration are often discussed together, as they both reflect a change in auditory perception, following an exposure phase to syllables or speech sounds. Just as the response patterns of the two phenomena go in opposite directions, and the two differ in numerous other ways as well. Vroomen and colleagues have compared an audiovisual form of selective speech adaptation to recalibration, and have found that the overall build-up and dissipation also tend to differ (Vroomen et al. 2006). The number of exposure trials has been found to share a log-linear relationship with selective speech adaptation, as the effect was observed to increase as exposure trials accumulate, whereas recalibration was found to have a curvilinear relationship in relation to the number of exposure trials, as it steadily increased until eight exposure trials, but reduced with additional exposure. Recalibration and selective speech adaptation are also differentially affected by the number of test trials, as visual recalibration effects are

short-lived and can be present only up until approximately six test trials, while selective speech adaptation effect can be continuously sustained for up to 60 test items (Vroomen et al. 2004).

Vroomen and Baart (2009b) also compared recalibration and selective speech adaptation in groups that viewed audiovisual sine-wave speech tokens as speech-like versus non-speech-like. Sine-wave speech (SWS) is constructed by starting from clear speech but stripped down until approximately three sinusoids that follow the central frequency and amplitude of the first three formants remain. These stimuli are often unintelligible unless listeners are explicitly told that the sounds have been extracted from actual speech. In this experiment, all of the ambiguous and clear sounds typical of recalibration and selective speech adaptation studies were replaced with SWS versions, so a continuum between two clear phonemes was converted into SWS. For exposure phases, these SWS sounds were still paired with videos of a speaker's corresponding lip movements, but were presented without video for test phases. One "speech-mode" group viewed ambiguous SWS tokens paired with videos, that identified the tokens as /onso/ or /omso/, and showed recalibration effects. A "non-speech mode" group viewed the same stimuli but categorized the ambiguous SWS tokens as "1"' or "2", and did not show a recalibration effect, so a "speech-mode" did impact any possible recalibration. In contrast, for selective speech adaptation, participants viewed videos coupled with endpoint SWS tokens (rather than ambiguous), and adaptation effects were observed. In this instance, listeners who performed a categorization test on SWS-versions of the ambiguous tokens heard them as the opposite phoneme to the one biased for by the preceding exposure (i.e. hearing more /omso/ after exposure to SWS-versions of a clear /onso/ paired with video). Selective speech adaptation was still measureable in another non-speech mode group, who underwent the same types of exposure, but categorized the subsequent test phase ambiguous sounds as 1 or 2. Essentially, selective speech adaptation was unaffected by either set of labels, so "speech-mode" had no impact and listeners still adapted accordingly. The awareness of speech-like qualities was crucial for successful recalibration, but selective speech adaptation was not hindered by this lack of this awareness. While recalibration and selective speech adaptation can reshape speech

sound representations, based on these comparisons, it appears the two may be controlled by distinct but related substrates. The authors concluded that audiovisual recalibration may emerge from speech and language networks while selective speech adaptation is purely a bottom-up process that does not require higher-level feedback. Potential neural mechanisms will be discussed in more detail in sect. 5.

3.3 Specificity of Audiovisual Recalibration

Whether recalibration can be generalized has been addressed with regard to audiovisual information as well, just as it has with lexical context. Audiovisual recalibration tends to be token-specific (Reinisch et al. 2014), as exposure to either visual /aba/ or /ada/ tokens dubbed with ambiguous audio had no effect on listeners' categorization of continua of either /ibi/-/idi/ or /ama/-/ana/ sounds during test. As such, audiovisual recalibration appears to be constrained by the acoustics features, as learning did not transfer to other phonemes, or even to the same phonemes paired with different vowels. The ear itself can also limit recalibration (Keetels et al. 2016), as in one study, the effect was optimal if exposure and test stimuli were presented into the same ear, but was diminished for test stimuli presented into the opposite ear, and locations in between resulted in a gradient of responses as the presentations moved further away from the original ear. The authors argue that this is further evidence that recalibration is strongly tied to the token and context, and the encoding process even accounts for the exact location of the presented sound (neural mechanisms will be addressed further in sect. 5). Notably, listeners also have the capacity to recalibrate each ear in opposite directions using the same ambiguous sounds, that is, one ear recalibrated towards /aba/, the other towards /ada/, with test sounds presented into the corresponding ears of the exposure phase (Keetels et al. 2015). Again, this outcome seems to be in favor of the argument that recalibration is context-specific.

While audiovisual recalibration may be restricted in some respects, it is not necessarily specific to the speaker, as listeners can recalibrate to another speaker's pronunciation of the same phoneme, although to a substantially lesser extent

compared to the speaker during exposure (van der Zande et al. 2014). Still, recalibration is generally maximal in response to the sound used during exposure, which suggests that it generally tends to be constrained by the acoustic features of the exposure sound.

It is also worth noting that the studies described in this section so far have focused on consonant contrasts, but a recent study (Franken et al. 2017) has found that audiovisual recalibration may also be possible using a vowel contrast pair of /e/-/ø/. The majority of these studies have also been centered on adults, but audiovisual recalibration can also be adopted early in life, and has been observed in children as young as eight years old. van Linden and Vroomen (2008) measured recalibration effects in two groups of children and determined that children at eight years old could recalibrate with audiovisual stimuli, but children at five years old could not, so the ability may be developed within this window of three years. Dyslexia does not pose a limitation either (Baart et al. 2012), as adults with dyslexia were compared with fluently-reading adults, and the dyslexic group showed no deficit in their ability to recalibrate. This finding was especially remarkable, given that children with dyslexia often experience difficulties in speech-reading (van Laarhoven et al. 2018).

Section 3 described audiovisual recalibration, originally described by Bertelson et al. (2003), and its various attributes. Later studies by Vroomen and colleagues have established the general build-up and dissipation, as well as similarities and differences with another perceptual learning effect, called selective speech adaptation. Audiovisual recalibration tends to both build up following a few exemplars during exposure and diminish with increasing numbers of test items as well. In contrast, selective speech adaptation requires much longer exposure phases, but subsequent effects can last for longer durations. Recalibration also tends to be token- and context-specific, even to the extent that listeners can recalibrate each ear in opposite directions. It also does not easily generalize to other speakers, phonemes, or to other similar instances of the same phoneme, so it is considerably restricted by the acoustic features present during exposure. Nevertheless, it has shown to be utilized by a variety of listeners, including children

and adults with dyslexia, and remains to be a helpful tool for listeners when the auditory signal is inadequate.

4 Audiovisual Versus Lexical Recalibration

Sections 2 and 3 have discussed audiovisual recalibration and lexical retuning separately, but because they share several overlapping aspects, it is worth examining the two together as well. In realistic situations, listeners are likely to encounter lexical and visual information simultaneously, so it is possible that these two sources may interact while influencing speech perception. The designs of the two types of experiments share overlap in many respects, with exposure phases consisting of stimuli embedded with ambiguous phonemes, followed by forced-choice test phases where the ambiguous sounds are presented without lexical or speech-reading contextual cues. Even the response patterns between the two original studies by Bertelson et al. (2003) and Norris et al. (2003) paralleled each other, so it may appear that phoneme categories are affected comparably by both audiovisual and lexical information. Brancazio (2004) probed the influence of lexical and speech-reading information in audiovisual speech perception but found that speech-reading exerted a stronger influence on phoneme categorization. This effect was sustained for both fast and slow responses, while lexical information showed a weaker effect and was observed most often during slower responses.

Based on this, van Linden and Vroomen (2007) proposed that audiovisual information may induce recalibration more effectively than lexical cues, and conducted a study comparing lexical and audiovisual recalibration to test this hypothesis. Two forms of recalibration were compared in native Dutch speakers using a /p/-/t/ phoneme contrast. One group was exposed to lexical stimuli, which consisted of audio Dutch words typically ending in either /op/ or /ot/ (such as *bioscoop*, or movie theater, and *idioot*, or idiot), with all endings replaced by an ambiguous token halfway between /op/ and /ot/. Another group was exposed to audiovisual stimuli, comprised of videos of pseudo-words, where lip-movements indicated a /op/ or /ot/ ending, and were also dubbed with audio of the ambiguous phoneme at the end of the token. Participants were also exposed to both /op/- and

/ot/-biased stimuli, to explore whether they could recalibrate in both directions of the phoneme pair, such that half of the exposure blocks would induce a bias towards /p/, and the remaining half were biased towards /t/. Test phase judgments indicated that recalibration was indeed successful in both groups and in response to both phonemes as well. As the authors originally proposed, audiovisual information was largely more effective in producing recalibration than lexical information. The discrepancy may have resulted from the inherent differences in the stimuli and the processing levels affected, as lexical information might only induce a phoneme preference with the help of top-down influences, whereas the incoming audiovisual information already contained a visual bias towards one phoneme. Theories of top-down and bottom-up processing will be discussed in more depth in Section 5.

In contrast to previous studies on lexical retuning, both audiovisual and lexical recalibration dissipated at the same rate. Although audiovisual recalibration has been known to dissipate relatively quickly (Vroomen et al. 2007b), other studies have found that lexically-guided perceptual learning can be long lasting (Eisner and McQueen 2006). Participants in the van Linden and Vroomen (2007) study were flexibly adjusting the phoneme boundary back and forth between the two phonemes, throughout the duration of the experiment, so the faster dissipation of lexical recalibration may have resulted from constant switching between the two phonemes. However, this was refuted in a follow-up experiment with a between-subjects design, where each group of participants were only exposed to one phoneme-modality combination, and no improvements to recalibration were found. Still, the chosen phoneme pair is also worth noting, as plosives or stop consonants such as /p/ and /t/ may be more amenable to adjustment than fricatives (as mentioned in sect. 2), such as /f/ and /s/ (Kraljic and Samuel 2007). Overall, lexical and audiovisual recalibration seem to be markedly similar, although the pathways supporting them may not be identical, and may only overlap.

The two types of perceptual learning also tend to differ in their stability, as lexical retuning has been shown to be stable over time, but audiovisual recalibration can be more susceptible to decay with the passage of time. After a standard exposure phase, participants were tested after a 24-hour gap and effects

had dissipated (Vroomen et al. 2007a), even if participants were tested both immediately after the exposure phase and again 24-hours later (Vroomen and Baart 2009b). Audiovisual recalibration effects have also been shown to diminish within the test phase, as responses that corresponded with the preceding visual exposure (such as /b/ responses after viewing /aba/ videos) were maximal at the start of the test phase, but consistently decreased as the test phase progressed (Vroomen and Baart 2009b). In contrast, lexical retuning effects can be preserved throughout longer testing sessions, often containing approximately 30 test items (Kraljic and Samuel 2009), or up to 12 hours later (Eisner and McQueen 2006). As mentioned earlier in sect. 2, lexical retuning is capable of generalizing to new speakers and certain phonemes, while audiovisual recalibration is most often token-specific and may generalize if the critical phonemes are plosives/stop consonants.

Despite these differences, lexical retuning and audiovisual recalibration share many similarities in terms of how the subsequent effects are exhibited, how the experiments measuring them are designed, as well as the resulting response patterns to presentations of ambiguous sounds. Both approaches are useful for adapting to speech in noise, even if their origins and functions may differ.

5 Theoretical and Neural Explanations of Recalibration

5.1 Theories of Speech Perception

The mechanisms that enable the auditory system to adjust phoneme boundaries are often debated. Numerous theories of speech perception have been invoked in explanations of recalibration and perceptual retuning as well. Cutler, McQueen, Norris and colleagues (Norris et al. 2000) originally proposed a feed-forward model of speech perception called Merge, and argued that listeners can retune phoneme categories through a bottom-up abstraction process, which does not rely upon online feedback from the lexicon, not unlike the COHORT model which also states that word recognition primarily relies on bottom-up processes (Gaskell and Marslen-Wilson 1997). COHORT presents a modular, unidirectional explanation, where word recognition is initiated first by acoustic information, triggering a possible “cohort” of matches, and later, other features such as context

and semantics allow the listener to narrow down the possibilities. Similarly, according to the Merge model, top-down feedback during speech recognition and phoneme categorization is not essential, and these processes operate at a pre-lexical level. Feedback during categorization could be time-consuming, so interactions between lexical and pre-lexical processing would not be beneficial. Phonemic decisions can be made based on both lexical and pre-lexical information, but does not necessitate interactions between the processes. Cutler et al. (2010) also emphasized that perceptual retuning cannot be explained purely by episodic information, and that abstraction from such events must be involved as well. A more recent model by Norris et al. (2016) has been updated to include predictions of perception based on Bayesian inference, but still does not rely upon online feedback during phoneme processing. Acoustic information and lexical knowledge are combined to calculate probable phonemes, but again, the two processes are not proposed to interact.

Others have described top-down (Davis et al. 2005; Davis and Johnsrude 2007) and bidirectional influences on speech perception (McClelland and Elman 1986; McClelland et al. 2006). A classical, interactive model of speech perception, TRACE (McClelland and Elman 1986), derives its name from a structure called "The Trace", a perceptual processing tool. McClelland and Elman proposed that top-down feedback modulates connections between three layers; from words, to phonemes, down to features. Phoneme identification can be influenced by lexical and speech-reading contexts, and can also be improved through experience. According to TRACE, this influence is due to feedback from higher levels of processing. Similarly, McClelland et al. (2006) contend that both top-down and bottom-up information streams are essential for speech perception. Phoneme representations can be influenced by both lexical and acoustic features, and vice versa.

Some have argued that phonemes cannot be represented abstractly, as retuning can be dependent on episodic features from the exposure phase. As discussed earlier, retuning does not always generalize to new speakers, even those with the same accent (Reinisch and Holt 2014; Xie and Myers 2017). Thus, phoneme representations may not be completely abstracted from the input received, and may

retain token- and context-specific details. Studies of audiovisual recalibration have also raised similar opinions, that phoneme representations cannot be fully abstracted during recalibration. As mentioned in sect. 3.3, the ear in which stimuli were heard or the spatial location can determine the extent of recalibration (Keetels et al. 2015; Keetels et al. 2016). Keetels et al. (2015) argue that this could be due to the perceptual system striking a balance between generalizing too often and too rarely. If recalibration is employed when speech is unclear, then it is may be only necessary to apply the newly learned boundary position to other instances that are similar both in acoustic and contextual features, so as to not unnecessarily over-generalize.

Likewise, Kleinschmidt and Jaeger (2015) have put forth a belief-updating model based on Bayesian inference, of both audiovisual recalibration and selective speech adaptation, called the Ideal Adaptor Framework. As described in sect. 3.2, audiovisual recalibration and selective speech adaptation are two forms of perceptual learning, but their response profiles are in direct contrast to each other. In the Ideal Adaptor Framework, both recalibration and selective speech adaptation are described as forms of statistical learning, as a result of exposure to various distributions of phonemes. Listeners can create speaker-specific models of phoneme categories which allow for initial speaker-level adaptation, but can eventually generalize to more speakers with additional experience and if they are also acoustically close. The authors also posit recalibration and selective speech adaptation as two response patterns along a continuum ranging from ambiguous to prototypical sounds. As mentioned earlier in sect. 2.2, recalibration effects tend to peak after approximately eight exposure tokens and slowly diminish with additional exposures, while selective speech adaptation tends to continuously build in a linear manner with increasing exposure. According to the model, recalibration reflects a response to ambiguous sounds, but with increasing amounts of exposure tokens and as speech sounds become more prototypical, selective adaptation effects can be observed.

5.2 Neural Correlates of Recalibration and Perceptual Retuning

While theoretical frameworks and models have been useful in understanding recalibration and retuning, neuroimaging studies have shed additional light on areas of the brain where these changes occur and how they might explain the levels of processing involved. More general models of speech perception drawn from neuroimaging data and primate studies (Scott and Johnsrude 2003; Rauschecker and Scott 2009) have described the hierarchical and topographic nature of processing in the auditory cortex and surrounding areas.

Hickok and Poeppel (2007) proposed the dual-stream processing model of speech, with certain features equivalent to those found in visual-processing models. According to the model, areas of the brain along a ventral pathway, including medial temporal gyrus (MTG) and inferior temporal sulcus (ITS), are geared towards connecting phonological and lexical representations, while regions along a dorsal pathway, including parietal-temporal, (pre)motor, and inferior frontal regions are geared towards connecting phonological with sensorimotor and articulatory representations. Jäncke et al. (2002) also identified structures of the brain specific to phoneme perception, in the planum temporale (PT) and middle superior temporal gyrus (STG). STG and the primary auditory cortex can also encode fine-tuned phonetic information (Mesgarani et al. 2008; 2014), with evidence for speaker-invariant phoneme representations distributed across both of these regions (Formisano et al. 2008; Bonte et al. 2014). Other regions implicated in categorical perception of speech sounds include the parietal-temporal and inferior parietal cortex (Davis and Johnsrude 2007; Raizada and Poldrack 2007).

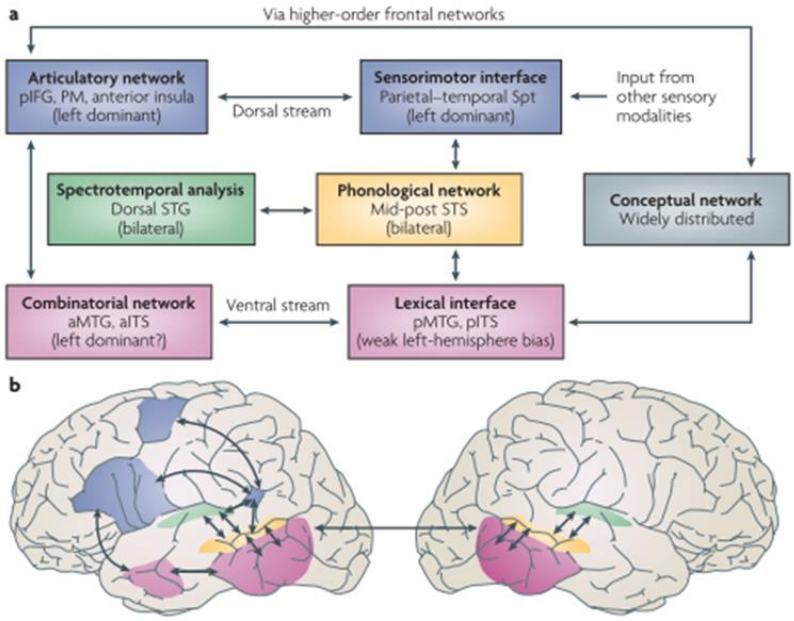


Figure 4. The dual-stream processing model of language, by Hickok and Poeppel (2007). A number of regions primarily within and around the temporal lobe are proposed to be responsible for the levels of linguistic processing, and splitting between a ventral and a dorsal pathway.

While these studies paved the way towards delineating a network of regions possibly implicated in recalibration, they may still be insufficient, as this process relies on the integration of both acoustic and contextual information, which are often lexical or visual. In light of this, Obleser and Eisner (2009) proposed a model of pre-lexical abstraction, reminiscent of the Merge model, based on prior neuroimaging studies of speech perception. Pre-lexical abstraction may appear to resemble recalibration, but it also implies that the phoneme representation can be fully disentangled from the acoustic input and thereby abstracted. Pre-lexical abstraction could be implemented probabilistically, primarily along the STG, resulting in phoneme likelihoods rather than definitive phoneme identification. Likelihoods could be calculated by weighing various acoustic features, first processed by primary auditory cortex, and could be updated with talker and context-specific information.

A recent study (Holdgraf et al. 2016) has also found evidence for a mechanism of perceptual enhancement, through spectro-temporal receptive field (STRF)

mapping on eletrocorticography recordings (ECoG) of the auditory cortex. Responses of cortical populations had increased sensitivity to speech-like, spectro-temporal features of degraded speech, after exposure to intact speech. This sensitivity could reflect how listeners encode rudimentary acoustic features that also allow the listener to interpret less intelligible speech, or how listeners "fill in the gaps".

The merits of these models of speech perception can be reexamined in light of functional-MRI (fMRI) studies of recalibration itself. Kilian-Hütten et al. (2011b) had participants undergo audiovisual recalibration using the classic /aba/-/ada/ stimuli while fMRI data was collected. It was discovered that a higher-order network of areas in and around the auditory cortex, including bilateral inferior parietal lobe (IPL), inferior frontal sulcus (IFS), STS/STG, and posterior MTG were all active in recalibration. These areas showed overlapping activation during both the exposure phase and the subsequent test phase. These regions are also known to be involved in audiovisual integration and constructive processes, which would account for their increased activation during recalibration. Kilian-Hütten et al. (2011a) were also able to investigate audiovisual recalibration using MVPA, or multivariate pattern analysis, a technique using fMRI data to train an algorithm to recognize differences in patterns of brain activity. They were successfully able to decode whether a participant perceived /aba/ or /ada/ while presented with the ambiguous sounds during the test phase of the same audiovisual recalibration experiment, solely using the activation patterns. Some of the areas that most effectively predicted the percepts, typically viewed as low level auditory areas, included clusters in and around left planum temporale (PT) and left Heschl's gyrus and sulcus, but evidently, they were influenced by information other than elementary acoustics features.

More recently, Lüttke et al. (2016) investigated a form of adaptation induced by McGurk-style adaptors with fMRI. Exposure to McGurk adaptors, or clear auditory /aba/ paired with video of /aga/, resulted in the percept of /ada/. These stimuli led to an effect much like selective speech adaptation, where follow-up presentations of clear auditory /aba/ were incorrectly perceived as /ada/ as a result. This mistaken /ada/ percept showed closely related neural patterns to those elicited

by correctly perceived auditory /ada/, and more so than to patterns associated with correct perception of clear /aba/ tokens. Again, neural activations echoed a shift in auditory perception due to adaptation through contextual cues.

fMRI has also been used to explore lexically-driven perceptual learning and other related phenomena. Activation in posterior left STG and STS has been recorded in listeners receiving instructions to switch from an acoustic mode to speech mode while listening to sine-wave speech stimuli (Dehaene-Lambertz et al. 2005). While stimuli remained the same, instructions alone could induce a shift in both perception and the resulting activation patterns. Similarly, activity in left pSTS has also been associated with identification of non-phonemic, short-term sound categories, while left mSTS may store long-term representation of phoneme patterns already known to the listener (Liebenthal et al. 2010). Myers and Blumstein (2008) investigated the Ganong effect (described in sect. 1.1), or the impact of lexical knowledge on perception of ambiguous speech tokens. Participants heard auditory items with ranging voice onset time (VOT) from *gift* to *kift* (i.e. word to nonword) and another continuum ranging from *giss* to *kiss* (from nonword to word). Activity in STG was modulated by the lexical effect, such that boundary tokens that were perceived as words showed higher activations compared to acoustically similar tokens from the other continuum that were not perceived as words. As STG was engaged in both phonological and lexical processing, the authors suggested that this was evidence in support of top-down models similar to TRACE that accommodate higher-level information during processing. (Myers and Blumstein 2008)

Similarly, Myers and Mesite (2014) tested participants in a classic lexically-guided perceptual retuning experiment with the addition of fMRI, alternating between exposure phases containing edited words ending in an ambiguous phoneme, followed by a forced-choice test phase on a continuum of the same ambiguous sounds. Participants were separated into two groups with the stimuli biased towards /s/ for one group, and towards /ʃ/ (the "sh" in shop) for the other. Behavioral results indicated a boundary shift, so over the course of the successive test phases, participants' perception of the ambiguous /s/-/ʃ/ phoneme had changed. Increased activity in left IFG and STG was measured with boundary

shifted items. These items reflected the perceptual shift, and were categorized as the biasing phoneme in test blocks following the exposure, but not during the earlier blocks at the start of the experiment. Activity both within the auditory cortex and in higher-level cognitive areas suggests that top-down information may have influenced the learning process, and may also have been responsible for creating connections between phonetic information and the speaker. Together, these results imply that perceptual learning may not be accomplished in a unidirectional manner, due to the involvement of areas encompassing both lower and higher levels of processing in the perception of these sounds.

Combined magneto-encephalogram (MEG) and electro-encephalogram (EEG) data have also confirmed that activity in STG can reduce over time, as participants learn to improve in identification of degraded speech sounds combined with matching text (Sohoglu and Davis 2015). Furthermore, the results were framed within a model of predictive coding, not unlike Bayesian inference, such that the listener can learn to reduce prediction errors as a consequence of learning. STG is proposed to encode acoustic features and receive predictions of phonological categories from higher-level frontal areas, and predictions are continuously updated with experience.

While many of the studies discussed thus far have identified STG to be involved in perceptual learning or recalibration, a recent study has also found evidence from the cerebellum, a sub-cortical area most well-known for sensorimotor functioning (Guediche et al. 2015). Listeners learned to identify words distorted by noise vocoding, and consequently, cerebellar regions showed changes in activity, as well as functional connections to cortical language and auditory regions. Stemming in part from this finding, another model of adaptation to speech has been proposed, also relying on a predictive coding mechanism, but supervised by the cerebellum (see Guediche et al. 2014 for a complete review).

Section 5 detailed various theories of speech perception as well as supporting neuroimaging data that propose the channels through which recalibration and perceptual retuning may operate. Proponents of these speech perception theories have debated the nature of how phoneme categories can be reshaped, as some argue that this is a unidirectional, bottom-up abstraction process (Merge, COHORT),

while others postulate that both top-down and bottom-up processes contribute (TRACE). Theories incorporating distributional and statistical learning, such as the Ideal Adaptor Framework (Kleinschmidt and Jaeger 2015) have also been useful for understanding how listeners adapt to variability. Neuroimaging data suggest that both top-down and bottom-up influences are involved, based on the areas of the brain that tend to be active during perception of ambiguous tokens, such as STS/STG and IFS/IFG. Sophisticated analysis techniques such as MVPA have also been useful for pinpointing specific patterns of neural activity associated with the shifts in perception, but the directionality of influences upon these percepts remain unclear and may require more advanced neuroscientific methods.

6 Conclusion and Future Directions

The literature described throughout this chapter has focused on lexical and audiovisual information as contextual influences on speech perception, as well as their dimensions and limitations. Section 2 highlighted the seminal findings regarding lexical retuning, starting from Norris et al. (2003) and the studies since then that have illuminated the strengths and drawbacks. Section 3 discussed audiovisual recalibration, first described by Bertelson et al. (2003) and expanded upon by others, most often Vroomen and colleagues.

These two contextual sources can differ in terms of their impact on perception, as lexical information can potentially lead to more stable and longer lasting shifts in perception, while audiovisual information results in adjustments in shorter durations that are not easily generalizable and are often either (or both) context and token-dependent. The phoneme categories themselves can also impose restrictions, as plosives (also known as stop consonants) may allow for generalization to other speakers more so than other types of phonemes, such as fricatives or liquids. Evidently, contextual cues alone do not drive these phoneme boundary shifts, and acoustic information still modulates learning effects to a great extent. Theories of speech perception have also been helpful for understanding the basis of phoneme boundary adjustments, but disagreements exist with regard to the stages of processing that are thought to be involved.

Although questions remain in the field as to the precise details of retuning, researchers continue to pursue the answers with behavioral and neuroimaging studies. Related works may also shed light upon how exactly these perceptual shifts may occur. Recent studies have investigated another related form of text-based recalibration. Reading text of syllables while listening to ambiguous phonemes can also contribute to changes in phoneme categorization (Keetels et al. 2016), and this has also been tested using fMRI (Bonte et al. 2017). Just as in audiovisual and lexical experiments, participants viewed either /aba/ or /ada/ written in text, while hearing an ambiguous blend of the two, and participants were able to effectively recalibrate depending on the text they viewed (Keetels et al. 2016). In addition, fMRI results showed that text-based recalibration was linked to activity in posterior superior temporal cortex, and percepts of /aba/ and /ada/ during test could also be decoded with MVPA, primarily based on patterns of activity in left posterior STG and planum temporale and right STS (Bonte et al. 2017). Functional connectivity was observed between IPL and left STG during exposure, and may be indicative of higher-order influences leading to eventual retuning. While lexical and audiovisual recalibration studies have been useful for understanding how listeners adapt to ambiguity in speech, this new paradigm illuminates how mappings are acquired between auditory and written representations, and may also have the potential to detect disruptions of reading networks during development, particularly in individuals with dyslexia.

Together, these approaches using lexical and audiovisual information, and more recently with text, have proven useful in understanding the plasticity of speech sounds. These non-acoustic sources of information can not only sway how speech tokens are perceived, but moreover, can restructure the units of speech. Evidently, these units are malleable and are continuously updated with experience; they are susceptible to change even within short windows of time and with relatively little input required to do so. This adaptive tool is beneficial for adjusting to speakers, noise, or other obstacles that could impede successful speech comprehension, although the acoustic features of the input may restrict the extent to which recalibration can be generalized. Still, stimulus-specificity may be advantageous, as a complete overhaul of speech sounds in response to deviations

from the norm would be impractical. Speech perception theories and neuroimaging studies have highlighted the possible processing streams involved, and both lexical and speech-reading influences appear to share significant similarities in terms of the brain areas being recruited. The relative contributions of top-down and bottom-up information in processing the acoustic input is still hotly debated, but the continued application of advanced neuroimaging techniques, as well as statistical modeling may aid in building a more cohesive picture of perceptual retuning.

7 Outline of the Dissertation

Questions remain unanswered regarding the ways in which listeners can exploit contextual information in order to guide category adjustments. As such, the goal of this dissertation was to reconcile the gaps between audiovisual recalibration and lexical retuning, to search for a coherent understanding of the two processes. To do so, the studies sought to measure the two processes under similar testing conditions, with paradigms that were suitable for either audiovisual or lexical perceptual learning and could reveal similarities and differences between them. With an appropriate design, the application of fMRI was incorporated to explicate the brain regions which modulate these perceptual adjustments and the resulting implications concerning the functional organization of speech perception. In **Chapter 2**, the two forms of perceptual learning were compared in a novel design that had listeners switch between the two cue types in order to measure the subsequent retuning and recalibration effects under short time constraints. Listeners who switched between audiovisual and lexical cues showed recalibration and retuning effects comparable to listeners who only received one cue type, and audiovisual cues overall resulted in larger aftereffects. **In Chapter 3,** retuning and recalibration were compared by compounding the audiovisual and lexical cues together, to see whether additive learning effects were possible when listeners had both audiovisual and lexical cues available. Listeners did not show additive effects, but rather showed effects similar to audiovisual cues alone. Once an appropriate design was established, that could be applied to measure either recalibration or retuning, **in Chapter 4,** audiovisual and lexical perceptual learning were then

compared in an fMRI study, to determine the underlying neural processes and whether the two showed overlap or differed in the brain areas recruited. Retuning and recalibration showed similar patterns of neural activity, particularly in the temporal cortex, but audiovisual recalibration showed strong activation in the visual cortex, despite the absence of any visual stimuli. In the **Chapter 5,** a summary of the empirical chapters and their findings will be discussed, as well as an outlook on the research field and the implications for later studies.

References

Baart M, de Boer-Schellekens L, Vroomen J (2012) Lipread-induced phonetic recalibration in dyslexia. Acta Psychol 140(1):91-95. doi: 10.1016/j.actpsy.2012.03.003

Baart M, Samuel AG (2015) Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing. J Mem Lang 85:42-59. doi: 10.1016/j.jml.2015.06.008

Baart M, Vroomen J (2010b) Phonetic recalibration does not depend on working memory. Exp Brain Res 203:575-582. doi: 10.1007/s00221-010-2264-9

Bertelson P, Vroomen J, De Gelder B (2003) Visual recalibration of auditory speech identification: a McGurk aftereffect. Psychol Sci 14(6):592-597. doi: 10.1046/j.0956-7976.2003.psci_1470.x

Bonte M, Correia JM, Keetels M, Vroomen J, Formisano E (2017) Reading-induced shifts of perceptual speech representations in auditory cortex. Sci Rep 7:1-11. doi: 10.1038/s41598-017-05356-3

Bonte M, Hausfeld L, Scharke W, Valente G, Formisano E (2014) Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. J Neurosci 34(13):4548-4557. doi: 10.1523/JNEUROSCI.4339-13.2014

Brancazio L (2004) Lexical influences in audiovisual speech perception. J Exp Psychol Human 30(3):445-463. doi: 10.1037/0096-1523.30.3.445

Bradlow AR, Bent T (2008) Perceptual adaptation to non-native speech. Cognition 106(2):707-729. doi: 10.1016/j.cognition.2007.04.005

Clarke CM, Garrett MF (2004) Rapid adaptation to foreign-accented English. J Acoust Soc Am 116(6):3647-3658. doi: 10.1121/1.1815131

Cutler A, Eisner F, McQueen JM, Norris D (2010) How abstract phonemic categories are necessary for coping with speaker-related variation. In: Fougeron C, Kühnert B, D'Imperio M, Vallée N (eds), Laboratory phonology, vol. 10. de Gruyter, Berlin, pp 91-111.

Cutler A, McQueen JM, Butterfield S, Norris D (2008) Prelexically-driven perceptual retuning of phoneme boundaries. In Fletcher J, Loakes D, Goecke R, Burnham D, Wagner M (eds), Proceedings of Interspeech, Brisbane, 2008.

Davis MH, Johnsrude IS (2007) Hearing speech sounds: top-down influences on the interface between audition and speech perception. Hearing Res 229(1-2):132-147. doi: 10.1016/j.heares.2007.01.014

Davis MH, Johnsrude IS, Hervais-Adelman AG, Taylor K, McGettigan C (2005) Lexical information drives perceptual learning of distorted speech: Evidence from

the comprehension of noise-vocoded sentences. J Exp Psychol Gen 134(2):222-241. doi: 10.1037/0096-3445.134.2.222

Dehaene-Lambertz G, Pallier C, Serniclaes W, Sprenger-Charolles L, Jobert A, Dehaene S (2005) Neural correlates of switching from auditory to speech perception. Neuroimage 24(1):21-33. doi: 10.1016/j.neuroimage.2004.09.039

Drozdova P, van Hout R, Scharenborg O (2015) Lexically-guided perceptual learning in non-native listening. Biling-Lang Cogn 19(5):914-920. doi: 10.1017/S136672891600002X

Eimas PD, Corbit JD (1973) Selective adaptation of linguistic feature detectors. Cognitive Psychol 4:99–109. doi: 10.1016/0010-0285(73)90006-6

Eisner F, McQueen JM (2005) The specificity of perceptual learning in speech processing. Atten Percept Psycho 67:224–238. doi: 10.3758/BF03206487

Eisner F, McQueen JM (2006) Perceptual learning in speech: Stability over time. J Acoust Soc Am 119:1950–1953. doi: 10.1121/1.2178721

Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what"? Brain based decoding of human voice and speech. Science 322(5903):970-973. doi: 10.1126/science.1164318

Franken MK, Eisner F, Schoffelen JM, Acheson DJ, Hagoort P, McQueen JM (2017) Audiovisual recalibration of vowel categories. In: Proceedings of Interspeech, Stockholm, p 655-658. doi: 10.21437/Interspeech.2017-122

Gaskell MG, Marslen-Wilson WD (1997) Integrating form and meaning: a distributed model of speech perception. Lang Cognitive Proc 12(5-6):613-656. doi: 10.1080/016909697386646

Ganong WF (1980) Phonetic categorization in auditory word perception. J Exp Psychol Human 6(1): 110-125. doi: 10.1037/0096-1523.6.1.110

Guediche S, Blumstein SE, Fiez JA, Holt LL (2014) Speech perception under adverse conditions: insights from behavioral, computational, and neuroscience research. Front Syst Neurosci 7:1-16. doi: 10.3389/fnsys.2013.00126

Guediche S, Holt LL, Laurent P, Lim S, Fiez JA (2015) Evidence for cerebellar contributions to adaptive plasticity in speech perception. Cereb Cortex 25:1867-1877. doi: 10.1093/cercor/bht428

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393-402. doi: 10.1038/nrn2113

Holdgraf CR, de Heer W, Pasley B, Rieger J, Crone N, Lin JJ, Knight RT, Theunissen FE (2016) Rapid tuning shifts in human auditory cortex enhance speech intelligibility. Nat Commun 7:13654. doi: 10.1038/ncomms13654

Holt LL, Lotto AJ (2008) Speech perception within an auditory cognitive science framework. Curr Dir Psychol Sci 17(1):42-46. doi: 10.1111/j.1467-8721.2008.00545.x

Jäncke L, Wüstenberg T, Scheich H, Heinze HJ (2002) Phonetic perception and the auditory cortex. Neuroimage 15(4):733-746. doi: 10.1006/nimg.2001.1027

Keetels MN, Pecoraro M, Vroomen J (2015) Recalibration of auditory phonemes by lipread speech is ear-specific. Cognition 141:121-126. doi: 10.1016/j.cognition.2015.04.019

Keetels MN, Schakel L, Bonte M, Vroomen J (2016a) Phonetic recalibration of speech by text. Atten Percept Psycho 78:938–945. doi: 10.3758/s13414-015-1034-y

Keetels MN, Stekelenburg JJ, Vroomen J (2016b) A spatial gradient in phonetic recalibration by lipread speech. J Phonetics 56:124–130. doi: 10.1016/j.wocn.2016.02.005

Kilian-Hütten N, Valente G, Vroomen J, Formisano E (2011a) Auditory cortex encodes the perceptual interpretation of ambiguous sound. J Neurosci 31(5):1715-1720. doi: 10.1523/JNEUROSCI.4572-10.2011

Kilian-Hütten N, Vroomen J, Formisano E (2011b) Brain activation during audiovisual exposure anticipates future perception of ambiguous speech. Neuroimage 57(4):1601-1607. doi: 10.1016/j.neuroimage.2011.05.043

Kleinschmidt DF, Jaeger TF (2015) Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. Psychol Rev 122(2):148-203. doi: 10.1037/a0038695

Kraljic T, Brennan SE, Samuel AG (2008a) Accommodating variation: Dialects, idiolects, and speech processing. Cognition 107:51–81. doi: 10.1016/j.cognition.2007.07.013

Kraljic T, Samuel AG (2005) Perceptual learning for speech: Is there a return to normal? Cognitive Psychol 51:141–178. doi: 10.1016/j.cogpsych.2005.05.001

Kraljic T, Samuel AG (2006) Generalization in perceptual learning for speech. Psychon B Rev 13:262–268. doi: 10.3758/BF03193841.

Kraljic T, Samuel AG (2007) Perceptual adjustments to multiple speakers. J Mem Lang 56:1-15. doi: 10.1016/j.jml.2006.07.010.

Kraljic T, Samuel AG (2009) Perceptual learning for speech. Atten Percept Psycho 71(3):1207-1218. doi: 10.3758/APP.71.6.1207.

Kraljic T, Samuel AG, Brennan SE (2008b) First impressions and last resorts: How listeners adjust to speaker variability. Psychol Sci 19:332–338. doi: 10.1111/j.1467-9280.2008.02090.x.

Lecumberri MLG, Cooke M, Cutler A (2010) Non-native speech perception in adverse conditions: a review. Speech Commun 52(11-12):864-886. doi: 10.1016/j.specom.2010.08.014.

Liebenthal E, Desai R, Ellingson MM, Ramachandran B, Desai A, Binder JR (2010) Specialization along the left superior temporal sulcus for auditory categorization. Cereb Cortex 20(12):2958-2970. doi: 10.1093/cercor/bhq045.

Maye J, Aslin RN, Tanenhaus MK (2008) The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. Cognitive Sci 32(3):543-562. doi: 10.1080/03640210802035357.

Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. Science 343(6174):1006-1010. doi: 10.1126/science.1245994

Mesgarani N, David SV, Fritz JB, Shamma SA (2008) Phoneme representation and classification in primary auditory cortex. J Acoust Soc Am 123(2):899-909. doi: 10.1121/1.2816572

McClelland JL, Elman JL (1986) The TRACE model of speech perception. Cognitive Psychol 18:1-86. doi: 10.1016/0010-0285(86)90015-0

McClelland JL, Mirman D, Holt LL (2006) Are there interactive processes in speech perception? Trends Cogn Sci 10(8):363-369. doi: 10.1016/j.tics.2006.06.007

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

McQueen JM, Cutler A, Norris D (2006a) Phonological abstraction in the mental lexicon. Cognitive Sci 30:1113–1126. doi: 10.1207/s15516709c0g0000_79

McQueen JM, Norris D, Cutler A (2006b) The dynamic nature of speech perception. Lang Speech 49(1):101-112. doi: 10.1177/00238309060490010601

Mitterer H, Scharenborg O, McQueen JM (2013) Phonological abstraction without phonemes in speech perception. Cognition 129:356-261. doi: 10.1016/j.cognition.2013.07.011

Myers EB, Blumstein SE (2008) The neural basis of the lexical effect: an fMRI investigation. Cereb Cortex 18:278-288. doi: 10.1093/cercor/bhm053

Myers EB, Mesite LM (2014) Neural systems underlying perceptual adjustment to non-standard speech tokens. J Mem Lang 76:80-93. doi: 10.1093/cercor/bhm053

Norris D, Cutler A, McQueen JM, Butterfield S (2006) Phonological and conceptual activation in speech comprehension. Cognitive Psychol 53(2):146-193. doi: 10.1016/j.cogpsych.2006.03.001

Norris D, McQueen JM, Cutler A (2000) Merging information in speech recognition: feedback is never necessary. Behav Brain Sci 23:299–325. doi: 10.1017/S0140525X00003241

Norris D, McQueen JM, Cutler A (2003) Perceptual learning in speech. Cognitive Psychol 47:204–238. doi: 10.1016/S0010-0285(03)00006-9

Norris D, McQueen JM, Cutler A (2016) Prediction, Bayesian inference and feedback in speech recognition. Lang Cogn Neurosci 31(1):4-18. doi: 10.1080/23273798.2015.1081703

Obleser J, Eisner F (2009) Pre-lexical abstraction of speech in the auditory cortex. Trends Cogn Sci 13(1):14-19. doi: 10.1016/j.tics.2008.09.005

Raizada RD, Poldrack RA (2007) Selective amplification of stimulus differences during categorical processing of speech. Neuron 56(4):726-740. doi: 10.1016/j.neuron.2007.11.001

Reinisch E, Holt LL (2014) Lexically-guided phonetic retuning of foreign-accented speech and its generalization. J Exp Psychol Human 40(2):539-555. doi: 10.1037/a0034409.

Reinisch E, Weber A, Mitterer H (2013) Listeners retune phoneme categories across languages. J Exp Psychol Human 39:75-86. doi: 10.1037/a0027979

Reinisch E, Wozny D, Mitterer H, Holt LL (2014) Phonetic category recalibration: What are the categories? J Phonetics 45:91-105. doi: 10.1016/j.wocn.2014.04.002

Roberts M, Summerfield Q (1981) Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. Atten Percept Psycho 30(4):309-314. doi: 10.3758/BF03206144

Samuel AG, Frost R (2015) Lexical support for phonetic perception during non-native spoken word recognition. Psychon B Rev 22(6):1746-1752. doi: 10.3758/s13423-015-0847-y

Sjerps MJ, McQueen JM (2010) The bounds on flexibility in speech perception. J Exp Psychol Human 36:195–211. doi: 10.1037/a0016803

Sohoglu E, Davis MH (2016) Perceptual learning of degraded speech by minimizing prediction error. P Natl Acad Sci USA 113(12):1747-1756. doi: 10.1073/pnas.1523266113.

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212-215. doi: 10.1121/1.1907309

Van der Zande P, Jesse A, Cutler A (2014) Hearing words helps seeing words: A cross-modal word repetition effect. Speech Commun 59:31-43. doi: 10.1016/j.specom.2014.01.001

Van Laarhoven T, Keetels M, Schakel L, Vroomen J (2018) Audio-visual speech in noise perception in dyslexia. Developmental Sci 21(1):e12504. doi: 10.1111/desc.12504.

Van Linden S, Vroomen J (2007) Recalibration of phonetic categories by lipread speech versus lexical information. J Exp Psychol Human 33(6):1483-1494. doi: 10.1037/0096-1523.33.6.1483

Van Linden S, Vroomen J (2008) Audiovisual speech recalibration in children. J Child Lang 35(4):809-822. doi: 10.1017/S0305000908008817

Vroomen J, Baart M (2009a) Phonetic recalibration only occurs in speech mode. Cognition 110(2):254-259. doi: 10.1016/j.cognition.2008.10.015

Vroomen J, Baart M (2009b) Recalibration of phonetic categories by lipread speech: Measuring aftereffects after a twenty-four hours delay. Lang Speech 52:341-350. doi: 10.1177/0023830909103178

Vroomen J, van Linden S, Baart M (2007a) Lipread aftereffects in auditory speech perception: measuring aftereffects after a twenty-four hours delay. In: Vroomen J, Swerts M, Krahmer E (eds), Auditory-Visual Speech Processing, Hilvarenbeek, p P05.

Vroomen J, van Linden S, de Gelder B, Bertelson P (2007b) Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. Neuropsychologia 45(3):572-577. doi: 10.1016/j.neuropsychologia.2006.01.031

Vroomen J, van Linden S, Keetels M, de Gelder B, Bertelson P (2004) Selective adaptation and recalibration of auditory speech by lipread information: dissipation. Speech Commun 44:55-61. doi: 10.1016/j.specom.2004.03.009

Winn M (2018) Speech: It's not as acoustic as you think. Acoustics Today 14(2):43-49.

Xie X, Myers EB (2017) Learning a talker or learning an accent: acoustic similarity constrains generalization of foreign accent adaptation to new talkers. J Mem Lang 97:30-46. doi: 10.1016/j.jml.2017.07.005

Zhang X, Samuel AG (2015) Perceptual learning of speech under optimal and adverse condition. J Exp Psychol Human 40(1), 200-217. doi: 10.1037/a0033

2

Interleaved lexical and audiovisual information can retune phoneme boundaries

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

Abstract

To adapt to situations in which speech perception is difficult, listeners can adjust boundaries between phoneme categories using perceptual learning. Such adjustments can draw on lexical information in surrounding speech, or on visual cues via speech-reading. In the present study, listeners proved able to flexibly adjust the boundary between two plosive/stop consonants, /p/-/t/, using both lexical and speech-reading information and given the same experimental design for both cue types. Videos of a speaker pronouncing pseudo-words and audio recordings of Dutch words were presented in alternating blocks of either stimulus type. Listeners were able to switch between cues to adjust phoneme boundaries, and resulting effects were comparable to results from listeners receiving only a single source of information. Overall, audiovisual cues (i.e., the videos) produced the stronger effects, commensurate with their applicability for adapting to noisy environments. Lexical cues were able to induce effects with fewer exposure stimuli and a changing phoneme bias, in a design unlike most prior studies of lexical retuning. While lexical retuning effects were relatively weaker compared to audiovisual recalibration, this discrepancy could reflect how lexical retuning may be more suitable for adapting to speakers than to environments. Nonetheless, the presence of the lexical retuning effects nonetheless suggests that it may be invoked at a faster rate than previously seen. In general, this technique has further illuminated the robustness of adaptability in speech perception, and offers the potential to enable further comparisons across differing forms of perceptual learning.

Key words: phoneme boundary, recalibration, perceptual retuning, lexical, audiovisual

Introduction

Listeners often encounter situations where they must understand a speaker they have never heard before, and must rapidly adapt to the unique acoustic characteristics of the individual's speech. In such scenarios, information other than the auditory signal itself can be utilized to assist the listener and can influence listeners' interpretation of what they are hearing. Early studies demonstrated that knowledge of the lexicon and speech-reading can create an immediate bias in what listeners perceive (Ganong, 1980; McGurk & MacDonald, 1976). More recent studies of perceptual retuning have shown that listeners can learn to disambiguate speech or speech-like sounds, by adjusting the boundary of a phoneme category and expanding the criteria used to identify a phoneme. Both lexical and speech-reading information have been established as sources that can facilitate this process, and thus enable the famously robust adaptability of human speech perception (Cutler, 2012; Vroomen & Baart, 2012).

In the initial experiments on perceptual retuning, listeners heard and viewed speech or speech-like stimuli edited to remove clear instances of a critical phoneme which were then replaced by an ambiguous phoneme blend nearly indistinguishable from a natural version (Bertelson et al., 2003; Norris et al., 2003). In lexically-guided perceptual learning, recordings of words ending in a particular phoneme (e.g., /s/, as in *carcass*), are edited to end in an ambiguous phoneme instead, such as an /s/-/f/ blend (Norris et al., 2003; Samuel & Kraljic, 2009). Following exposure to such stimuli, listeners perform a categorization task on the ambiguous token and other neighboring sounds along an /s/-/f/ continuum and are likely to report hearing more sounds in accordance with the preceding exposure stimuli (i.e., as an /s/). Listeners are also likely to perceive an /s/-/f/ blend as /f/, if they hear recordings of /f/-final words (e.g., *paragraph*) with the ambiguous token replacing the /f/. Likewise, in visually-guided recalibration, participants are presented with video recordings of a speaker pronouncing a syllable (/aba/ or /ada/) paired with an audio recording of an ambiguous token (/aba/-/ada/ blend) (Bertelson et al., 2003). After sufficient exposure to these videos, participants perform a categorization task on the ambiguous token, and are also likely to report

perceiving it as the phoneme it was replacing (as /aba/ if coupled with videos of /aba/, or as /ada/ with videos of /ada/). Note that we have used recalibration here to refer to the audiovisual form, retuning to refer to the lexical version, and perceptual learning when referring to both. This is in correspondence with the terminology used by the researchers who have developed and deployed the two approaches, and we will maintain the distinction throughout our report for the convenience of the reader.

The lexical and visual approaches are certainly similar in that they both reveal how internal representations of speech sounds can be reshaped during perceptual experience by reference to existing knowledge. However, despite this similarity in the resulting effects, the course of the learning can vary across other dimensions, such as build-up and dissipation, or the extent to which the effects are still measureable. Lexical retuning studies typically use longer exposure phases with critical items embedded into a lexical decision task or other listening material containing filler words as well, while audiovisual recalibration studies often repeat videos of a single syllable and eight exposure tokens can be enough to induce after-effects (see Samuel & Kraljic, 2009, for an overview). Eisner and McQueen (2006) have shown that the retuning effects from lexical information can be present up to 12 hours after exposure, both during the daytime or after a night of sleep, while Baart and Vroomen (2009) noted that audiovisual recalibration effects can quickly diminish with increasing numbers of items during the follow-up categorization task, and are not observable after 24 hours.

Van Linden and Vroomen (2007) sought to quantify these differences between lexical and audiovisual perceptual learning by exposing participants to both forms in two separate sessions, with the categorization task immediately following each such exposure phase. Retuning effects were larger after audiovisual exposure than after lexical, but could build up and dissipate in a similar fashion when the exposure and test phases were structured consistently.

What is as yet unknown is whether both forms of perceptual learning can also be called upon within the same circumstances and under the same experimental constraints. Perceptual systems must be flexible so as to accommodate possible variability in speech, so listeners should be capable of

switching between available contextual cues depending on the needs of the situation, but conversely, may also find that switching between two cue types does not allow perceptual learning effects to build up sufficiently. The present study addresses this question by comparing perceptual learning effects following lexical and visual/speech-reading exposure, both within participants and within a single session. In order to compare them within a single session, the study also explored whether lexical retuning can take place under more restricted conditions, with short exposure blocks in two possible biasing directions, rather than a long exposure pointing towards only one phoneme. Following brief exposures to stimuli ending in an ambiguous phoneme (a /p/-/t/ blend), wherein the direction of the bias was changing throughout the session, participants were expected to continuously adjust the phoneme boundary between two clear phonemes, based on their responses during categorization tasks on ambiguous phoneme blends. The same procedure for both audiovisual recalibration and lexical retuning was maintained in order to compare them directly. It further allowed us to determine whether lexical retuning was possible under more restricted conditions more typical of audiovisual recalibration, by presenting only 8 items per exposure block. The design, adapted from van Linden and Vroomen (2007), incorporated pseudo-words and words for audiovisual and lexical recalibration, respectively, by presenting interleaved exposure blocks of the two types of stimuli, each followed by test blocks containing ambiguous phonemes without context.

Methods

Participants

Sixty healthy native Dutch speakers were recruited from Maastricht University. All participants (37 female and 23 male; mean age = 22, standard deviation = 3 years) had normal hearing, normal or corrected-to-normal vision, and received study credits or monetary compensation for participating. The study was approved by the university ethical research board. Participants were randomly selected to be in one of three groups; exposure to audiovisual/speech-reading stimuli, to lexical stimuli, or to both.

Materials

The materials for the experiment were modeled on those used previously by van Linden and Vroomen (2007). Digital audio and video of a female native Dutch speaker were recorded in a sound-proof booth. Recordings of the syllables /op/ and /ot/ were made, as well as a set of 16 Dutch words (e.g., *siroop* 'syrup', or *walnoot*, 'walnut') and 16 pseudo-words (e.g. *miloop, geroot*). The words varied in number of syllables and stress pattern and contained a range of segments, and the pseudo-words were matched in these respects to the real words and thereby creating varying input which could counteract possible selective adaptation effects from repetitive stimuli (Vroomen et al., 2007). All items were recorded with both /op/ and /ot/ endings.

A 10-step continuum ranging from clear /op/ to clear /ot/ was created using the Praat speech editing program (Boersma & Heuven, 2001), and adapted from a procedure devised by McQueen (1991), based on earlier work by Repp (1981). The endpoints of the continuum were excised from two recordings of the Dutch pseudo-words /soop/ and /soot/ with equal durations and a sampling frequency of 44 kHz. To prepare the continuum, the durations of the consonant (plosive) bursts of /op/ and /ot/ were spliced out and equated to 186ms, and the averaged pitch contour was calculated to replace the original. The intermediate sounds were created by concatenating the amplitudes of waveforms in 10% increments with each token after the first (e.g. 90% /op/ with 10% /ot/, etc.)**.** The preceding vowels of the two tokens were equated to 50ms and also interpolated using the same procedure as the consonants. As a result, the second and third formants of the vowel were systematically decreased from the /ot/-token to the /op/-token. All items of the continuum were then spliced onto a recording of /soo/, resulting in 10 items varying from /soop/ to /soot/. Multiple sets of lexical and audiovisual stimuli, or words and pseudo-words respectively, were created by with the middle steps of the continuum (steps 4, 5, 6, 7, and 8), which were most likely to be perceived as most ambiguous. These sounds were spliced into the stimuli at the zero-crossing closest to the last 50ms of the vowel preceding the final consonant, to eliminate any co-articulatory cues from the preceding vowel. The appropriate stimuli set was individually chosen

for each participant, during a categorization pre-test prior to the experiment, based on the sound perceived as /op/ or /ot/ for 50% of the responses, or as close as possible.

Lexical stimuli. Lexical stimuli were 16 Dutch words with word-final voiceless stop consonants, eight ending in /op/ and eight ending in /ot/. In the edited versions, the final phoneme was replaced with the ambiguous phoneme blend. Each set of eight contained one monosyllable, three disyllables, and three trisyllables. Stimuli lasted 1300ms on average with a standard deviation of 160ms. /p/-final words had an average word frequency of 421 per million, while /t/-words had an average word frequency of 367 per million.

Audiovisual stimuli. Audiovisual stimuli consisted of 16 videos of a speaker pronouncing Dutch pseudo-words, which were matched with the lexical stimuli for number of syllables. Pseudo-words were created using the program WinWordGen 1.0 for Dutch (Duyck et al., 2004), and were recorded with videos centered around the mouth of the speaker. The edited audio recordings containing the ambiguous final phoneme replaced the original audio of the video recordings. Based on the speaker's lip movements, eight of the videos indicated an /op/ ending, and the other eight an /ot/ ending. Each video lasted 1400 ms on average with a standard deviation of 100ms and no stimuli were longer than 1500 ms. Videos were approximately 24 frames per second with 1920x1080 pixels per frame.

Procedure

Participants were individually tested in a sound-proofed room. Stimuli were delivered using Presentation software and sound stimuli were presented through Philips Sensimetric earphones at a comfortable listening volume. Participants first underwent a pre-test in order to determine the step of the /op/-/ot/ continuum perceived to be most ambiguous. The items of the continuum were presented 100 times in total, with more presentations of medial steps than endpoints. For each sound, participants indicated with a button press if the sound resembled /ot/ or /op/. The step of the continuum reported as /op/ or /ot/ for 50% of trials, or as close as possible, was used to determine the appropriate stimuli set

to use in the exposure blocks of the experiment, as well as the sound used during the test blocks. All participants' perceived midpoints ranged between steps 4 and 8. All of the audio endings of the audiovisual and lexical stimuli would contain the individually selected ambiguous token. Individual ambiguous-token selection (as typically used for audiovisual studies since Bertelson et al., 2003) ensures that each participant will receive an equivalently effective stimulus set, but direct comparisons have shown that the perceptual learning process is unaffected by the choice between this method versus the simpler method (as typically used for audio-only studies since Norris et al., 2003) in which all participants receive the same ambiguous stimulus based on a pre-test with a separate group of listeners (Bruggeman & Cutler, in press).

Once the appropriate midpoint and its corresponding stimuli were selected, participants began the main experiment, which consisted of 32 blocks in total, each block beginning with eight exposure stimuli, followed by six test stimuli. Four unique exposure stimuli were presented, each repeated twice, and within a block, all had either /op/ or /ot/ endings so as to induce a bias in one direction at a time. For lexical stimuli, a gray fixation cross was present on the screen, while a black screen was present between video clips during audiovisual exposure. Each exposure trial lasted 1600ms in total, including the sound/video presentation and a brief silence. During the test phase, the ambiguous token from the continuum and its two neighbors (one more /op/-sounding, the other closer to /ot/) were each presented twice. After each sound presentation, the participants were prompted to respond with a button press to indicate the sound it most resembled (/op/ or /ot). Blocks were presented in pseudo-random order, where no more than two blocks with the same phoneme bias followed one another. Participants were randomly assigned to the three possible experimental conditions (lexical stimuli only, audiovisual stimuli only, or both types of stimuli). In the third group, blocks contained either lexical or audiovisual stimuli, and order was counterbalanced, such that the stimulus type changed every four blocks. For all three groups, the phoneme bias switched every one or two blocks. In total, 256 exposure trials were presented (for the third group, 128 of each exposure type), and 192 test trials. Examples of the testing procedure are shown in Figure 1.

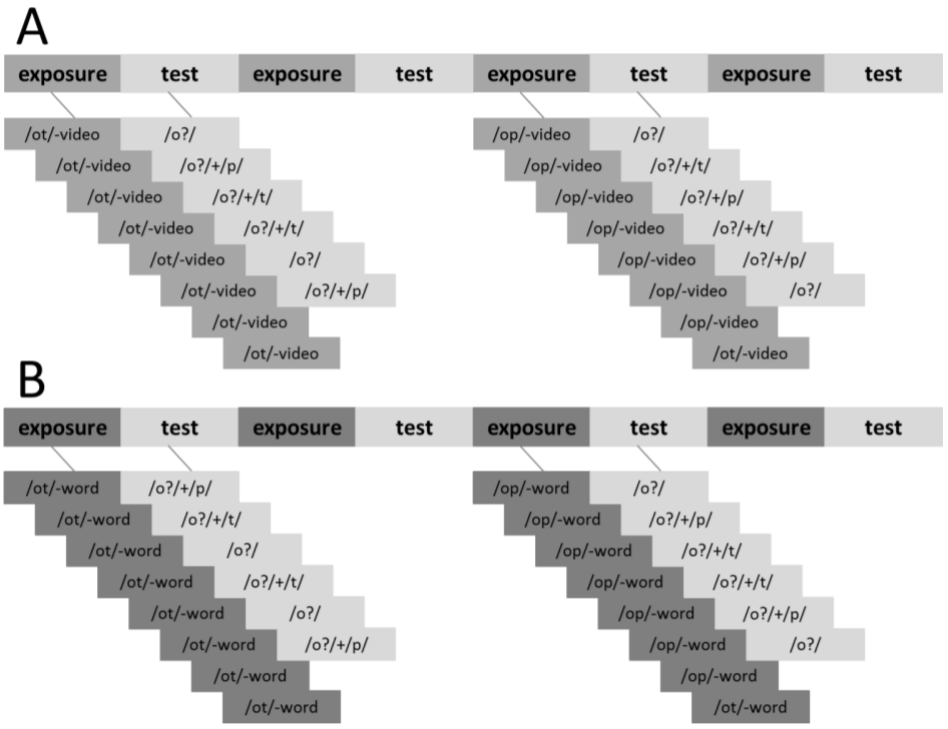


Figure 1. Examples of testing procedure. Participants received audiovisual (A), lexical (B), or both types of stimuli during exposure blocks, followed by test blocks. Participants who underwent single exposures would follow the procedure outlined in either panel A or B repeatedly for 32 blocks, while the third group received both A and B for the duration of the experiment. Any given exposure block aimed to elicit a bias towards either /p/ or /t/. Test items orders were randomized for every block.

Blocks alternated between presenting exposure and test stimuli. Exposure blocks consisted of eight items, either audio recordings of words or videos of pseudo-words, inducing a bias towards either /op/ or /ot/ during each block. Two groups received only one of the two types of stimuli (audiovisual or lexical), while a third group was presented with both types of stimuli (changing every 4 blocks). Each exposure block was followed by a test block containing the most ambiguous sound along the continuum, and its two perceptual neighbors, to which listeners responded depending on whether it was perceived as /op/ or /ot/.

Results

Pre-test Results

All participants underwent a pre-test to determine the most ambiguous sound along the /op/-/ot/ continuum. On average, the seventh step was closest to 50% perceptual midpoint and was the most frequent choice across participants. Pre-test results averaged across participants over the 10 steps are shown in Figure 2.

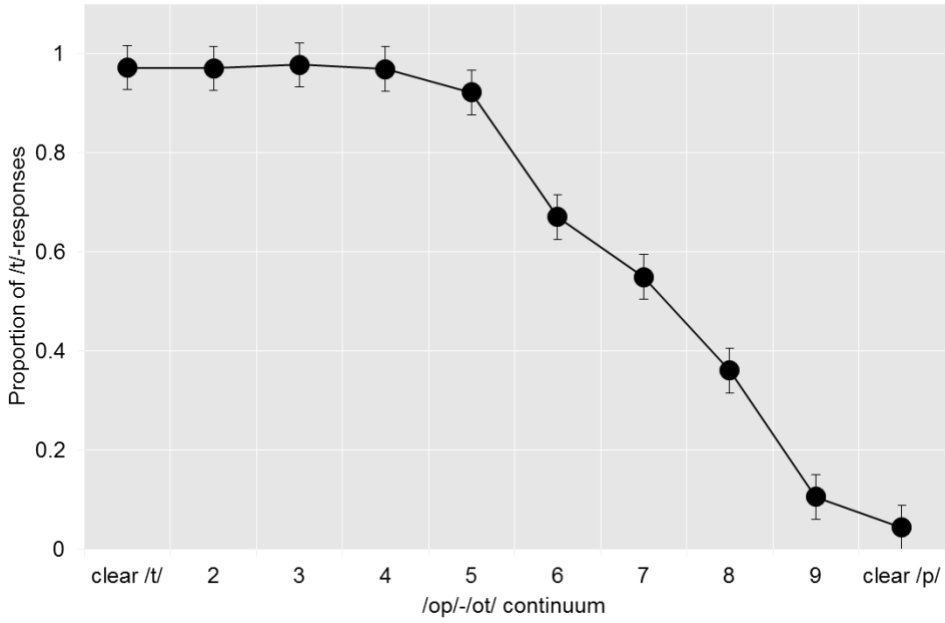


Figure 2. Pre-test results. Proportions of /t/-responses for each of the ten continuum sounds presented during the pre-test, averaged across all participants (n=60).

Perceptual Learning Results

Results were analyzed using the statistical package, *R,* with the lme4 library. All variables were entered into a generalized linear mixed-effects model with a logistic linking function for a binomial distribution. Four independent variables were entered into the model. *Phoneme bias* referred to the direction of the bias induced by the stimuli, being either /op/ or /ot/, while the *conditions* were either lexical or audiovisual. One out of the three participant groups was exposed to both audiovisual and lexical stimuli, while the other two groups only underwent

one form of exposure, so the model accounted for this with a variable of *switch,* by coding the two single exposure groups as one value and the third group (double exposure) as another. A variable was included for the three different *sounds* used during the test phases; the most ambiguous sound (selected during the pre-test) and its two surrounding neighbors from the continuum. Finally, the serial *block position* was also included, to see whether retuning effects varied from the start to the end of the experiment. All variables were numerically coded to be centered around 0. *Phoneme bias, condition, and switching* were entered as fixed effects, while the within-subject factors *phoneme bias, sounds, block position,* and an additional variable of *subject* were included as random effects as well. The dependent variable was the response to the test tokens, with "0" and "1" representing /op/ and /ot/, respectively. A maximal model containing all variables was created, as well as random slopes for all within-subjects variables and their interactions. The resulting model of best fit was: Response ~ 1 + Phoneme bias * Condition * Switching * Sound * Block position + (1 + Phoneme bias * Sound * Block position || Subject). Fixed effects correlations were checked to ensure the validity of the model, and all were less than 0.2.

The model showed a significant negative effect of the intercept, or general tendency to respond with /p/ across all test blocks. A significant main effect of *phoneme bias* and significant interactions between *phoneme bias* and *condition, block position* and *phoneme bias,* and between *block position, phoneme bias, and condition* were also found.

The main effect of *phoneme bias* revealed that more /t/ responses were seen after blocks biased towards /t/ than blocks biased towards /p/, which confirmed that listeners showed perceptual learning effects after audiovisual and lexical exposure. Bonferroni-corrected pairwise contrasts were performed on the factors in the 3-way interaction, between *block position, phoneme bias,* and *condition*. Significantly more /t/-responses were found after /t/-biased blocks than /p/-biased blocks in the audiovisual condition than in the lexical condition (shown in Figure 3). More specifically, significant differences between /t/-responses following /p/- and /t/-biased blocks were found across all block positions in the audiovisual condition ($p<0.002$), and for all blocks in lexical condition ($p<0.05$) although

slightly less at the first block (p=0.06). According to the model results, perceptual learning effects did not vary significantly across the testing session in either condition, although a statistically non-significant reduction in audiovisual recalibration was found from block 5 to block 6 (shown in Figure 4). As no significant main effects were found for the remaining factors of *switching* or *sound*, we concluded that perceptual learning effects did not vary due to either of these factors.

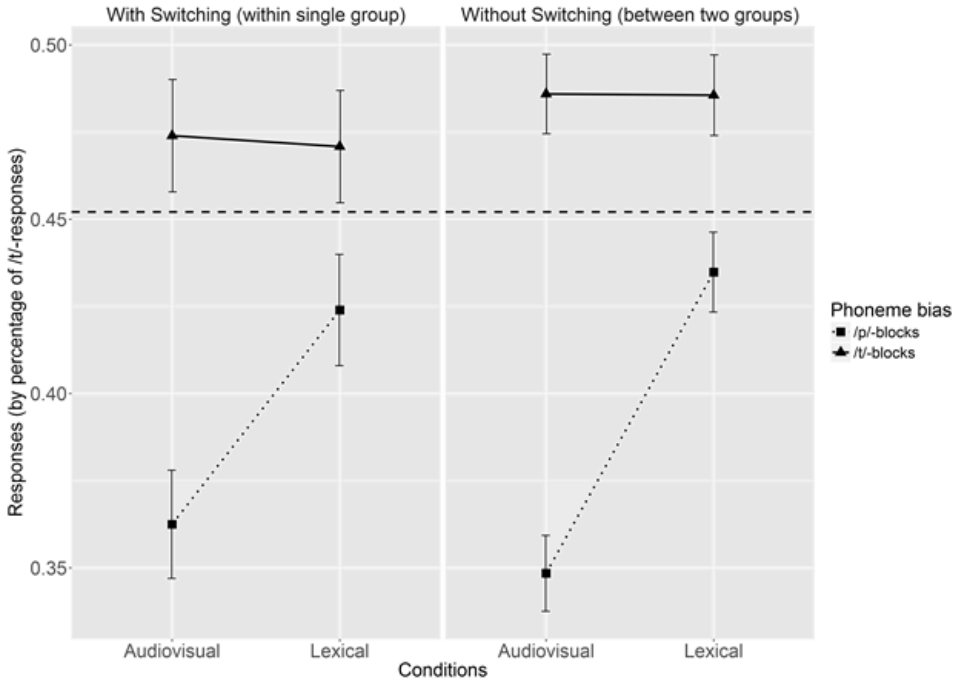


Figure 3. Audiovisual and lexical perceptual learning effects. Proportions of /t/-responses collapsed across the three test sounds, split by group that received both exposures (left panel) and single condition groups (right panel), and separated by exposure type. The dashed line indicates the pre-test average of /t/-responses over all participants to the individually-selected midpoint (=0.4528).

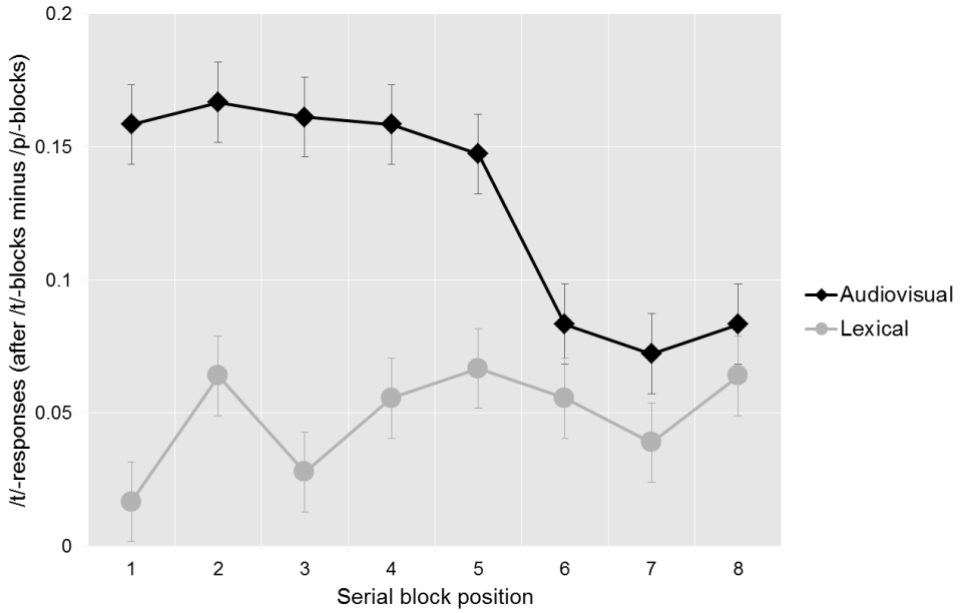


Figure 4: Subtracted perceptual learning effects from first to last block. The subtracted difference in the proportion of /t/-responses for each block are shown (/t/-responses following /t/-biased blocks minus /t/-responses after /p/-biased blocks). Lexical- and audiovisual-only groups are collapsed every two blocks, while responses from the double exposure group are averaged per block.

Discussion

In the present study, listeners could adjust phoneme boundaries using both lexical and audiovisual information, and switch between these two sources of information within a session. Comparison groups that underwent only one form of exposure showed similar levels of after-effects to the group that received both exposure types. Although the interleaved exposure blocks could have potentially led to interference between the two forms of perceptual learning, no such deficit was shown. Audiovisual recalibration and lexical retuning thereby appear to be separate processes, and do not necessarily interact with each other even while being measured in alternation and with the same phoneme pair. Neither form of perceptual learning showed significant variation over the course of the experiment, with the exception of lexical retuning effects at the first test block. A reduction in audiovisual recalibration was found between the fifth and sixth test blocks (from a 15% to 8% difference in subtracted /t/-responses), although it was not statistically significant. While audiovisual recalibration was more robust overall, it appears that

the effects may not be sustained with increasing numbers of test blocks, perhaps due to fatigue with repeated testing. Vroomen et al. (2004) have also reported reductions in audiovisual recalibration with increasing numbers of test items. Nevertheless, perceptual learning effects were largely stable throughout the testing session, and short and alternating exposures still led to observable effects on a block-to-block basis.

The experimental design used in the present study is in several ways more common in audiovisual recalibration experiments than in audio-only lexical retuning experiments. As noted, the two types of task typically differ in whether the ambiguous sound is customized to the individual participant, as in the present case, or is based on a separate pre-test with a separate participant group, as in most lexical retuning studies; but the two ambiguity determination methods have been shown to produce equivalent learning effects (Bruggeman & Cutler, 2019). In addition, lexical retuning studies commonly use longer exposure phases combined with a distractor activity, such as a lexical decision task, a counting task, or listening to a story (see Cutler et al., 2010, for an overview), and only induce a bias towards only one particular phoneme, instead of repeatedly changing the bias direction (Kraljic & Samuel, 2009). Lexical retuning effects with such designs have been found to be robust and even measurable up to 12 hours later (Eisner & McQueen, 2006). Lexical retuning effects in the present study may not have been as pronounced due to the experimental design, as listeners were continuously adapting the category boundary in two opposing directions. Although it is therefore arguable that such a design may be more suitable for audiovisual recalibration and may not have been optimal for inducing lexically-driven retuning, perceptual shifts in all conditions were still clearly evident.

The interleaved design still allowed lexical information to adjust phoneme boundaries using the same phoneme pair in either direction, with no reduction resulting from switching between exposure types, or due to short exposure blocks (which may not have given listeners adequate time to allow effects to accumulate). Krajlic & Samuel (2007) have reported that lexical retuning can take place in a speaker-specific manner, such that one particular phoneme pair is adjusted with one speaker, and another pair with another speaker, befitting the role of retuning

in social conversations with potentially many participants. Similarly, the flexibility of lexical retuning observed in the present study is consistent with the hypothesized value of lexical retuning for ensuring such adjustment to newly encountered interlocutors is rapid. Audiovisual recalibration can occur between multiple speakers (Mitchel, Gerfen, & Weiss, 2016) and even in two different directions by each ear (Keetels, Pecoraro, & Vroomen, 2015; Keetels, Stekelenburg, & Vroomen, 2016). In the present study, the original finding by van Linden & Vroomen (2007) was replicated, where lip-reading pseudo-words led to recalibration, but in addition, could take place while interleaved with lexical retuning. Note that the use of pseudo-words and interleaved exposure in our study may be the source of the lack of significance between test sounds (e.g. most /t/-responses for the most /t/-sounding token, etc.). Pseudo-words, rather than single syllables, were less specific to the phoneme at hand and could have led to a minor detraction in sound-specific recalibration.

The interleaved design would also lend itself well to neuroimaging studies. With the advancement of neuroimaging techniques such as functional MRI (fMRI), this design allows for exploration of the neural underpinning of multiple phoneme percepts induced by multiple cue types, all while presenting the same acoustic token during and after various contextual conditions. The paradigm could be used to explore how other phoneme pairs may fare, and how the learning effects would vary depending on the types of phonemes being manipulated (i.e. plosives/stops versus fricatives).

Audiovisual information proved more effective than lexical cues in inducing subsequent retuning effects, in line with prior findings (Lüttke et al., 2018; Mitterer & Reinisch, 2016; van Linden & Vroomen, 2007). This difference is predicted given the visual salience of the /p/-/t/ contrast (a bilabial versus an alveolar plosive) compared to the subtlety of the auditory difference between the same two sounds (both voiceless, both plosive). Any potential advantage to lip-reading cues is thus tied to the phonemes at hand, as they must be visually distinguishable in order for audiovisual cues to be a source of guidance. Prior studies have noted variation in the nature of lexical retuning across phoneme pairs in audio-only presentation (Kraljic & Samuel, 2007), as well within-pair differences

in shift effect size (Cutler et al., 2010); other contrasts may display varying patterns of relative effect. Notably, the difference in the magnitude of audiovisual and lexical perceptual learning effects in the present study was largely due to the difference in responses after /p/-biased blocks. The proportions of /t/-responses after audiovisual and lexical /t/-biased blocks were rather similar, whereas audiovisual /p/-blocks elicited fewer /t/-responses than lexical /p/ blocks. This strong /p/ response in the audiovisual /p/ blocks is as expected; not only is the /p-/t/ distinction visually salient, this salience is effectively carried by the /p/, so that the audiovisual contrast effectively amounts to plus versus minus lip closure. The possibility remains that the lexical information contained in the /p/-biased blocks may not have been as effective in inducing a shift in perception as the lexical information in the /t/-biased blocks; and as previously mentioned, each individual phoneme can vary in the extent that its boundary can be shifted by contextual cues. However, the reliability of the lip cues to /p/ for conversational participants is evidently the strongest effect.

The asymmetry between the sizes of the observed lexical and audiovisual retuning effects highlights how their intrinsic purposes may differ. Lexical cues can lead to retuning in response to static speaker characteristics that are unlikely to change, such as accents or idiosyncratic pronunciations unique to a particular speaker (Cutler et al., 2010). A speaker's pronunciation of a particular word is unlikely to change within a short amount of time. Lexical retuning effects may be more optimal in one particular direction, as was indeed seen in this study. In contrast, recalibration driven by speech-reading may be particularly useful and reliable in environmental circumstances that are not tied to a specific speaker, such as the presence of noise (Macleod & Summerfield, 1987; Massaro & Jesse, 2007; Sumby & Pollack, 1954). Thereby, the retuning resulting from audiovisual cues may be more malleable and more easily reconfigured across phonemes. In real-world scenarios, this means that listeners can attend to cues according to the needs of the situation, but are capable of switching between the two if required, as is suggested by the results of this study.

As noted in the methods section, the materials were designed to avoid selective adaptation effects, which typically occur when listeners have undergone

repeated exposure to a clear sound, but as a result are likely to perceive similar ambiguous sounds as a contrasting phoneme to the original (Eimas & Corbit, 1973). For example, after repeated presentations of clear auditory /op/, sounds on a continuum of /op/-/ot/ are more likely to be perceived as /ot/ than as /op/, i.e., the reverse of the exposure (Kleinschmidt & Jaeger, 2015; Vroomen et al., 2004; 2007). Selective adaptation can thus be viewed as the opposite of perceptual learning effects. Interestingly, one previous study (Samuel 2001) found that listeners who underwent short exposures to 10 words containing an ambiguous phoneme, similar to the design of the present study, showed selective adaptation effects during the subsequent test phases (ambiguous tokens presented without context). In this particular case, it is possible that the stimuli involved were insufficiently ambiguous, and could have been perceived as clear phonemes even when embedded in mismatching stimuli; this could potentially have induced a contrasting percept for a subsequently presented isolated sound. Importantly, the pattern of results in the current study clearly resemble perceptual learning, and not selective adaptation (which would have led to the opposite pattern of results, i.e. *fewer* /t/ responses after /t/-biased blocks than after /p/-biased blocks.). The observed results showed significantly more /t/-responses after /t/-biased blocks and significantly fewer after /p/-biased blocks. The average proportion of /t/-responses to the individually-selected midpoint (during the pre-test) was used to verify whether there were more or less /t/-responses after /t/- and /p/-biased blocks respectively, relative to the proportion of /t/-responses during the pre-test. As shown in Figure 3, more /t/-responses after /t/-biased blocks were seen compared to the baseline of the pre-test, and fewer /t/-responses compared to baseline were found after /p/-biased blocks as well. Therefore, it appears unlikely that listeners could have undergone selective adaptation effects, which would have been in the opposite directions compared to baseline as well. The study design that was adapted from Van Linden and Vroomen (2007) also reported lexical retuning effects with short exposures containing ambiguous sounds.

Overall, the results of the present study suggest that it is possible to compare audiovisual and lexical retuning under similar constraints and that listeners are capable of using both sources of information within a short period of

time to adjust phoneme boundaries. While audiovisual cues were, as expected, able to elicit larger recalibration effects, our results indicate that lexical retuning may be flexible in a manner not previously shown, using short exposures to create shifts in two opposing directions, all within a single session. Both lexical and audiovisual perceptual learning were achieved with interleaved exposure blocks and consequently, we suggest that phoneme boundary retuning can be utilized as a short-term solution for listeners' perceptual difficulties, and can be updated rapidly in accordance with the available contextual cues. The robustness of adaptability in speech perception becomes more apparent with every new investigative technique. In conclusion, the present technique would allow itself to be deployed in the future to explore the neural underpinnings of perceptual retuning, and to investigate potential differences in the multiple percepts induced by lexical and visual/speech-reading information.

Open Practices: The data and materials for all experiments are available at https://hdl.handle.net/10411/RWVUTN. None of the experiments were pre-registered.

References

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

Boersma, P., & Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International*, *5*(9/10), 341–347. https://doi.org/10.1097/AUD.0b013e31821473f7

Bruggeman, L. & Cutler, A. (2019). No L1 privilege in talker adaptation. *Bilingualism,*

Language and Cognition. https://doi.org/10.1017/S1366728919000646

Cutler, A. (2012). Native listening: the flexibility dimension. *Dutch Journal of Applied Linguistics*, *1*(2), 169–187. https://doi.org/10.1075/dujal.1.2.02cut

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. *Laboratory Phonology, 10*, 91–111. https://doi.org/10.1017/CBO9781107415324.004

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: a tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *36*(3), 488–499. https://doi.org/10.3758/BF03195595

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*. https://doi.org/10.1016/0010-0285(73)90006-6

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: stability over time. *The Journal of the Acoustical Society of America, 119*(4), 1950-1953. https://doi.org/10.1121/1.2178721

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology. Human Perception and Performance*, *6*(1), 110–125. https://doi.org/10.1037/0096-1523.6.1.110

Keetels, M., Pecoraro, M., & Vroomen, J. (2015). Recalibration of auditory

phonemes by lipread speech is ear-specific. *COGNITION*, *141*, 121–126. https://doi.org/10.1016/j.cognition.2015.04.019

Keetels, M., Stekelenburg, J. J., & Vroomen, J. (2016). A spatial gradient in phonetic recalibration by lipread speech. *Journal of Phonetics*, *56*, 124–130. https://doi.org/10.1016/j.wocn.2016.02.005

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

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15. https://doi.org/10.1016/j.jml.2006.07.010

Kraljic, T., & Samuel, A. G. (2009). Perceptual learning for speech. *Attention, Perception & Psychophysics*, *71*(3), 481–489. https://doi.org/10.3758/APP

Lüttke, C. S., Pérez-Bellido, A., & de Lange, F. P. (2018). Rapid recalibration of speech perception after experiencing the McGurk illusion. *Royal Society Open Science*, *5*(3), 170909. https://doi.org/10.1098/rsos.170909

Macleod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*(2), 131-141. https://doi.org/10.3109/03005368709077786

Massaro, D. W., & Jesse, A. (2007). Audiovisual speech perception and word recognition. *The Oxford handbook of psycholinguistics,* 19-36. https://doi.org/10.1093/oxfordhb/9780198568971.013.0002

McGurk, H., & MacDonald, M. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746.

McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(2), 433–443. https://doi.org/10.1037/0096-1523.17.2.433

Mitchel, A. D., Gerfen, C., & Weiss, D. J. (2016). Audiovisual perceptual learning with multiple speakers. *Journal of Phonetics*, *56*, 66–74. https://doi.org/10.1016/j.wocn.2016.02.003

Mitterer, H., & Reinisch, E. (2017). Visual speech influences speech perception

immediately but not automatically. *Attention, Perception & Psychophysics*, *79*(2), 660–678. https://doi.org/10.3758/s13414-016-1249-6

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

Repp, B. H. (1981). Perceptual equivalence of two kinds of ambiguous speech stimuli. *Bulletin of the Psychonomic Society*, *18*(1), 12-14. https://doi.org/10.3758/BF03333556

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, *12*(4), 348–351. https://doi.org/10.1111/1467-9280.00364

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26(2), 212-215. https://doi.org/10.1121/1.1907309

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

Vroomen, J. & Baart, M. (2009). Recalibration of phonetic categories by lipread speech: measuring aftereffects after a 24-hour delay. *Language and speech,* 52(2-3), 341-350. https://doi.org/10.1177/0023830909103178

Vroomen, J., & Baart, M. (2012). Phonetic recalibration in audiovisual speech. in M. M. Murray and M. T. Wallace (Eds.) The Neural Bases of Multisensory Processes. (pp. 363–379). Boca raton (FL): CRC Press.

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: contrasting build-up courses. *Neuropsychologia*, *45*(3), 572–577. https://doi.org/10.1016/j.neuropsychologia.2006.01.031

Vroomen, J., van Linden, S., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Communication*, *44*(1-4), 55–61. https://doi.org/10.1016/j.specom.2004.03.009

Appendix to Chapter 2

Table A1: Word list & definitions

Word	Meaning
Hoop	Hope
Aanloop	Approach
Afkoop	Surrender
Siroop	Syrup
Wanhoop	Despair
Geweerloop	Gun barrel
Horoscoop	Horoscope
Kussensloop	Pillowcase
Vloot	Fleet
Afsloot	Closed-off
Vennoot	Partner
Vergroot	Increases
Walnoot	Walnut
Hazelnoot	Hazelnut
Levensgroot	Life-size
Middenmoot	Mid-range

Table A2: Pseudo-word list

	/p/-final	/t/-final
One syllable	smoop	vroot
Two syllable	aaroop, miloop, onsoop, weloop	faloot, geroot, mevoot, neuloot
Three syllable	senkenloop, acenkoop, lakeroop	leuveroot, frieseloot, sanekoot

Table A3: Model results

Response ~ 1 + Block position*Phoneme bias*Condition*Switch*Sound + (1 + Block position*Phoneme bias*Sound || Subject)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.284633	0.067149	-4.239	2.25E-05	***
Block position	0.0200379	0.011382	1.76	0.07833	
Phoneme bias	0.2031495	0.02696	7.535	4.87E-14	***
Condition	0.0900698	0.056043	1.607	0.10802	
Switch	0.0096086	0.094854	0.101	0.91931	
Sound	-0.0188647	0.033564	-0.562	0.57408	
Block position*Phoneme bias	-0.0171471	0.007315	-2.344	0.01908	*
Block position*Condition	-0.0161389	0.010218	-1.58	0.11422	
Phoneme bias*Condition	-0.0945418	0.024829	-3.808	0.00014	***
Block position*Switch	0.0046131	0.016024	0.288	0.77343	
Phoneme bias*Switch	0.0261111	0.037972	0.688	0.49167	
Condition*Switch	0.0224973	0.059394	0.379	0.70485	
Block position*Sound	0.0099187	0.012321	0.805	0.42081	
Phoneme bias*Sound	-0.0185316	0.035223	-0.526	0.5988	
Condition*Sound	0.0098987	0.030862	0.321	0.74841	
Switch*Sound	-0.0694648	0.047303	-1.469	0.14196	
Block position*Phoneme bias*Condition	0.01454	0.007309	1.989	0.04666	*
Block position*Phoneme bias*Switch	-0.0067358	0.010292	-0.654	0.51281	
Block position*Condition*Switch	-0.0097491	0.012515	-0.779	0.43598	
Phoneme bias*Condition*Switch	-0.0199224	0.031677	-0.629	0.52939	
Block position*Phoneme bias*Sound	-0.0070255	0.015005	-0.468	0.63963	
Block position*Condition*Sound	0.0050501	0.011332	0.446	0.65585	
Phoneme bias*Condition*Sound	-0.006144	0.032067	-0.192	0.84806	
Block position*Switch*Sound	-0.0280237	0.017362	-1.614	0.10651	
Phoneme bias*Switch*Sound	-0.0367522	0.049621	-0.741	0.4589	
Condition*Switch*Sound	0.036105	0.03924	0.92	0.35752	
Block position*Phoneme bias*Condition*Switch	-0.0066206	0.010278	-0.644	0.51947	
Block position*Phoneme bias*Condition*Sound	-0.0002666	0.013329	-0.02	0.98404	
Block position*Phoneme bias*Switch*Sound	-0.0310419	0.021158	-1.467	0.14234	
Block position*Condition*Switch*Sound	-0.001025	0.014418	-0.071	0.94333	
Phoneme bias*Condition*Switch*Sound	0.0160756	0.040193	0.4	0.68918	
Block position*Phoneme bias*Condition*Switch*Sound	0.0020348	0.016035	0.127	0.89902	

*Significance: ***p<0.0001; *p<0.05*

3

Audiovisual and lexical cues do not additively enhance perceptual adaptation

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

Abstract

When listeners experience difficulty in understanding a speaker, lexical and audiovisual (or lip-reading) information can be a helpful source of guidance. These two types of information embedded in speech can also guide perceptual adjustment, also known as recalibration or perceptual retuning. With retuning or recalibration, listeners can use these contextual cues to temporarily or permanently reconfigure internal representations of phoneme categories to adjust to and understand novel interlocutors more easily. These two types of perceptual learning, previously investigated in large part separately, are highly similar in allowing listeners to use speech-external information to make phoneme boundary adjustments. This study explored whether the two sources may work in conjunction to induce adaptation, thus emulating real life, in which listeners are indeed likely to encounter both types of cue together. Listeners who received combined audiovisual and lexical cues showed perceptual learning effects similar to listeners who only received audiovisual cues, while listeners who received only lexical cues showed weaker effects compared to the two other groups. The combination of cues did not lead to additive retuning or recalibration effects, however, suggesting that lexical and audiovisual cues operate differently with regard to how listeners utilize them for reshaping perceptual categories. Reaction times did not significantly differ across the three conditions, so none of the forms of adjustment were either aided or hindered by processing time differences. Mechanisms underlying these forms of perceptual learning may diverge in numerous ways despite similarities in experimental applications.

Key words: recalibration, perceptual retuning, lip-reading, lexical, audiovisual

Introduction

Contextual information can impact what listeners perceive they are hearing, and can be helpful when, due to unfamiliar accents, background noise, or idiosyncratic pronunciations, speech is unclear. To adapt to such situations, listeners can draw on cues outside the speech signal, such as lip-reading information or lexical knowledge. The lexical Ganong effect, in which *?esk*, with an ambiguous /d/-/t/ blend replacing /d/, is often heard as *desk* (Ganong, 1980) shows how listeners' perception of an ambiguous phoneme is influenced by the word in which it occurs. Similarly, in the McGurk effect (where audio of /ba/ accompanying a speaker pronouncing /ga/ prompts a combined percept of /da/; McGurk & MacDonald, 1976), lip-reading information determines what listeners believe they are hearing.

Not only can lexical and audiovisual cues influence the perception of individual speech tokens, but each cue type can reconfigure the listener's perceptual system. Thus, listeners who heard words such as *giraffe* where an /f/-/s/ blend replaced the /f/ were then more likely to report this blend and similar sounds along a /f/-/s/ continuum as /f/ (Norris, McQueen, & Cutler, 2003). Likewise, listeners who viewed stimuli of a speaker pronouncing /aba/ paired with an auditory /aba/-/ada/ blend then reported hearing /aba/ even when given the ambiguous blend without visual context (Bertelson, Vroomen, & De Gelder, 2003). This audiovisual effect has been termed "recalibration" of phoneme decisions; it can be a conscious action by the listener, and indeed is even taught as a listening strategy (e.g., for taking dictation in second languages). In contrast, the lexical effect, of which listeners are typically unaware, has been referred to as "retuning" to interlocutor-specific articulation. We will here retain this distinction when referring to the two types of adjustment.

McGurk-style fusion percepts between auditory /b/ and visual /g/ (perceived together as /d/) can also result in similar shifts of the perceived boundary along a VOT continuum compared to isolated auditory stimuli without visual accompaniment (Green & Kuhl, 1989). The boundary shift determined by exposure to these fusion percepts can also vary depending on the phoneme pairs tested, such as in a /b/-/p/ pair compared to a /g/-/k/, even though both pairs also

vary along the same VOT dimension (Brancazio, Miller, & Paré, 2003). Visual representations of phonetic categories can also undergo shifts guided by lexical information (van der Zande, Jesse, & Cutler, 2013).

Perceptual recalibration and retuning have been extensively studied using lexical and lip-reading cues, but separately, and often with slightly differing experimental designs. Audiovisual recalibration can take place after exposure to as few as eight biasing stimuli (Vroomen, van Linden, de Gelder, & Bertelson, 2007). In contrast, lexically-driven retuning studies have typically used longer exposure phases with around 20 critical items, often embedded into a lexical decision task containing other filler words (see Cutler, Eisner, McQueen, & Norris, 2010, for a review), although Kraljic & Samuel (2007) showed that as few as 10 critical items can also induce lexical retuning. While audiovisual information can induce strong recalibration effects in a short period of time, the effects can dissipate quickly, with increasing numbers of categorization test items (Vroomen et al., 2004). However, lexical retuning appears robust and longer-lasting, measureable up to 24 hours later, again in designs with long exposure phases and usually by inducing a bias towards one particular phoneme (Eisner & McQueen, 2005, 2006; Kraljic & Samuel 2009). The two cue types may therefore operate on different timescales and thus require differing amounts of exposure (Eisner & McQueen, 2006; Vroomen, et al., 2007). Van Linden and Vroomen (2007) directly compared the two processes with matched designs but separate sessions for each cue type; audiovisual cues produced slightly larger effects than lexical cues.

Related research on audiovisual speech processing (see Massaro & Jesse, 2007; Rosenblum, 2010; for overviews) has established that lip-reading information can enhance speech comprehension, especially when the available auditory signal is unclear (Macleod & Summerfield, 1987; Sumby & Pollack, 1954). Lip-reading cues can also enhance the perception of certain types of phonetic information, such as the place of articulation, particularly for bilabial consonants, and can even be available to the listener prior to the onset of auditory phoneme cues (Massaro & Cohen, 1993). Such visual cues however affect reported perception more if a word results (e.g., auditory *besk* with visually presented *desk),* in contrast to auditory *desk*, visual *besk* where the visual choice makes a non-word (Brancazio, 2004). It

has been shown that visual cues can also enhance phoneme perception if visual information is available before auditory signal onset (Mitterer & Reinisch, 2016); but listeners performing a simultaneous interpretation task received no benefit from the presence of lip-reading cues when the auditory signal was clear and free of noise (Jesse, Vrignaud, Cohen, & Massaro, 2000).

Despite this substantial evidence of audiovisual effects on speech perception, prior research has not investigated the perceptual learning effects resulting from combined audiovisual and lexical cues. It remains unknown whether combined cues can induce effects larger than those elicited by either cue on its own. Redundant audiovisual and lexical cues, as listeners are most likely to encounter in real-life, could be more informative and could potentially lead to stronger adaptation effects than either cue in isolation. It may be beneficial for listeners to utilize as many available cues as possible when speech is unclear in order to interpret the ambiguous signal with ease, and thereby shift the underlying categories, rather than to rely on one source of information. However, visual cues may not significantly enhance perceptual learning if the auditory cues alone are sufficiently informative to the listener, or because the necessary exposure for a cue type has not been achieved. By mapping how these cues influence perceptual learning, we hope to enable the extension of current theories of speech perception to account for the role of such information in the process of speech comprehension and speaker adaptation. Although Massaro and Cohen (1993) and Rosenblum (2008) have argued that integrating acoustic and non-acoustic information is crucial for speech comprehension, accounts of speech perception have largely overlooked the contributions of non-acoustic information, especially with regard to perceptual learning (see Weber & Scharenborg, 2012 for a review).

The present study provides the first examination of phoneme boundary retuning given combined lexical and audiovisual information. If multiple sources of biasing information can be additive, we would expect to observe enhanced perceptual learning effects. However, if these cue types differ in the optimal conditions needed (i.e. differences in the amount of exposure needed for effects to be induced) or if one of the two cues can already induce ceiling-level results, then the combination may produce no benefit. To test this, three participant groups

were exposed to blocks of either lexical, audiovisual, or combined stimuli containing an ambiguous final phoneme, and in following test phases, ambiguous tokens were presented in a forced-choice categorization task.

Methods

Participants

Sixty participants were recruited from Maastricht University (32 female; mean age = 23, SD = 2.5 years). All were native Dutch speakers with normal hearing, normal or corrected-to-normal vision, and were compensated monetarily or with study credits. Participants were assigned to one of the three possible conditions (audiovisual, lexical, or combined) randomly, with 20 participants in each group.

Stimuli

Three sets of stimuli were constructed for the experiment. All stimuli were created using digital audio and video recordings of a female native Dutch speaker. A set of 16 real Dutch words and 16 pseudo-words were recorded with both /op/ and /ot/ endings, as well as two isolated recordings of the pseudo-words /soop/ and /soot/. For a full list of stimuli with their pronunciations, see Table 1.

The two syllables /op/ and /ot/ (long vowel plus voiceless stop-consonants) were the basis of a ten-step continuum, containing eight steps between these two endpoints, and were created using the *Praat* speech-editing program (Boersma & van Heuven, 2001) based on prior work by McQueen (1991). Similar procedures have been applied by Mitterer, Scharenborg, & McQueen (2013) and Reinisch & Holt (2014) using the STRAIGHT algorithm by Kawahara, Masuda-Katsuse, & De Cheveigné (1999). The two syllables were equated in duration with a 44kHz sampling frequency and with the original pitch contour replaced with an averaged one. The consonant bursts of the two syllables were scaled to have the same peak amplitude and were blended in 10% increments starting from one endpoint. Vowel durations were equated to 186 ms and morphed together in the same manner as consonants. These morphed syllables were spliced onto the ends of the recordings

of the words and pseudo-words, with joins made at the zero-crossing closest to the final 50 ms of the vowel to eliminate any co-articulatory cues.

The lexical stimuli were recordings of 16 Dutch words, with eight typically ending in /op/ and the other eight typically ending in /ot/, and matched in frequency and numbers of syllables. None of the selected words could be words if they ended in the alternative phoneme, and none contained any other occurrences of either target phoneme or, with a single exception, of the phonemes /b/ and /d/ that differ from the morphed phonemes only in voicing.

The pseudo-words generated for the audiovisual stimuli, using WinWordGen (Duyck, Desmet, Verbeke, & Brysbaert, 2004), were matched with the words for numbers of syllables. The audio endings of the pseudo-words replaced by the ambiguous steps from the /op/-/ot/ continuum. Video recordings of the pseudo-words contained only the speaker's mouth pronouncing the items to emphasize the lip-movements, half of which indicated /op/ ending and the other half /ot/ ending. Videos lasted 1200ms on average and no longer than 1500ms**.** The combined audiovisual-lexical stimuli consisted of the same words as the lexical stimuli, with the addition of the video of the speaker pronouncing the words (still centered around the speaker's mouth). These stimuli contained both lip-movement and lexical cues, while still containing the ambiguous audio ending. All videos had the original audio replaced with the corresponding audio token containing the ambiguous final phoneme.

Procedure

Participants were seated in front of a computer in a quiet testing room with audio presented over earphones set to a comfortable volume, using Presentation software (Neurobehavioral Systems). All participants first underwent a pretest by hearing the 10 continuum sounds ranging from /op/ to /ot/ to determine the sound most ambiguous to them. Stimuli sets that are tailored individually allow for equally ambiguous perception across participants, and are comparable in effect size to a pre-selected single midpoint used for all participants (Bruggeman & Cutler, 2019). Each sound was presented 10 times on average, with endpoint sounds presented six

to eight times while sounds towards the center were presented 10 to 12 times, and all sounds were presented in random order. Participants responded with a button press for each sound depending on whether they perceived it as /op/ or /ot/. The most ambiguous sound, perceived as either /op/ or /ot/ for the closest average to 50% of responses, was used to select the particular participant's stimuli set for the retuning experiment.

Following the pre-test, exposure and test stimuli were presented in alternating blocks, for a total of 32 exposure blocks and 32 test blocks. Exposure blocks contained four unique stimuli, each presented twice, for eight items total. Either audio-only recordings of words, videos of pseudo-words, or videos of words were presented in the lexical, audiovisual, and combined conditions, respectively. For the lexical condition, a gray fixation cross was centered on the screen during the eight audio-only trials. In the audiovisual and combined conditions, eight videos were presented during the exposure block. Each individual exposure block induced a bias towards one particular phoneme, (i.e. towards /op/ by presenting only words ending in /op/ in the lexical condition). The phoneme bias of the exposure block was pseudo-randomly alternated every one or two blocks, with 16 blocks inducing a bias towards /p/ and the other 16 towards /t/, in order to enable a within-subject measure of perceptual learning results (rather than two separate groups; i.e. one group receiving ambiguous /p/ and the other receiving ambiguous /t/).

A test block followed every exposure block in all conditions, consisting of a categorization task upon the individually-selected ambiguous token from the /op/-/ot/ continuum, and its immediately preceding and following sounds: one more /p/-sounding, one more /t/-sounding. Each sound was presented twice, for six presentations total. After each sound, participants signaled with a button press what they reported hearing (/p/ or /t).

Exposure and test trials lasted 1600 ms each, while test trials were followed by a 1400 ms gap for response. For test blocks in all conditions, a red fixation cross was presented during the sound presentation, followed then by a green fixation cross prompting the participant's response. Figure 1 provides an overview of the experimental procedure.

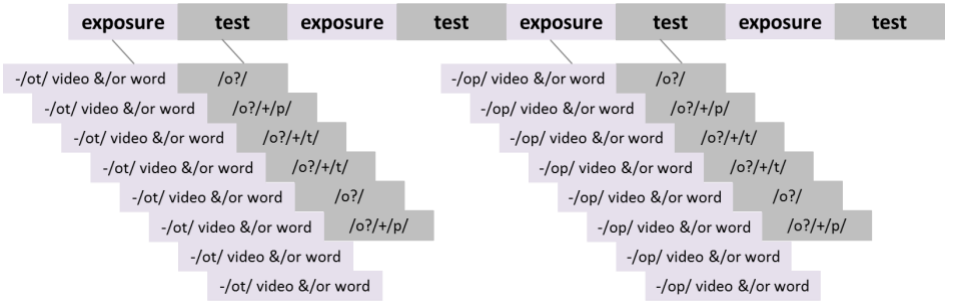


Figure 1. Example of blocked exposure-test procedure. In exposure blocks, listeners were presented with eight stimuli (audio recordings of words, videos of pseudo-words, or the combination [videos of words], depending on assigned condition), biased towards /op/ or /ot/ per block. The phoneme bias in each exposure block changed every one or two blocks. In the test blocks following each exposure, listeners heard the most ambiguous sound and its two neighbors (one more /p/-sounding and one more /t/-sounding), and responded whether each sound resembled /op/ or /ot/. The procedure depicted was repeated eight times over the course of the experiment (with pseudo-randomized alternation of phoneme bias in the exposure blocks), such that listeners would be consistently shifting the boundary between the two phoneme endpoints throughout the session.

A separate group of six listeners provided goodness ratings of all of the exposure stimuli (lexical, audiovisual, and combined). Participants were presented with each item three times, and rated them on a scale from 1 to 7, with 1 indicating a clear /p/-ending and 7 indicating a clear /t/-ending (4 if the item was ambiguous). The resulting ratings are shown in Table A2 in the Appendix. These listeners replicated the asymmetry reported by van Linden and Vroomen (2007), where audiovisual stimuli received the highest goodness ratings, followed by the combined stimuli, and with lexical items receiving relatively lower ratings.

Results

Pre-test responses

Responses during the pre-test were averaged per test sound to determine the most ambiguous token per subject, in order to determine the most appropriate stimulus set. On average, the seventh step was marked as /t/ for 50% of responses and most ambiguous for the majority of participants. Pre-test results are shown in Figure 2. For the individually selected midpoints, the average of /t/ responses for the selected token were 0.41458, 0.44792, and 0.38333, for the audiovisual, lexical, and combined groups respectively.

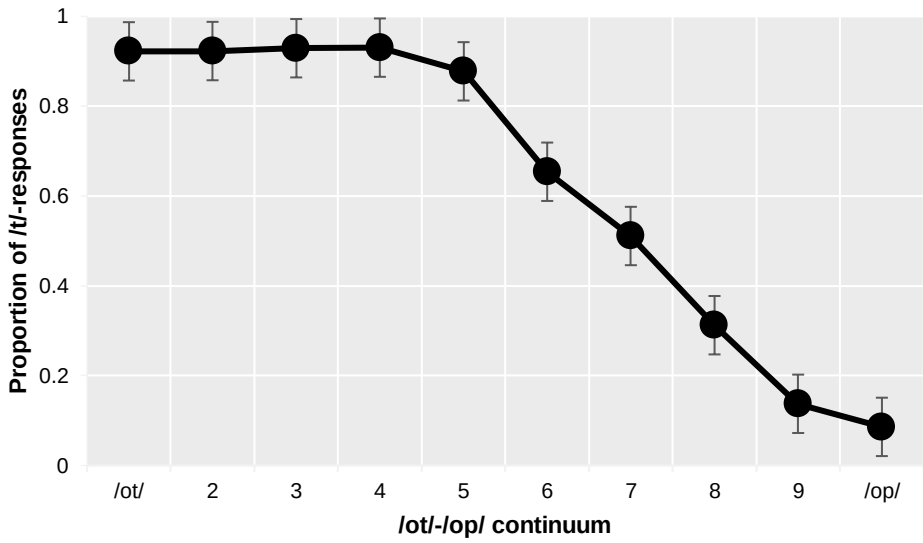


Figure 2. Pre-test /t/-responses averaged across participants (n = 60) for each sound along the continuum, ranging from clear /ot/ to clear /op/.

Retuning responses

Responses during test blocks were entered into a generalized linear mixed model, using the lme4 package in R. *Phoneme bias* during the preceding exposure blocks, *condition* (lexical, audiovisual, or combined), *sound* (the three types of sounds presented during test blocks), and block position (collapsed to range from 1 to 8) were entered into the model as fixed effects. All factors were coded to be centered around zero, except for the test block responses, which were coded as 0 (for /p/) and 1 (for /t/). Within-subjects factors including *phoneme bias, sound, and block position* in addition to *subjects* were entered as random effects. Random slopes were fitted for within-subjects factors of *phoneme bias, sound, and block position,* as well as their interactions. All variables were coded to be centered around zero, but responses were entered as zeroes (/p/) and ones (/t/). The model was created by entering all possible random effects and interactions, while ensuring that the model converged, where all fixed effects correlations were no larger than 0.4. The resulting model was: Response ~ 1 + Phoneme bias * Condition * Sound * Block position + (1 + Phoneme bias * Sound * Block position || Subject; see Table A3).

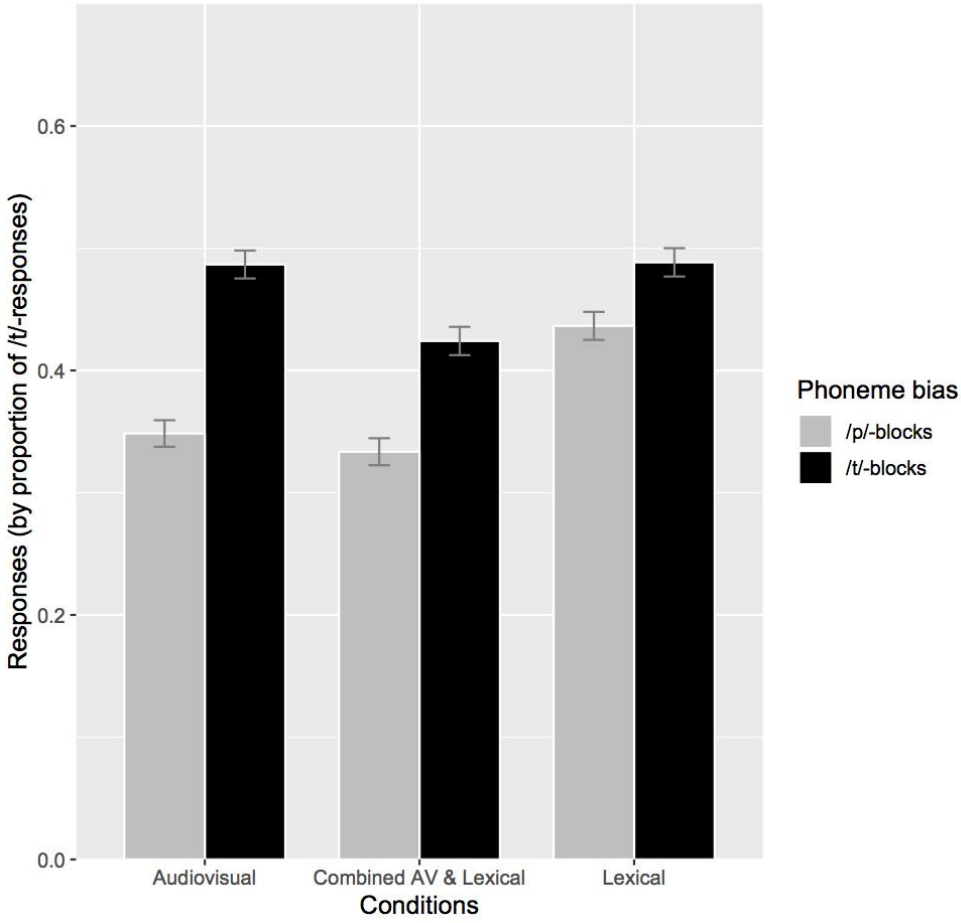


Figure 3. Recalibration/retuning effects across test sounds for each condition, by proportions of /t/-responses during test blocks, separately by phoneme bias during exposure block.

Effects across the three conditions are depicted in Figure 3. The model showed a significant main effect of *phoneme bias* and the intercept, as well as significant interactions between *phoneme bias* and *condition* and between *phoneme bias* and *block position*. Due to the significant intercept, participants generally had a bias towards responding with /p/ throughout the experiment. However, the main effect of *phoneme bias* indicated that participants responded with significantly more /t/ following /t/-biased exposure, and with /p/ following /p/-biased exposure, demonstrating the retuning/recalibration effect. Due to the interactions between *phoneme bias* and *condition* as well as *phoneme bias* and *block position*, post-hoc t-tests were conducted, and showed that the effect of *phoneme bias* differed between

the three conditions and over the series of blocks. On average across the three test sounds, the difference in /t/-responses following /t/- and /p/-biased blocks was larger for the audiovisual and combined conditions ($p < 0.0001$) while to a lesser extent in the lexical condition ($p < 0.01$). In addition, the difference in /t/-responses between /t/- and /p/- blocks varied over the block positions, and was significant for all positions in the audiovisual and lexical conditions ($p<0.0001$), but in the lexical condition, was significant for all blocks ($p<0.05$) except for the 5th and 7th blocks ($p=0.07$ and $p=0.1316$). The subtracted percentage of responses between /t/- and /p/- blocks per block position is shown in Figure 4. The factor *sound* showed no significant main effect or interactions; i.e., the three test sounds did not differ significantly in the proportion of responses elicited.

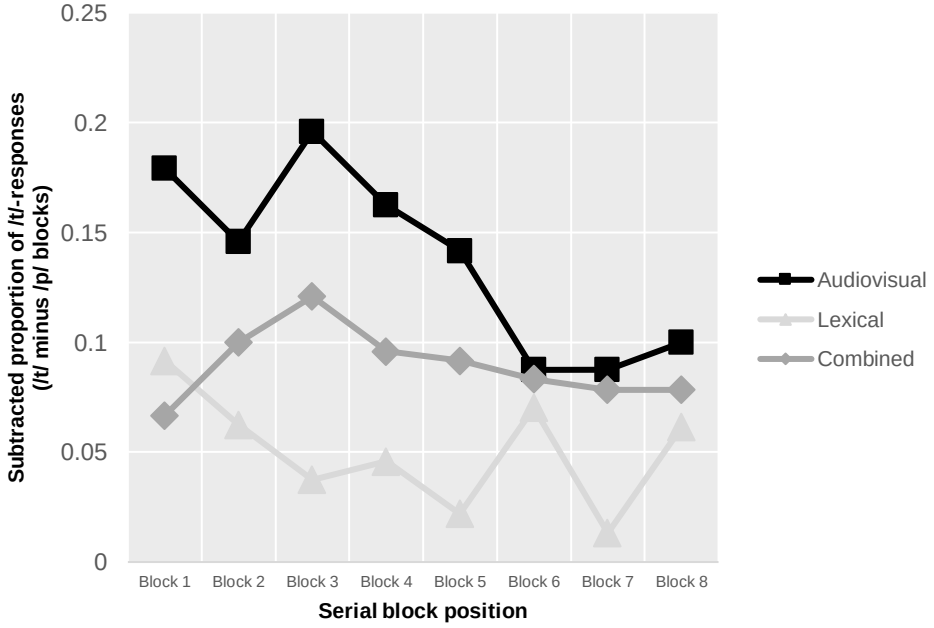


Figure 4. Perceptual learning effects from first to last block. Subtracted proportion of /t/-responses (i.e. /t/-responses after /t/-blocks minus /t/-responses after /p/-blocks) are shown for each block position, separated by the three conditions (audiovisual, lexical, and combined).

Discussion

In this study, participants underwent three forms of phoneme boundary adjustments using lexical, audiovisual, or combined stimuli. All three groups successfully showed perceptual learning effects in accordance with the exposure stimuli presented. Audiovisual and combined groups showed stronger effects than the lexical group, but the three groups did not differ significantly from each other. Combined cues resulted in perceptual learning effects similar to audiovisual cues and were numerically larger than lexical retuning effects. An overall bias towards /p/ was observed in all conditions, most likely as a result of the visually noticeable place of articulation of /p/ (bilabial) compared to /t/ (alveolar), as well as the greater lexical information provided by /p/ in word-final positions than /t/. In Dutch, /t/ is often a morphological verb suffix, and does not always carry as useful lexical information in the same manner as /p/. Nevertheless, significant shifts were seen following the phoneme-biased exposure blocks and relative to the pre-test averages to the individually selected ambiguous token as well. From block to block, there was some variation in the amount of perceptual learning effects, particularly as lexical retuning showed some slight reductions in effects (at the fifth and seventh block positions).

Although lexical retuning took place in the study, the observed effects were weaker than those of audiovisual and combined effects**.** The fast, alternating design used in this study may not have provided optimal conditions to elicit such retuning. Previous studies of lexical retuning have often used a single exposure phase, biased only towards one particular phoneme, embedded in a distractor task containing filler words as well (Cutler, et al., 2010). In contrast, in the present study, the phoneme bias was changing throughout the experiment, and was presented in short exposure blocks quickly followed by test blocks. With this design, lexical cues may have insufficient time to build up their potential retuning effects, which are potentially measurable up to 24 hours later in more optimal designs (Eisner & McQueen, 2006). The smaller magnitude of the lexical retuning effect seemed to be driven largely by the lack of /p/-responses after /p/-biased blocks, more so than the /t/-responses after /t/-biased blocks (see Figure 3). The greater proportion of /p/-

responses following audiovisual and combined exposure may result from the salience of the visual /p/ more strongly indicating the final /p/ in comparison to the lexical /p/. This finding may also demonstrate the relative rigidity of lexical retuning under the constraints of this study design. Lexical retuning presumably exists for situations involving an unfamiliar pronunciation or accent in which the phoneme bias is in a constant direction. When listeners must continuously update the phoneme category boundary, as in the present study, they may experience difficulty in shifting the boundary in differing directions rather than only in one. Still, lexical retuning can still be accomplished under these restricted conditions of the current study, albeit less robustly.

Audiovisual and combined audiovisual-lexical recalibration were comparable in the obtained effects, and both were larger in comparison to lexically-guided retuning. Notably, combined audiovisual/lexical cues did not result in larger learning effects than audiovisual cues. Although real-life circumstances were more closely emulated by combining lexical and audiovisual cues, which could also allow listeners to readjust faster and more effectively, no such benefit was observed in the pattern of results. It was hypothesized that the compounded cues could have led to an enhanced effect, as listeners had two informative sources available to steer their perceptual adjustments. Instead, the results pointed towards an averaging effect between lexical retuning and audiovisual recalibration. The lexical cues may not have provided any additional benefit to the audiovisual cues during the listeners' perception of the ambiguous phonemes. If the audiovisual cues alone were enough to induce a perceptual shift in the listeners, then the lexical cues may not have given the listeners any additional support not already available. Audiovisual cues may have therefore produced a ceiling effect, which the addition of lexical cues could not further enhance. Audiovisual integration can also occur at an earlier stage than lexical access (Ostrand et al. 2016), and as the phoneme pair could be distinguished visually by the place of articulation (a bilabial /p/ versus an alveolar /t/) and at an earlier point in time as well, then the subsequent lexical information may not have been able to a further enhance perception. However, relative contributions of visual and lexical information while interpreting ambiguous sounds may also be phoneme-dependent. For example, confusable

phonemes sharing the same place of articulation (e.g., /b/, /p/) may be aided more by lexical cues, whereas confusable phonemes that are visually discrepant (e.g., /m/, /n/) may benefit more from lip-reading cues. Thus, adaptation effects may be driven by whichever cues are most salient in a given situation.

Perceptual learning effects per block showed some variation, especially for lexical retuning at the fifth and seventh block positions. As previously mentioned, the design may not be optimal for maximizing lexical retuning, and the variation is a likely consequence. Audiovisual recalibration also showed variation over the blocks, and seemed to decrease from the sixth block towards the end, although not significant statistically. Combined audiovisual-lexical learning appeared more stable over the course of the blocks and less prone to variation. Overall, all perceptual learning effects showed some decreases with prolonged testing, as Vroomen et al. (2004) have previously reported.

Reaction times across the three groups also did not differ significantly (see figure in Appendix). Previously, Brancazio (2004) reported slower responses associated with a visual cue versus an auditory cue for a phoneme within a word, so in the present study we were also interested in whether slower responses would arise with combined audiovisual and lexical effects compared to lexical effects alone. However, Brancazio (2004) did not include phonemes presented without audiovisual or lexical context, whereas in the present study, ambiguous phonemes were presented in test blocks isolated from audiovisual and lexical cues. Our results suggest that Brancazio's finding reflected a processing time increase to allow for lexical activation; responses in the case of perception of isolated phonemes have no need for such activation, and indeed we found no indication of such reaction time differences.

The combination of ambiguous audio, rather than clear audio, with the audiovisual and lexical cues appears effective in inducing phoneme boundary shifts. One previous study combined both audiovisual and lexical cues in McGurk-style fusion percepts (e.g. auditory *armabillo* paired with visual *armagillo* resulting in a percept of the word *armadillo)* but these stimuli did not induce significant perceptual shifts (Samuel & Lieblich, 2014). McGurk-style fusion stimuli can lead to perceptual shifts (Lüttke, Pérez-Bellido, & de Lange, 2018; Roberts & Summerfield,

1981; Saldaña & Rosenblum, 2005), but such stimuli often combine clear audio of a syllable (/ba/) with an incongruent video of another syllable (such as /ga/), leading to an entirely new percept (/da/). The combination of lexical and audiovisual cues in these McGurk percepts may not allow for perceptual adjustments. In the present study, however, the combination of ambiguous audio with audiovisual and lexical information did prompt a shift in the perceptual boundary. Some relevant acoustic information appears to be necessary to activate lexical and audiovisual representations that allow for recalibration and retuning, even when auditory signals are ambiguous.

Our results show that lexical and audiovisual cues in combination do not jointly enhance perceptual learning. We suggest that the inherent differences in timing between audiovisual and lexical cues is likely to play an important role in how the two cues are integrated to elicit perceptual adjustments. The discrepancy between audiovisual and lexical effects may also be indicative of differences in their underlying structures and networks. Despite the clear similarities between the perceptual learning effects, lexical and audiovisual information seem to diverge in how they operate to adjust phoneme boundaries.

Open Practices: The data and materials for the experiments reported here are available at (https://hdl.handle.net/10411/UT7PGU) and none of the experiments were preregistered.

References

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

Boersma, P., & van Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International*, *5*(9/10), 341–347. https://doi.org/10.1097/AUD.0b013e31821473f7

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology. Human Perception and Performance, 30*(3), 445-463.

Brancazio, L., Miller, J. L., & Paré, M. A. (2003). Visual influences on the internal structure of phonetic categories. Perception and Psychophysics, 65(4), 591–601. https://doi.org/10.3758/BF03194585

Bruggeman, L. & Cutler, A. (2019). No L1 privilege in talker adaptation. *Bilingualism, Language and Cognition*. https://doi.org/10.1017/S1366728919000646

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. *Laboratory Phonology, 10*, 91–111. https://doi.org/10.1017/CBO9781107415324.004

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: a tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, and Computers*, *36*(3), 488–499. https://doi.org/10.3758/BF03195595

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

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: stability over time. *The Journal of the Acoustical Society of America, 119*(4), 1950-1953. https://doi.org/10.1121/1.2178721

Ganong, W. F. (1980). Phonetic categorization in auditory word perception.

Journal of Experimental Psychology. Human Perception and Performance, *6*(1), 110–125. https://doi.org/10.1037/0096-1523.6.1.110

Green, K. P., & Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*, *45*(1), 34-42. https://doi.org/10.3758/BF03208030

Jesse, A., Vrignaud, N., Cohen, M. M., & Massaro, D. W. (2000). The processing of information from multiple sources in simultaneous interpreting. *Interpreting*, *5*(2), 95-115. https://doi.org/10.1075/intp.5.2.04jes

Kawahara, H., Masuda-Katsuse, I., & De Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, *27*(3), 187–207. https://doi.org/10.1016/S0167-6393(98)00085-5

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56*(1), 1–15. https://doi.org/10.1016/j.jml.2006.07.010

Kraljic, T., & Samuel, A. G. (2009). Perceptual learning for speech. *Perception & Psychophysics*, *71*(3), 481–489. https://doi.org/10.3758/APP

Lüttke, C. S., Pérez-Bellido, A., & de Lange, F. P. (2018). Rapid recalibration of speech perception after experiencing the McGurk illusion. *Royal Society Open Science*, *5*(3). https://doi.org/10.1098/rsos.170909

Macleod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 21*(2), 131-141. https://doi.org/10.3109/03005368709077786

Massaro, D. W., & Cohen, M. M. (1993). Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables, *Speech Communication, 13*(1-2), 127–134.

Massaro, D. W., & Jesse, A. (2007). Audiovisual speech perception and word recognition. *The Oxford Handbook of Psycholinguistics,* 19-36. https://doi.org/10.1093/oxfordhb/9780198568971.013.0002

McGurk, H., & MacDonald, M. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746.

McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(2), 433–443. https://doi.org/10.1037/0096-1523.17.2.433

Mitterer, H., & Reinisch, E. (2016). Visual speech influences speech perception immediately but not automatically. *Perception & Psychophysics*, *79*(2), 660–678. https://doi.org/10.3758/s13414-016-1249-6

Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition*, *129*(2), 356–361. https://doi.org/10.1016/j.cognition.2013.07.011

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

Ostrand, R., Blumstein, S. E., Ferreira, V. S., & Morgan, J. L. (2016). What you see isn't always what you get: auditory word signals trump consciously perceived words in lexical access. *Cognition*, *151*, 96–107. https://doi.org/10.1016/j.cognition.2016.02.019

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign accented speech and its generalization. *Journal of Experimental Psychology. Human Perception and Performance*, *40*(2), 539–555. https://doi.org/10.1037/a0034409

Roberts, M., & Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Perception & Psychophysics*, *30*(4), 309–314. https://doi.org/10.3758/BF03206144

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, *17*(6), 405–409. https://doi.org/10.1111/j.1467-8721.2008.00615.x

Rosenblum, L. D. (2010) *See what I'm saying: the extraordinary powers of our five senses.* W. W. Norton & Company, New York, NY.

Saldaña, H. M., & Rosenblum, L. D. (1994). Selective adaptation in speech perception using a compelling audiovisual adaptor. *The Journal of the Acoustical Society of America, 95*(6), 3658-3661.

https://doi.org/10.1121/1.409935

Samuel, A. G., & Lieblich, J. (2014). Visual speech acts differently than lexical context in supporting speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(9), 1740–1747. https://doi.org/10.3174/ajnr.A1650.Side

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*. *26*(2), 212-215. https://doi.org/10.1121/1.1907309

Van der Zande, P., Jesse, A., & Cutler, A. (2013). Lexically guided retuning of visual phonetic categories. *The Journal of the Acoustical Society of America*, *134*(1), 562–571. https://doi.org/10.1121/1.4807814

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

Vroomen, J., Van Linden, S., Keetels, M., De Gelder, B., & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Communication*, *44*(1-4 SPEC. ISS.), 55–61. https://doi.org/10.1016/j.specom.2004.03.009

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective speech adaptation in auditory-visual speech perception: contrasting build-up courses. *Neuropsychologia, 45*(3), 572-577.

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(3), 387–401. https://doi.org/10.1002/wcs.1178

Appendix to Chapter 3

Table A1: Words & pseudowords

/op/-words:	
Hoop	[hoʊp]
Siroop	[sɪʀoʊp]
Aanloop	[a:nloʊp]
Afkoop	[ɑfkoʊp]
Wanhoop	[ʋɑnhoʊp]
Geweerloop	[ɣəʋe:rloʊp]
Horoscoop	[ɦɔʀɔscoʊp]
Kussensloop	[kʏsənsloʊp]

/ot/-words:	
Vloot	[vloʊt]
Afsloot	[ɑfsloʊt]
Vennoot	[vɛnoʊt]
Vergroot	[vəʀɣʀoʊt]
Walnoot	[ʋa:lnoʊt]
Hazelnoot	[ɦɑzəlnoʊt]
Levensgroot	[lɛvənsɣʀoʊt]
Middenmoot	[mɪdənmoʊt]

/op/-pseudowords:	
Smoop	[smoʊp]
Aaroop	[a:ʀoʊp]
Miloop	[mɪloʊp]
Onsoop	[ɔnsoʊp]
Weloop	[ʋəloʊp]
Acenkoop	[ɑsəŋkoʊp]
Lakeroop	[lɑkəʀoʊp]
Senkenloop	[sɛŋkənloʊp]

/ot/-pseudowords:	
Vroot	[vʀoʊt]
Faloot	[fɑloʊt]
Geroot	[ɣəʀoʊt]
Mevoot	[məvoʊt]
Neuloot	[nø:loʊt]
Frieseloot	[fʀisəloʊt]
Leuveroot	[lø:vəʀoʊt]
Sanekoot	[sɑnəkoʊt]

Table A2: Stimuli ratings
Ratings of the stimuli (n=6) on a scale from 1-7 (1 for clear /p/, 7 for clear /t/, 4 for ambiguous).

	/op/-ending	/ot/-ending
Lexical (audio words)	3.2917	4.9167
Audiovisual (audio+video pseudowords)	2.3611	5.5625
Combined (audio+video words)	2.6458	5.4028

Table A3: Retuning/recalibration results
Model: Response ~ 1 + Phoneme bias * Condition * Sound * Block position + (1 + Phoneme bias * Sound * Block position || Subject)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.38632	0.077687	-4.973	6.60E-07	***
Phoneme	0.219164	0.027841	7.872	3.49E-15	***
Condition	0.098318	0.095028	1.035	0.30085	
Sound	0.004709	0.034309	0.137	0.89083	
Block	0.021641	0.011294	1.916	0.05534	
Phoneme*Condition	-0.10528	0.033877	-3.108	0.00189	**
Phoneme*Sound	-0.02038	0.037177	-0.548	0.58361	
Condition*Sound	0.031938	0.041826	0.764	0.4451	
Phoneme*Block position	-0.01588	0.007372	-2.154	0.03125	*
Condition*Block position	-0.02189	0.013761	-1.591	0.1117	
Sound*Block position	0.010039	0.013084	0.767	0.44291	
Phoneme*Condition*Sound	0.013674	0.045333	0.302	0.76292	
Phoneme*Condition*Block position	0.011169	0.008955	1.247	0.21234	
Phoneme*Sound*Block position	-0.01966	0.01462	-1.345	0.17866	
Condition*Sound*Block position	0.006478	0.015955	0.406	0.68475	
Phoneme*Condition*Sound*Block	0.003955	0.017842	0.222	0.82458	

*Significance: ***p<0.0001; **p<0.01; *p<0.05*

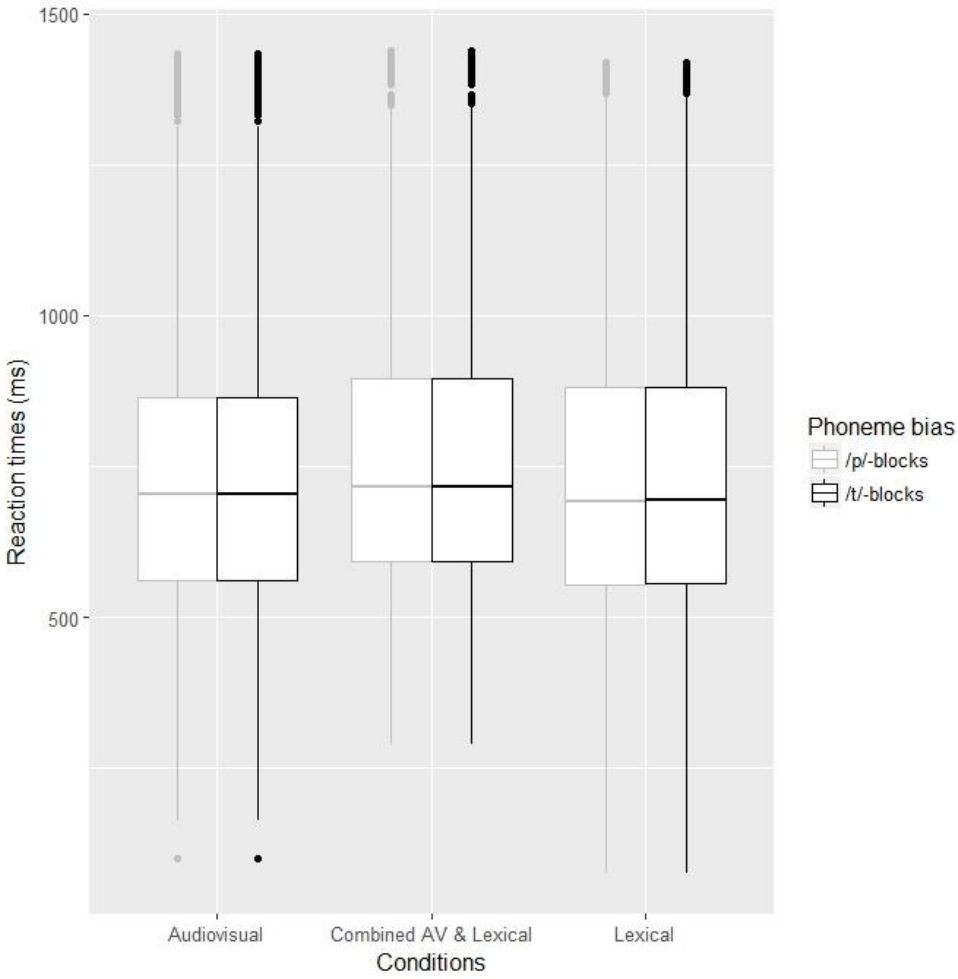


Figure A1. Reaction times across the three testing groups, separately by phoneme bias during the preceding exposure block.

4

Neural correlates of phonetic adaptation as induced by lexical and audiovisual context

Ullas, S., Hausfeld, L., Cutler, A., Eisner, F., & Formisano, E. (under review). Neural correlates of phonetic adaptation as induced by lexical and audiovisual context. *Journal of Cognitive Neuroscience*.

Abstract

When speech perception is difficult, one way listeners adjust is by reconfiguring phoneme category boundaries, drawing on contextual information. Both lexical knowledge and lip-reading cues are used in this way, but it remains unknown whether these two differing forms of perceptual learning are similar at a neural level. The present study compared phoneme boundary adjustments driven by lexical or audiovisual cues, using ultra-high field 7T functional MRI. During imaging, participants heard *exposure* stimuli and *test* stimuli. Exposure stimuli for lexical retuning were audio recordings of words, and for audiovisual recalibration were audio-video recordings of lip-movements during utterances of pseudowords. Test stimuli were ambiguous phonetic strings presented without context and listeners reported what phoneme they heard. Reports reflected phoneme biases in preceding exposure blocks (e.g., more reported /p/ after /p/-biased exposure). Analysis of corresponding brain responses indicated that both forms of cue use were associated with a network of activity across the temporal cortex, plus parietal, insula, and motor areas. Audiovisual recalibration also elicited significant occipital cortex activity despite the lack of visual stimuli. Activity levels in several regions of interest also co-varied with strength of audiovisual recalibration, with greater activity accompanying larger recalibration shifts. Similar activation patterns appeared for lexical retuning, but here no significant regions of interest were identified. Audiovisual and lexical forms of perceptual learning thus induce largely similar brain response patterns. However, audiovisual recalibration involves additional visual cortex contributions, suggesting that previously acquired visual information (on lip movements) is retrieved and deployed to disambiguate auditory perception.

Introduction

Speech perception is influenced by information other than the acoustic signal itself, such as seeing concurrent lip-movements, or the listener's lexical knowledge. These contextual cues not only support speech comprehension, but can also create categorically different and novel percepts; consider, for example, the McGurk effect, whereby an auditory syllable (such as /ba/) paired with video of a speaker pronouncing an incongruent syllable (such as /ga/) leads to a perceived new syllable (often /da/) (McGurk & MacDonald, 1976). Similarly, when presented with a word containing an unclear syllable (such as a /d/-/t/ blend instead of /d/ in *desk*), listeners are more likely to report hearing a word rather than a non-word (*desk* rather than *tesk*) (Ganong, 1980). Audiovisual lip-reading cues and lexical knowledge can guide and disrupt perception, but can also alter the categorical boundaries of presented phonemes.

Through audiovisual recalibration, listeners presented with video of a speaker pronouncing a syllable, such as /aba/, paired with an ambiguous auditory stimulus (an /aba/-/ada/ mixture) are, after sufficient exposure to the combination, likely to perceive the auditory blend without visual cues as /aba/ (Bertelson, Vroomen, & De Gelder, 2003). Similarly, in lexically-guided perceptual retuning, listeners presented with an ambiguous phoneme embedded within words (such as an /s/-/f/ blend in place of /s/ in words such as *horse*), are later likely to identify the /s/-/f/ phoneme blend when it is heard without lexical context as /s/ (Norris, McQueen, & Cutler, 2003).

Both of these approaches allow a glimpse into how speech sound categories can be shifted using contextual cues in addition to the acoustic signal. As audiovisual recalibration can operate through an additional sensory modality (vision), unlike lexical retuning which relies on word recognition within the same sensory channel (audition), the two forms of perceptual learning tend to differ in how they can be induced. In audiovisual processing, the visual cues such as lip movements are available earlier to the listener (Jesse & Massaro, 2010) and thus strong perceptual shifts can be observed after only a few exposure items, but these effects also diminish quickly (Vroomen et al., 2004), while lexical cues can lead to

longer-lasting, more robust effects, but following long exposures towards one particular phoneme (Eisner & McQueen, 2006). When lexical and audiovisual effects are compared under the same exposure and testing conditions, with short exposures (i.e. 8 biasing items) in alternation with short categorization tests on ambiguous items, both adaptation effects occur, with audiovisual cues generating larger perceptual shifts than lexical cues (van Linden & Vroomen, 2007; Ullas, Formisano, Eisner, & Cutler, 2020a); the behavioral effects are however not additive (Ullas, Formisano, Eisner, & Cutler, 2020b).

The application of neuroimaging techniques such as functional MRI (fMRI) has indicated some of the brain regions involved in category retuning. In general, speech perception employs a network of primarily left-lateralized regions in and around the temporal cortex, particularly within Heschl's gyrus (HG) and planum temporale (PT) (Binder, 2000; Zatorre et al., 1992; Zatorre, Belin, & Penhune, 2002). Phonetic perception has been linked to activation in HG and PT (Jäncke et al., 2002) as well as the superior temporal gyrus (STG) and sulcus (STS) (Buchsbaum, Hickok, & Humphries, 2001; Formisano, De Martino, Bonte, & Goebel, 2008); these areas are also responsible for encoding low-level acoustic-phonetic features and phonemes (Chang et al., 2011; Leonard & Chang, 2014; Mesgarani et al., 2008; 2014; Rutten et al., 2019). STG and STS are also implicated in distinguishing intelligible speech from distorted speech (Davis & Johnsrude, 2003), recognizing consonant-vowel syllables (Liebenthal et al., 2005) and identifying phonemic sounds (Liebenthal & Bernstein, 2017). Dual streams of processing may be responsible for acoustic feature processing and gestural motor processing, separated by an anterior-ventral and posterior-dorsal pathway, respectively (Hickok & Poeppel, 2004; Scott & Johnsrude, 2003), although phoneme processing can be bilateral and shared between networks in both the left and right hemispheres (Formisano et al. 2008; Hickok & Poeppel, 2004).

Speech perception extends into frontal and parietal regions as well (Rauschecker & Scott, 2009). Pre-motor, motor, and parieto-temporal regions are pertinent for representing articulatory gestures and sensorimotor functions (Hickok & Poeppel, 2007), while the left inferior frontal gyrus (LIFG) is notably linked to speech comprehension and unifying various levels of linguistic

information, including phonemes, syllables, and semantics (Hagoort, 2005; Poldrack, Wagner, Prull, Desmond, & Glover, 1999; Sharp, Scott, Cutler, & Wise, 2005).

When lip movement cues accompany speech, creating audiovisual speech, a similar pattern of activity in the brain can be found across frontal, parietal, and temporal regions (Bernstein & Liebenthal, 2014; Dick, Solodkin, & Small, 2010), with the addition of occipito-temporal contributions (Skipper et al., 2007). Activity in STG and IFG has been observed while listeners experience the McGurk effect (Jones & Callan, 2003), and phoneme boundary shifts resulting from the McGurk effect have been located within STG (Lüttke et al., 2016). STS may also facilitate perception of noisy audiovisual speech (Beauchamp, 2005) and contextual influences from surrounding sentences on phoneme processing can be exerted by STG and left MTG (Guediche, Salvata, & Blumstein, 2013). Kilian-Hütten, Vroomen, & Formisano (2011) specifically investigated audiovisual recalibration using fMRI. These authors found that exposure to the audiovisual pairings of ambiguous syllables with videos of lip-movements elicited activity in STG, as well as in the inferior parietal lobe (IPL), inferior frontal sulcus (IFS), and posterior MTG. Interestingly, activity in response to exposure of adaptor sounds in the same regions predicted activity during test blocks, when ambiguous auditory stimuli were presented in isolation. Furthermore, Kilian-Hütten, Valente, Vroomen, & Formisano (2011) applied multivariate pattern analysis (MVPA) to show that unique patterns of auditory cortex activity reflected the syllable percept (/aba/ and /ada/) for the same acoustic stimulus presented during the test phase.

Similarly, the lexical or Ganong effect has been associated with activity across left and right STG as well as frontal and parietal regions (Myers & Blumstein, 2008). Lexically-driven perceptual learning appears to initially depend on frontal and middle temporal regions, followed by later activity in left superior temporal areas when listeners perceive tokens along a continuum of /g/-/k/ whose shift is mediated by exposure to lexical stimuli containing an ambiguous /g/-k/ (Myers & Mesite, 2014).

Although studies on lexical and audiovisual recalibration have thus indicated involvement of similar brain areas, prior studies did not directly compare

the neural underpinnings of the two phenomena. The recalibration or perceptual retuning paradigm allows for the use of the same stimuli during test blocks with either lexical or audiovisual exposure. The ambiguous phoneme blends, to be perceived differently depending on the prior exposure block, can consist of either edited words or videos. The exposure time can also be matched; while lexical retuning studies typically use longer exposure phases to induce a bias, such retuning can take place in shorter timespans and can be observed in shorter test blocks, similar to the typical audiovisual exposure, as well (van Linden & Vroomen, 2007; Ullas, et al., 2020a,b).

In this study, lexical and audiovisual recalibration were compared using fMRI, to determine the similarity between the underlying brain regions involved in the two processes using similar testing procedures. As noted above, the existing behavioral studies of audiovisual recalibration and lexical retuning have tended to differ in the amount of exposure time used to induce effects, but they have also differed in the constancy of the bias. Thus the long exposure phases in lexical retuning have usually served to induce a bias towards only a single phoneme; in contrast, audiovisual recalibration studies have not only used shorter blocks (e.g., eight stimuli) but have also induced a changing phoneme bias throughout the experiment (e.g., Eisner & McQueen, 2006; Vroomen et al., 2004). The present study maintained consistency between the two procedures by using exposure blocks of the same length for both types of stimuli, and also allowing the phoneme bias to vary for both. Ambiguous phonemes were presented in identical test blocks and participants indicated their percept to assess recalibration effects in the same way for each exposure type. This approach of alternating exposure (containing either audiovisual or lexical stimuli, with changing phoneme biases) and test blocks has been shown to be effective in producing both audiovisual recalibration and lexical retuning (see Ullas et al., 2020a for more details regarding the behavioral outcomes of this approach). By utilizing this procedure, the study aimed to identify the neural commonalities between lexical and audiovisual recalibration under similar experimental constraints, as well as potential unique contributions from multimodal or visual regions for audiovisual recalibration, in contrast to activity within areas of the language network for lexical retuning.

As these two processes likely involve similar cortical areas, we made use of ultra-high field MRI at 7 Tesla which provided increased sensitivity in detecting possible differences. While audiovisual and lexical recalibration have been shown to involve highly similar areas across the temporal cortex as well as parietal, motor, and insular areas, audiovisual recalibration seems in previous studies to have been influenced by visual cortex activity as well. For both lexical retuning and audiovisual recalibration, we investigated whether activity within regions of interest (in temporal, occipital, inferior-parietal, and insular regions), defined by activity during exposure, could distinguish test blocks with high and low adaptation effects, with higher activation associated with higher behavioral scores.

Methods

Participants

Twelve participants (nine female, three male) were recruited from Maastricht University to take part in the study (data from one participant was not analyzed due to excessive motion leading to poor quality MRI data). All participants had normal or corrected-to-normal vision and normal hearing. Participant age range was 21.7 to 27.3 years (mean age = 24.5). Participants gave written informed consent to be scanned and to have their data shared.

Stimuli

The stimulus sets contained a combination of exposure and test stimuli, where exposure stimuli were designed to induce a bias towards a particular phoneme using either lexical or audiovisual (lip-reading) information, while test stimuli were ambiguous phonemes presented without context, to which listeners could report what phoneme they heard. If recalibration/retuning were successful, responses to test stimuli would be in line with the phoneme bias contained in the prior exposure block (i.e. more perceived /p/ after /p/-biased exposure, etc.). Exposure stimuli consisted of audio recordings of words and audio-video recordings of pseudowords, to measure lexical retuning and audiovisual recalibration, respectively. Pseudo-words were used to isolate the influence of audiovisual cues without any additional

confounds, while also retaining the speech-like structure. All stimuli had the clear portions of the critical phoneme removed (either /op/ or /ot/) and replaced with an ambiguous /op/-/ot/ blend, which was individually chosen from a ten-step /op/-/ot/ continuum.

For lexical stimuli, sixteen Dutch words with eight /op/ and eight /ot/ endings were chosen. Most words did not contain any acoustically similar phonemes (i.e. /b/ or /d/) so as to limit retuning effects to the critical phonemes only. Importantly, words were chosen such that only one of the two critical phonemes in the final position could form a word (i.e. *siroop* is a word but *siroot* is not). There were four two-syllable words, three three-syllable words, and one monosyllabic word ending in /op/ and /ot/. All stimuli are listed in Table 1.

For audiovisual stimuli, 16 pseudo-words were created using WinWordGen (Duyck, Desmet, Verbeke, & Brysbaert, 2004). Pseudo-words were matched with words for number of syllables, and lip-movements of the speaker indicated /op/ or /ot/ endings, with eight of each.

/op/ words:		/ot/ words:	
Hoop	[hoʊp]	Vloot	[vloʊt]
Aanloop	[ˈaːnloʊp]	Afsloot	[ˈɑfsloʊt]
Afkoop	[ˈɑfkoʊp]	Vennoot	[vɛˈnoʊt]
Siroop	[sɪˈʀoʊp]	Vergroot	[vəʀˈɣʀoʊt]
Wanhoop	[ˈʋɑnhoʊp]	Walnoot	[ˈʋaːlnoʊt]
Geweerloop	[ɣəˈʋeːrˌloʊp]	Hazelnoot	[ˈɦɑzəlnoʊt]
Horoscoop	[ɦɔʀɔˈscoʊp]	Levensgroot	[ˈlɛvənsɣʀoʊt]
Kussensloop	[ˈkʏsənsloʊp]	Middenmoot	[ˈmɪdənmoʊt]
/op/ pseudowords:		**/ot/ pseudowords:**	
Smoop	[smoʊp]	Vroot	[vʀoʊt]
Aaroop	[ˈaːʀoʊp]	Faloot	[fɑˈloʊt]
Miloop	[ˈmɪloʊp]	Geroot	[ɣəˈʀoʊt]
Onsoop	[ˈɔnsoʊp]	Mevoot	[məˈvoʊt]
Weloop	[ʋəˈloʊp]	Neuloot	[ˈnøːloʊt]
Acenkoop	[ˈɑsəŋkoʊp]	Frieseloot	[ˈfʀisəloʊt]
Lakeroop	[ˈlɑkəʀoʊp]	Leuveroot	[ˈløːvəʀoʊt]
Senkenloop	[ˈsɛŋkənloʊp]	Sanekoot	[ˈsɑnəkoʊt]

Table 1. Stimuli used in the study, with corresponding IPA transcriptions.

All stimuli were recorded by a female native Dutch speaker in a sound-attenuated booth. Words and pseudo-words were all recorded with both /op/ and

/ot/ endings. In addition, *soop* and *soot* (not words in Dutch) were recorded to create an /op/-/ot/ continuum. Video recordings were centered around the speaker's mouth to highlight lip movements during audiovisual exposure.

A continuum of /op/ to /ot/ was created, using the *soop* and *soot* recordings, with the speech editing program Praat (Boersma & Heuven, 2001). The final portions of /op/ and /ot/ were each extracted, equated in duration at 44kHz sampling frequency and original pitch contours were replaced with the average (at about 230Hz), similar to previous morphing procedures (Mitterer, Scharenborg, & McQueen, 2013; van der Zande, Jesse, & Cutler, 2014). Consonant bursts and vowel durations of the /op/ and /ot/ tokens were scaled to the same peak amplitude and equated in duration (to 50ms for the vowel) and then blended together in 10% increments for each step of the continuum. The morphed /op/-/ot/ blends were spliced back onto the /s/ token of *soop*/*soot* for the pre-test and test block stimuli. Lexical and audiovisual exposure stimuli were created by splicing these blends at the zero crossing closest to the last 50ms of the vowel, to reduce potential effects of co-articulatory cues from the preceding vowel. For audiovisual stimuli, the edited pseudo-words replaced the audio of the original video recordings, so that the lip-movements of the final phoneme/p/ or /t/ were aligned with the ambiguous auditory phoneme. Multiple stimulus sets were created to be able to present listeners with the stimuli containing the phoneme blend perceived to be most ambiguous, on an individual basis.

Behavioral procedure

During each functional run of the MRI scanning session participants performed a categorization task on individually selected phonetically ambiguous blends. Prior to the start of the experiment, all participants underwent a pre-test to determine the sound along the /op/-/ot/ continuum they perceived to be most ambiguous, and to select the most appropriate stimulus set containing this token. The pre-test was conducted while participants were already placed in the scanner and using the MRI-compatible earphones, so that participants could become accustomed to the MR environment, sound presentation and stimuli as closely as possible to the actual

scanning session. Participants heard each sound on the continuum for a minimum of six times, with sounds at the middle of the continuum presented more often (six times for steps 1, 2, 9, and 10; eight times for steps 3 and 8; 12 times for steps 4, 5, 6, and 7). For each sound, participants responded with a button press to report whether they heard /op/ or /ot/.

The experimental design was adapted from a similar previous study by van Linden & Vroomen (2007). Stimuli were presented using Presentation software (version: 18.2; NeuroBehavioral Systems, Berkeley, CA). Lexical retuning and audiovisual recalibration were induced in a blocked, counterbalanced design. Each run consisted of eight exposure-test rounds, four rounds of inducing and testing audiovisual recalibration and four rounds of lexical recalibration. In each run, four blocks of audiovisual recalibration were followed by four blocks of lexical recalibration or vice versa. Half of the exposure blocks were biased towards /p/ and the other half towards /t/, so that each run contained two audiovisual-/p/ blocks, two audiovisual /t/-blocks, two lexical /p/-blocks and two lexical /t/-blocks. The phoneme bias of the exposure block alternated every two blocks. Although this procedure can successfully result in both audiovisual and lexical retuning effects, audiovisual cues, compared to lexical, can lead to larger effects (Ullas et al., 2020a).

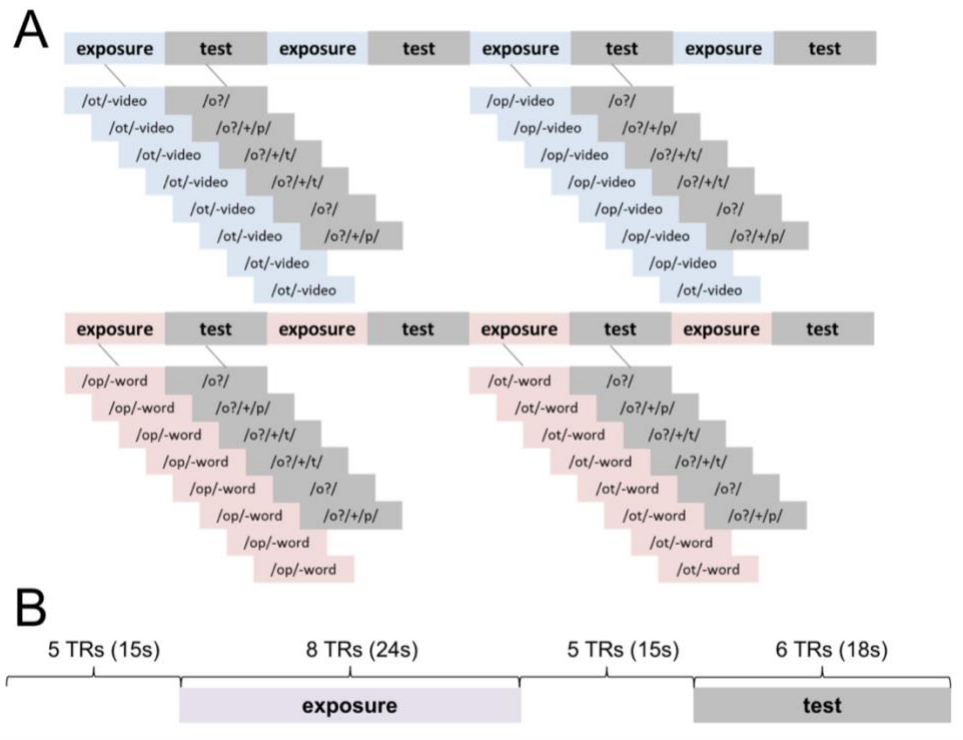


Figure 1. Sample scheme of a run (A). Half of the exposure blocks contained audiovisual stimuli, with half of those containing a bias towards /op/ or /ot/, and the same for the lexical blocks. The same test block followed every exposure block, with the most ambiguous token from the continuum selected from the pre-test and its two neighbors, each presented each twice. Participants were prompted to indicate by button press after every test item whether they heard /op/ or /ot/. Timings of exposure and test blocks are shown in (B); 15 seconds gaps, or 5 TRs (repetition time), were given between exposure and test blocks. Exposure and test items were presented within the silent gap of each TR.

In an exposure block, eight stimuli were presented with either /p/ or /t/-final bias, indicated by the lip-movements of the speaker in the audiovisual version, or by the phoneme the word would typically end in for lexical blocks. Four unique items were each presented twice without repetition of the same items. Following each exposure block was a test block, containing six stimuli reflecting the most ambiguous token from the /op/-/ot/ continuum and its two neighbors, each presented twice and in random order. Participants were instructed to respond during test blocks for each stimulus with a button press on a button-box as soon as the stimuli ended, signaling whether they heard /op/ or /ot/.

MRI data acquisition

Subjects were scanned in a Siemens 7 Tesla MRI scanner (Siemens Medical Systems, Erlangen, Germany) with a head coil (Nova Medical) at the Maastricht Brain Imaging Center (Maastricht, the Netherlands). Stimuli were presented binaurally through Sensimetrics MR-compatible earphones (Sensimetrics S14, Sensimetrics Corporation, Malden, MA) and played at a comfortable listening volume during silent gaps introduced within image acquisition (see below). Anatomical scans were acquired using a T1-weighted MPRAGE sequence at 0.6mm resolution, as well as a proton density image for inhomogeneity correction (TE = 2.52ms, TR = 3100ms, 192 slices). Functional scans were obtained using gradient echo (GE) sequence with Multiband 3 and GRAPPA 3 acceleration factor at 1.2mm resolution isotropic. 81 slices were collected per volume, with a 3000-ms TR (silent gap for sound presentation: 1500ms, acquisition time [TA] = 1500ms, echo time [TE] = 19ms, Field-of-View [FoV] = 229x229mm), and a total of 200 volumes per run. Five 10-minute runs were completed per participant. Two additional five-volume runs with opposite phase encoding directions (anterior-posterior and posterior-anterior, AP-PA) were collected for EPI distortion correction.

MRI data preprocessing

MRI and fMRI data were preprocessed using BrainVoyager QX v2.8 (BrainInnovation, Maastricht, the Netherlands). Anatomical T1 images were scaled using a proton density image to remove distortions. All images were transformed into Talairach space (Talairach & Tournoux, 1988) and interpolated to create 0.5mm anatomical and 1mm functional images. Motion correction and slice time correction was performed on all functional runs. To correct for EPI distortions, the data was corrected using the COPE plugin in BrainVoyager (version 0.5l) and the 5-volume AP-PA runs. Additional preprocessing steps included spatial smoothing (8mm FWHM) as well as temporal high-pass filtering (11 cycles per run) and linear trend removal. Gray-matter and white-matter segmentations were used for surface creation and functional data was projected onto vertices of the resulting cortical sheet.

MRI data analysis

Functional data were analyzed using a random-effects general linear model (GLM) including all runs of all participants with separate subject predictors, by convolving the time course of each condition with a hemodynamic response function. Here, predictors reflected six experimental conditions, with audiovisual and lexical exposure, high and low audiovisual test, and high and low lexical test, as well as a predictor for a baseline of neural activity in each run. Test blocks were defined as high or low based on behavioral performance, but the median number of correct responses (in the same direction as the bias of the prior exposure block, i.e. /p/ responses after a /p/-biased block) differed between lexical and audiovisual test blocks. For audiovisual recalibration (median correct = 4, range =1), if the participant responded with four or more correct responses then this was defined as a high recalibration test block, whereas blocks with fewer than four correct responses were defined as low recalibration test blocks. For lexical retuning (median correct = 3, range = 1), behavioral performance overall indicated a lower median of performance, so three or more correct responses were categorized as high test blocks, and fewer than three as low test blocks.

In addition to vertex-wise analyses, we conducted a region of interest (ROI) analysis to examine whether average activity within specific regions could distinguish high versus low recalibration test blocks. ROIs were defined based on individual fixed-effects GLMs using the activity during exposure phases. This produced five regions per participant in auditory cortex, parietal, insula, motor, and visual cortex (for audiovisual only) in both hemispheres. A contrast between high and low recalibration during the respective test blocks (i.e., audiovisual high versus low recalibration in regions defined by audiovisual exposure) was conducted for each ROI. Paired t-tests were performed on individual beta estimates reflecting activity during high and low recalibration test blocks within these ROIs.

Results

Behavioral

Pre-test responses on the 10-step continuum ranging from /op/ to /ot/ revealed that the sixth step was perceived to be most ambiguous on average.

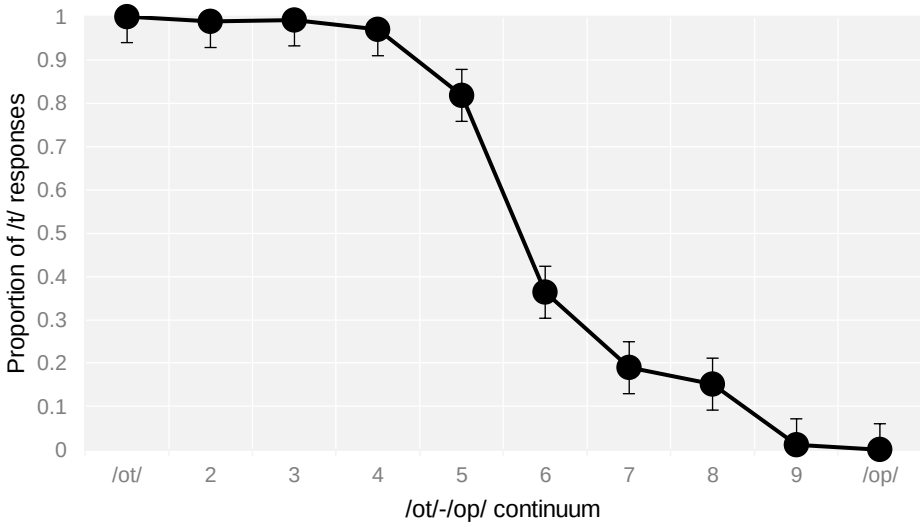


Figure 2. Pre-test responses. Responses to each of the 10 steps of the /op/-/ot/ continuum averaged across participants, with error bars indicating standard error.

Responses during test blocks were entered into a generalized linear mixed model with a logistic link using the *lmer* package in R (version 3.4.1). The factors *phoneme bias* during the exposure block, the type of exposure stimuli (lexical or audiovisual, as *condition*), and the three test *sounds* presented during the test blocks were entered as fixed effects into the model, and each individual *subject* was included as a random effect. Interactions were only modeled between the fixed effects variables. All variables were coded to be centered around 0, while responses during the test blocks were coded as 0 for /p/ and 1 for /t/. For model selection, the fitting was first performed for a full model including all possible main effects and interactions and followed by fitting of sparser models by iteratively removing slopes of random effects until the model converged and all fixed effects correlations were sufficiently low (less than 0.2). Results are shown in Table 2.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.24666	0.09079	-2.717	0.00659	**
Phoneme bias	0.46548	0.09633	4.832	1.35E-06	***
Condition	-0.28206	0.11301	-2.496	0.01257	*
Sound	0.52104	0.20483	2.544	0.01097	*
Phoneme bias*Condition	0.41149	0.18046	2.28	0.0226	*
Phoneme bias*Sound	-0.08412	0.11301	-0.744	0.45666	
Condition*Sound	-0.04486	0.1131	-0.397	0.69161	
Phoneme bias*Condition*Sound	0.10153	0.22575	0.45	0.6529	

*Significance: p<0.0001***; p<0.001**; p<0.01**

Table 2. Model results. Model: Response ~ Phoneme bias * Condition * Sound + (1 + Phoneme bias * Condition + Sound || Subject)

Model results showed a significant intercept, indicating a general tendency to respond with /p/ across all blocks, regardless of other factors. Main effects of *phoneme bias*, *sound*, and *condition,* were found to be significant. *Phoneme bias* was highly significant (p<0.0001), where more /t/ responses were found after /t/-biased exposure blocks than for /p/-biased exposure blocks, indicating successful recalibration with effects in the expected direction. *Sound* was also found to be significant, where more /t/-responses were observed for the more /t/-sounding test stimuli. The main effect of *condition* (p<0.001) indicated that subjects showed a stronger response bias towards /t/ across all lexical test blocks than across audiovisual test blocks. Pairwise contrasts were performed for *phoneme bias* and *condition,* and the difference in amounts of /t/-responses between /t/- and /p/-biased blocks was larger in the audiovisual condition (p<0.0001) compared to the lexical condition, where the difference was smaller (p<0.05). Behavioral results are displayed in Figure 3.

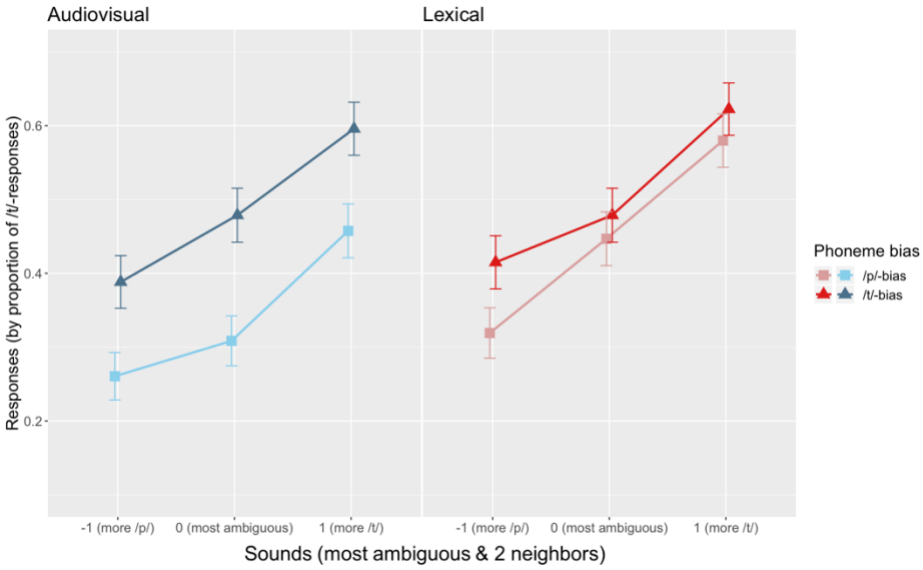


Figure 3. Behavioral results split by type of exposure in preceding block (lexical & audiovisual), across the three test sounds, and error bars for standard error.

FMRI results

GLM results

Group GLM results were projected onto a group-averaged brain, created using cortex-based alignment (Goebel, Esposito, & Formisano, 2006). First, contrasts between audiovisual and lexical exposure blocks versus baseline were performed (Figure 4A & 4C). In addition, contrasts between test blocks following audiovisual or lexical exposure, compared to baseline, were conducted (Figure 4B & 4D). To identify areas of overlap of conditions, conjunction maps between audiovisual and lexical exposure, and between audiovisual and lexical test were also created (Figure 5). All maps were corrected for multiple comparisons by cluster-size threshold (p_{corr}=.05), with an initial vertex-wise threshold of p=0.01. Cluster-size threshold correction was performed with Monte Carlo simulations to estimate the false positive rates at the cluster level (Goebel et al., 2006).

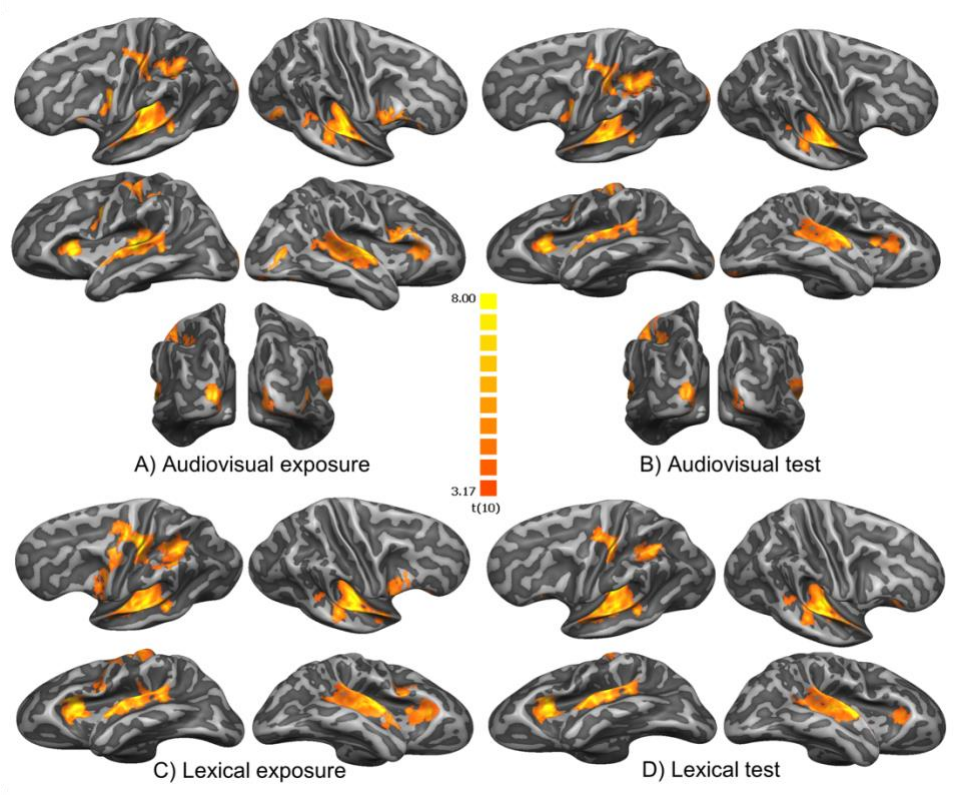


Figure 4. Audiovisual exposure (A), audiovisual test (B), lexical exposure (C), and lexical test (D) blocks versus baseline, with $t(10)>3.17$, $p<0.01$.

During audiovisual exposure blocks, significant bilateral engagement was observed in the temporal cortex, in Heschl's gyrus, PT and STG/STS, and in the occipital cortex between V1 and V2, as well as in IFG, insula, IPL, and postcentral gyrus in the left hemisphere and in an occipito-temporal cluster in the right hemisphere (Figure 4A). During lexical exposure blocks, bilateral activation of Heschl's gyrus, STG/STS, and insula was found, while postcentral gyrus/central sulcus, PP, PT, and IPL were also active in the left hemisphere (Figure 4B). Similarly, during test blocks following audiovisual exposure, significant activation was observed bilaterally in Heschl's gyrus/sulcus, PP, and STG/STS, in insula and between V1 and V2 as well. IPL and postcentral gyrus/central sulcus were also activated in the left hemisphere (Figure 4B). For test blocks after lexical exposure, significant activation was found across bilateral Heschl's gyrus, STG, PT, and insula, as well as postcentral gyrus/central sulcus, IPL, and PP in the left hemisphere

(Figure 4D). Activation during both exposure types (Figure 5A) and both tests (Figure 5B) were observed consistently in many of the same areas. Table 3 contains a list of all active regions & their respective coordinates (in Talairach space).

	Peak vertex			
Left hemisphere regions	X	Y	Z	Number of vertices
Temporal (HG, PT, PP, STG/STS)	-46	-25	6	6340
Frontal (IFG)	-45	3	22	2325
Insula	-27	17	7	1083
Motor (pre/postcentral gyrus, central sulcus)	-33	-24	44	2258
Occipital (V1/V2)	-12	-90	2	920
Parietal (IPL)	-32	-44	35	2221
Right hemisphere regions				
Temporal (HG, PT, PP, STG/STS)	54	-18	9	5128
Frontal (IFG)	45	5	16	742
Insula	30	24	10	972
Occipital (V1/V2)	10	-85	13	920
Occipito-temporal (BA19/V3)	39	-67	7	319

Table 3. List of active regions during exposure and test (as shown in Figure 4). All active regions are listed by hemisphere, with average Talairach coordinates of the peak vertex, and the average number of contiguous vertices per region, across participants.

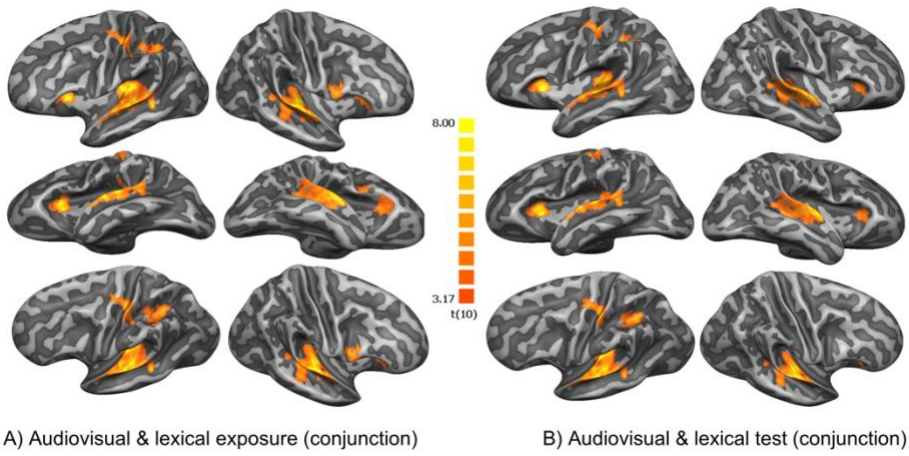
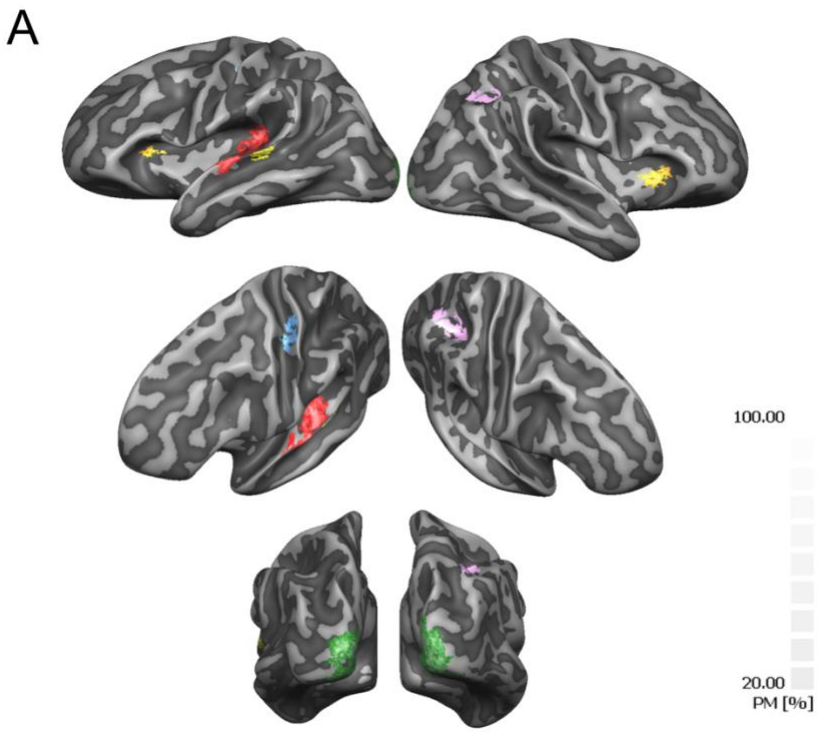


Figure 5. Conjunction maps between audiovisual and lexical exposure (A), and audiovisual and lexical test (B), with $t(10)>3.17$, $p<0.01$.

ROI analysis

For the analysis of ROIs (Figure 6A), defined based on activity during exposure blocks, significant differences between high and low recalibration test blocks were found for audiovisual recalibration but not for lexical retuning. As described in the Methods, test blocks were split into high and low based on the median number of correct responses per condition, which on average, resulted in 8.061 audiovisual low blocks (SD=2.833) and 9.129 lexical low blocks (SD=2.927), as well as 11.939 audiovisual high blocks (SD=2.561) and 10.871 lexical high blocks (SD=2.771) per participant. In addition, the positioning of high blocks was calculated to see whether high recalibration blocks may have been in positions where the phoneme bias of the previous exposure block could have had any effect on the recalibration, as the phoneme bias changed every two blocks. For example, if a /p/-biased block was followed by another /p/-biased block, we verified whether the second /p/-block may have potentially led to higher recalibration due to build-up, and if all of the high blocks were confounded by this. Of the two possible positions (the first being a change in phoneme bias versus the second being the same phoneme bias as the previous exposure), 67.78% of the first position blocks were high blocks and 70% of the second position blocks were high blocks for the audiovisual condition (p=0.344, paired *t*-test, two-tailed). For the lexical condition, 45.56% of the first position blocks and 51.11% of the second position blocks were categorized as high blocks (p=0.179, paired *t*-test, two-tailed). We concluded that there was no significant evidence that high recalibration blocks were confounded by the order of the phoneme biases in the exposures.

In ROIs defined by audiovisual exposure, temporal, insular, motor (central sulcus) regions, and STG in the left hemisphere showed a significant difference between high versus low test blocks, while insular and parietal clusters showed the same difference in the right hemisphere (Figure 6B). The contrast was also significant for both the left and right occipital ROIs.


A
100.00
20.00
PM [%]

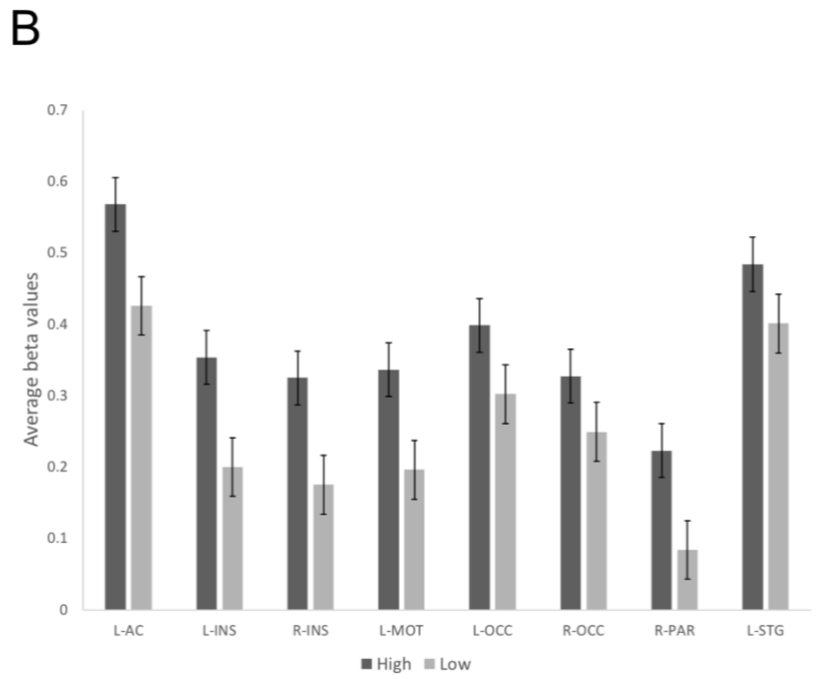


Figure 6. Significant ROIs for high versus low audiovisual recalibration. (A) Probabilistic maps (PM) are shown. Color shadings denote regions with an overlap of at least 3 participants showing a significant difference (p<0.01) between high and low audiovisual recalibration. (B) Average beta values by regions, for high and low audiovisual recalibration blocks. Significant differences between high and low blocks were found within temporal/auditory cortex (left), occipital/visual cortex (left & right), insula (left & right), motor (left), parietal (right) clusters, and STG (left). High recalibration referred to blocks where 4 or more correct responses, or responses that were in the same direction as the preceding exposure block (i.e. /p/ responses after /p/-biased exposure), whereas low recalibration included blocks with 0 to 3 correct responses. High versus low blocks per region were significant at p<0.05. Error bars indicate standard error.

Discussion

Phoneme category recalibration or retuning refers to a process that is an essential part of the celebrated robustness of human speech perception. Listeners can draw on information other than the acoustic signal – lip-movements, or lexical/semantic knowledge – to adjust boundaries between speech sound categories so that they fit the speech input they are currently hearing, which enables them to adapt to pronunciations they have perhaps never previously heard. Behavioral evidence (Ullas et al., 2020b) suggests that despite the apparent similarity, these two

adaptation processes may have distinct triggers (coping with noise in the case of audiovisual recalibration, coping with talker novelty in the case of lexical retuning), although both types of adaptation often occur conjointly in real-life. In the present study, fMRI data was collected as participants underwent both forms of phoneme category adjustments, using lexical and audiovisual cues respectively, in a counterbalanced, blocked design. The perceptual boundary between two phonemes, /p/ and /t/, was systematically shifted, using lexical and audiovisual cues, towards either /p/ or /t/. Note that the behavioral results had shown that this procedure resulted in significant effects in both conditions and towards both phonemes, although audiovisual recalibration effects were larger than lexical retuning, in line with previous findings as well (van Linden & Vroomen, 2007; Ullas, et al., 2020a).

The analysis of concurrent fMRI measurements showed similarities between audiovisual and lexical exposure blocks, particularly in the temporal cortex across bilateral HG, STG/STS, PT, as well as left IPL and right insula. HG and PT are most likely responsible for acoustic and rudimentary phonetic processing (Binder, 2000; Obleser & Eisner, 2009), while nearby STG and STS are likely to represent similar items such as syllables and phonemes (Jäncke et al., 2002; Mesgarani et al., 2008; Yi, Leonard, & Chang, 2019), although they may show overlap in their functions.

Outside of the lower-level perceptual areas, insula and IPL activity was also evoked during the audiovisual and lexical exposure blocks. The insula has been proposed to be a part of the articulatory network (Hickok and Poeppel, 2007). Oh, Duerden, & Pang (2014) suggest that the insula also oversees articulation, and other motor-like properties of speech, and is connected to other speech and language regions, including Broca's area. IPL activity may be related to processing audiovisual speech as well as words and pseudowords (Newman & Tweig, 2001; Ojanen et al., 2005) Some areas were uniquely engaged by audiovisual exposure, in the occipital cortex over V1 and V2, while lexical exposure was not associated with any unique brain areas. Naturally, the presentation of visual stimuli during the audiovisual blocks elicited activity within the visual/occipital cortex, unlike the lexical blocks where no visual stimuli were presented.

Similar patterns of activation were identified during test blocks following audiovisual and lexical exposure in the temporal cortex, again within HG, STG, and STS. As previously mentioned, these regions are responsible for representing phonemes, syllables, and low-level acoustic information. Activation in these early auditory regions has also been found to undergo top-down modulation by attention to task-relevant acoustic information, such as spectral or temporal features (Rutten et al., 2019). In addition to these functions, Myers and Mesite (2014) reported STG and MTG activity to be strongest for ambiguous items that had been perceptually shifted by exposure to lexical items. Kilian-Hütten, Vroomen, & Formisano (2011) similarly noted STG as well as IPL, insula, and IFS to be activated during audiovisual recalibration, and that IPL can coordinate higher-order constructive processes in perception. Regions in the parietal lobe may also be involved in detecting phonological changes, distinguishing words from pseudowords, and general linguistic comprehension (Binder et al., 1997; Newman & Tweig, 2001; Obleser & Eisner 2009). Similarly, the insula can assist in disambiguating degraded speech (Erb, Henry, Eisner, & Obleser, 2013). IPL and insula activation have been reported to underlie text-based recalibration as well (Bonte, Correia, Keetels, Vroomen, & Formisano, 2017). As IPL and insula lie outside of the core speech network, they may also be involved in less tangible functions, such as processing abstract linguistic information or multimodal integration (Dick et al., 2010; Guediche, Blumstein, Fiez, & Holt, 2014; Jones & Callan, 2003). The convergence of these regions in the present study, as well as the left-right asymmetry we observed in activation strength, consistently align with previous studies of speech perception and retuning/recalibration. Also, as expected from that prior work, audiovisual cues led to stronger effects than lexical cues.

Although additional activation was also elicited in postcentral gyrus and central sulcus for lexical and audiovisual test blocks, this most likely reflects activity related to the expected button presses. Therefore, it appears unlikely that the activity observed in these regions represents any functions beyond the button presses made during the test blocks, however, motor cortex activity may be reflective of gestural or articulatory movements triggered by speech sounds

(Hickok & Poeppel, 2007) and may ease the interpretation of ambiguous speech sounds (Guediche et al., 2014).

Both forms of perceptual learning showed a pattern of reactivation, where many of the same regions active during the exposure blocks were also active during the test blocks, despite the differences in stimuli and task between exposure and test blocks. This overlap was observed in namely HG, STG/STS, and left IPL for both audiovisual and lexical test blocks. Both exposure and test blocks evoked activity in the speech network as a result of the presentation of speech (and speech-like) sounds. Most notably however, the occipital cortex remained active during audiovisual test blocks, although no visual stimuli were presented and a sufficient amount of time was given between exposure and test blocks to allow the BOLD response to return to baseline. The sustained activation in visual cortex suggests that the visual information from the exposure blocks is salient enough to be retained during the subsequent test block, possibly as a form of mental imagery or within a short-term memory loop, as early visual areas are capable of contributing to visual mental imagery (Kosslyn, Ganis, & Thompson, 2001; Sparing et al., 2002). Associative learning may entail involuntary visual learning, or when an association is formed between two stimuli, and can take place within early visual areas such as V1 and V2 (Pearson, 2019). In the present study, listeners may thus have formed associations between the ambiguous phonemes and the preceding visual stimuli, with these associations being retrieved and deployed during the test blocks. Kilian-Hutten, Vroomen, & Formisano (2011) have also noted functional connectivity between occipital regions and left auditory cortex during audiovisual recalibration. Further, the strong activation of visual cortex during purely auditory test blocks suggests a functional role of visual cortex during audiovisual recalibration, and that the auditory cortex does not implement these perceptual shifts on its own.

An ROI analysis revealed a number of regions that were found to be modulated by audiovisual recalibration only, including clusters in left temporal, motor, insular regions, and in right insular and parietal clusters, as well as a larger region spanning V1 and V2. These regions showed significantly higher hemodynamic activity for test blocks where participants showed larger recalibration effects, and lower activity for weaker effects. The relative increase in

activity observed during high recalibration blocks points toward more efficient identification of the ambiguous sounds, facilitated by top-down contributions from these regions. A conjunction of both higher- and lower-order regions within and outside of the speech network appears capable of distinguishing high and low recalibration performance, which suggests that the process may not be unidirectional, requiring instead a combination of extraction of lower-level acoustic features plus recourse to higher-level semantic and cross-modal representations. The strength of neural activity in these regions seems to be associated with a larger category boundary shift in the same direction as the preceding exposure. Low recalibration blocks appear to be linked with lower levels of activation, however, the relationship between the two is unclear as the underlying cause could be due to a number of factors, such as a lack of attention paid during exposure, the combination of stimuli during exposure not effectively inducing a shift in perception, or fatigue with repeated testing.

The same analysis within the ROIs was not associated with any differences in lexical retuning, corresponding to neither high nor low performance in the test blocks. Participants' generally lower performance during lexical test blocks may have reduced the scope for a significant difference between high- and low-scoring lexical blocks in comparison to the audiovisual test blocks. This might then have translated into the lack of a neural difference as well. In contrast, behavioral audiovisual recalibration effects were larger than lexical, which could have led to higher activation overall compared to lexical test blocks, and thereby increased sensitivity to detecting differences between high and low recalibration within regions of interest. Nonetheless, lexical retuning was still elicited under the constraints of the task design (i.e., few exposure items and continuous boundary shifting) and evoked significant patterns of activation across regions known for acoustic-phonetic processing (HG, STG/STS) and higher-levels of cognitive engagement (IPL, insula).

Conclusion

The present study compared audiovisual recalibration and lexical retuning using high-field fMRI to investigate the underlying similarities and differences in their neural activity. A network of speech-related regions and other higher-order areas emerged as a result of the two forms of perceptual learning, while audiovisual recalibration specifically seems to evoke significant visual cortex input during the process, pointing towards a form of involuntary mental imagery, perhaps as a byproduct of associative learning taking place between the visual stimuli and the ambiguous phonemes. In addition, neural activity in several regions spread across the brain was found to be modulated in correspondence with the amount of audiovisual recalibration observed behaviorally. While lexical retuning did not display this pattern across the selected regions, remarkable overlap with audiovisual recalibration was found in temporal, parietal, and insular regions. Evidently, a number of both lower-level regions involved in acoustic-phonetic processing, as well as more complex semantic and cross-modal areas are involved in these perceptual adjustments. From within and extending beyond the speech network, the strength of the relationship formed between the exposure stimuli and the ambiguous phonemes may therefore be responsible for enabling the perceptual shifts. The precise timing and directionality of information processing remain to be investigated; however, our results suggest that not only do recalibration and retuning involve subtly different triggers, but the brain areas responsible for modulating them also involve multiple levels of perceptual organization.

References

Beauchamp, M. S. (2005). See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Current Opinion in Neurobiology*, *15*(2), 145–153. https://doi.org/10.1016/j.conb.2005.03.011

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in Neuroscience*, *8*, 386. https://doi.org/10.3389/fnins.2014.00386

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

Binder, J.R. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, *10*(5), 512–528. https://doi.org/10.1093/cercor/10.5.512

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience*, *17*(1), 353–362. https://doi.org/10.1523/jneurosci.17-01-00353.1997

Boersma, P., & Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International*, *5*(9/10), 341–347. https://doi.org/10.1097/AUD.0b013e31821473f7

Bonte, M., Correia, J. M., Keetels, M., Vroomen, J., & Formisano, E. (2017). Reading-induced shifts of perceptual speech representations in auditory cortex. *Scientific Reports*, *7*(1), 5134. https://doi.org/10.1038/s41598-017-05356-3

Buchsbaum, B. R., Hickok, G., & Humphries, C. (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cognitive Science*, *25*(5), 663–678. https://doi.org/10.1207/s15516709cog2505

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, M., & Knight, R. T. (2011). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience, 13*(11), 1428–1432. https://doi.org/10.1038/nn.2641

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language

comprehension. *The Journal of Neuroscience*, *23*(8), 3423–3431. https://doi.org/10.1523/jneurosci.23-08-03423.2003

Dick, A. S., Solodkin, A., & Small, S. L. (2010). Neural development of networks for audiovisual speech comprehension. *Brain and Language*, *114*(2), 101–114. https://doi.org/10.1016/j.bandl.2009.08.005

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: a tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, and Computers*, *36*(3), 488–499. https://doi.org/10.3758/BF03195595

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: stability over time. *The Journal of the Acoustical Society of America, 119*(4), 1950-1953.https://doi.org/10.1121/1.2178721

Erb, J., Henry, M., Eisner, F., & Obleser, J. (2013). The brain dynamics of rapid perceptual adaptation to adverse listening conditions. *Journal of Neuroscience, 33*(26), 10688-10697. https://doi.org/10.1523/JNEUROSCI.4596-12.2013

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science, 322*(5903), 970-973. https://doi.org/10.1126/science.1164318

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology. Human Perception and Performance*, *6*(1), 110–125. https://doi.org/10.1037/0096-1523.6.1.110

Goebel, R., Esposito, F., & Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with BrainVoyager QX: from single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Human Brain Mapping, 27*(5), 392-401. https://doi.org/10.1002/hbm.20249

Guediche, S., Blumstein, S. E., Fiez, J. A., & Holt, L. L. (2014). Speech perception under adverse conditions: insights from behavioral, computational, and neuroscience research. *Frontiers in Systems Neuroscience*, *7*, 126. https://doi.org/10.3389/fnsys.2013.00126

Guediche, S., Salvata, C., & Blumstein, S. E. (2013). Temporal cortex reflects effects

of sentence context on phonetic processing. *Journal of Cognitive Neuroscience, 25*(5), 706-718. https://doi.org/10.1162/jocn_a_00351

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences*, *9*(9), 416–423. https://doi.org/10.1016/j.tics.2005.07.004

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition, 92*(1-2), 67-99. https://doi.org/10.1016/j.cognition.2003.10.011

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews. Neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113

Jäncke, L., Wüstenberg, T., Scheich, H., & Heinze, H. J. (2002). Phonetic perception and the temporal cortex. *NeuroImage*, *15*(4), 733–746. https://doi.org/10.1006/nimg.2001.1027

Jesse, A., & Massaro, D.W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception, & Psychophysics, 72,* 209–225. https://doi.org/10.3758/APP.72.1.209

Jones, J. A., & Callan, D. E. (2003). Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *Neuroreport*, *14*(8), 1129–1133. https://doi.org/10.1097/00001756-200306110-00006

Kilian-Hütten, N., Valente, G., Vroomen, J., & Formisano, E. (2011). Auditory cortex encodes the perceptual interpretation of ambiguous sounds. *The Journal of Neuroscience*, *31*(5), 1715–1720. https://doi.org/10.1523/jneurosci.4572-10.2011

Kilian-Hütten, N., Vroomen, J., & Formisano, E. (2011). Brain activation during audiovisual exposure anticipates future perception of ambiguous speech. *NeuroImage*, *57*(4), 1601–1607. https://doi.org/10.1016/j.neuroimage.2011.05.043

Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2001). Neural foundations of imagery. *Nature Reviews Neuroscience, 2,* 635-642. https://doi.org/10.1038/35090055

Leonard, M. K., & Chang, E. F. (2014). Dynamic speech representations in the human temporal lobe. *Trends in Cognitive Sciences*, *18*(9), 472–479. https://doi.org/10.1016/j.tics.2014.05.001

Liebenthal, E., & Bernstein, L. E. (2017). Editorial: Neural mechanisms of perceptual categorization as precursors to speech perception. *Frontiers in Neuroscience, 11*, 69. https://doi.org/10.3389/fnins.2017.00069

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex, 15*(10), 1621-1631. https://doi.org/10.1093/cercor/bhi040

Lüttke, C. S., Ekman, M., Van Gerven, M. A. J., & De Lange, F. P. (2016). McGurk illusion recalibrates subsequent auditory perception. *Scientific Reports*, *6*, 32891. https://doi.org/10.1038/srep32891

McGurk, H., & MacDonald, M. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1010. https://doi.org/10.1126/science.1245994

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *The Journal of the Acoustical Society of America*, *123*(2), 899–909. https://doi.org/10.1121/1.2816572

Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition*, *129*(2), 356–361. https://doi.org/10.1016/j.cognition.2013.07.011

Myers, E. B., & Blumstein, S. E. (2008). The neural bases of the lexical effect: an fMRI investigation. *Cerebral Cortex, 18*(2), 278-288. https://doi.org/10.1093/cercor/bhm053

Myers, E. B., & Mesite, L. M. (2014). Neural systems underlying perceptual adjustment to non-standard speech tokens. *Journal of Memory and Language*, *76*, 80–93. https://doi.org/10.1016/j.jml.2014.06.007

Newman, S. D., & Tweig, D. (2001). Differences in auditory processing of words and pseudowords: an fMRI study. *Human Brain Mapping*, *14*(1), 39-47. https://doi.org/10.1002/hbm.1040

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

Obleser, J., & Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. *Trends in Cognitive Sciences*, *13*(1), 14–19. https://doi.org/10.1016/j.tics.2008.09.005

Oh, A., Duerden, E. G., & Pang, E. W. (2014). The role of the insula in speech and language processing. *Brain and Language, 135,* 96-103. https://doi.org/10.1016/j.bandl.2014.06.003

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T., & Sams, M. (2005). Processing of audiovisual speech in Broca's area. *NeuroImage*, *25*(2), 333–338. https://doi.org/10.1016/j.neuroimage.2004.12.001

Poldrack, R. A., Wagner, A. D., Prull, M. W., Desmond, J. E., Glover, G. H., & Gabrieli, J.D. (1999). Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. *Neuroimage*, *10*(1), 15–35. https://doi.org/10.1006/nimg.1999.0441

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718–724. https://doi.org/10.1038/nn.2331

Rutten, S., Santoro, R., Hervais-Adelman, A., Formisano, E., & Golestani, N. (2019). Cortical encoding of speech enhances task-relevant acoustic information. *Nature Human Behaviour, 3*, 974-987. https://doi.org/10.1038/s41562-019-0648-9

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, *26*(2), 100–107. https://doi.org/10.1016/S0166-2236(02)00037-1

Sharp, D. J., Scott, S. K., Cutler, A., & Wise, R. J. S. (2005). Lexical retrieval constrained by sound structure: the role of the left inferior frontal gyrus. *Brain and Language*, *92*(3), 309–319. https://doi.org/10.1016/j.bandl.2004.07.002

Skipper, J. I., Van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, *17*(10), 2387–2399. https://doi.org/10.1093/cercor/bhl147

Sparing, R., Mottaghy, F. M., Ganis, G., Thompson, W. L., Töpper, R., Kosslyn, S. M., & Pascual-Leone, A. (2002). Visual cortex excitability increases during visual mental imagery - A TMS study in healthy human subjects. *Brain Research, 938*(1-2), 92-97. https://doi.org/10.1016/S0006-8993(02)02478-2

Talairach, J., & Tournoux, P. (1988). Co-Planar Stereotaxic Atlas of the Human Brain. *Thieme, New York*.

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

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

Van der Zande, P., Jesse, A., & Cutler, A. (2014). Cross-speaker generalisation in two phoneme-level perceptual adaptation processes. *Journal of Phonetics*, *43*, 38–46. https://doi.org/10.1016/j.wocn.2014.01.003

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

Vroomen, J., van Linden, S., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Communication*, *44*(1-4), 55–61. https://doi.org/10.1016/j.specom.2004.03.009

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron*, *102*(6), 1096–1110. https://doi.org/10.1016/j.neuron.2019.04.023

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences*, *6*(1), 37–46. https://doi.org/10.1016/S1364-6613(00)01816-7

Zatorre, R. J., Evans, A. C., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science, 256*(5058), 846-849. https://doi.org/10.1126/science.1589767

5
Summary & general discussion

Each chapter of this dissertation has addressed an aspect of a perceptual strategy known as recalibration or retuning, a process through which listeners can learn to adapt to a speaker by attending to information other than the auditory signal itself. These sources can include the lip-movements of the speaker (also known as audiovisual cues) or the listener's own lexical knowledge, which can assist them in making assumptions as to what the speaker is most likely to be saying. Repeated experience with pairings between an ambiguous auditory signal and these contextual sources can shift boundaries between phoneme categories (Bertelson, Vroomen, & De Gelder, 2003; Norris, McQueen, & Cutler, 2003) and thereby allow the listener to understand a speaker with more ease (Sjerps & McQueen, 2010).

Summary

In Chapter 2, lexical retuning and audiovisual recalibration were compared with a novel paradigm where listeners switched between the two forms of perceptual learning within a single session. Switching did not lead to significant cost in learning effects, compared to groups that received only one type of cue. Audiovisual recalibration effects were stronger than lexical retuning, in a switching group and a single exposure group, but this was most likely due to the design of the study which contained short exposures in two possible acoustic directions, unlike most previous studies of lexical retuning (Cutler, Eisner, McQueen, & Norris, 2010). Nevertheless, listeners were able to show alternating forms of perceptual learning, indicating that both audiovisual recalibration and lexical retuning are flexible even under constrained conditions. The obtained results may reflect how listeners can switch between informative contextual sources depending on the needs of the listening situation.

In Chapter 3, lexical and audiovisual cues were combined to investigate whether and how the combination of cues would lead to perceptual shifts. The combined cues produced effects that were similar in magnitude to audiovisual recalibration effects, but were larger than lexical effects. Again, a constrained design was employed (with short and alternating exposure/test blocks), so lexical cues may have led to diminished effects with the atypical design, but lexical and

audiovisual cues also did not additively combine to induce perceptual boundary shifts. Rather, the combination of cues led to effects larger than lexical retuning alone and comparable to audiovisual recalibration. The pattern of results suggests that lexical and audiovisual cues do not operate together when inducing categorical shifts, and the two types of cues may be relied upon for different purposes.

Finally, in Chapter 4, lexical retuning and audiovisual recalibration were compared in an fMRI study, to pinpoint the neural correlates underlying the two processes, and to identify how much overlap they share. Once again, an alternating blocked design was used in order to have participants undergo both forms of perceptual learning with two phonemes within a short window of time. During exposure blocks, audiovisual and lexical cues elicited similar patterns of activity in the temporal cortex, across Heschl's gyrus (HG), planum temporale (PT), superior temporal gyrus (STG) and sulcus (STS). These regions are involved in acoustic and phonemic processing (HG/PT/STG) as well as higher-level syllabic and semantic information (STG/STS) (Buchsbaum, Hickok, & Humphries, 2001; Formisano, De Martino, Bonte, & Goebel, 2008; Jäncke, Wüstenberg, Scheich, & Heinze, 2002). Significant activation was also found in the inferior parietal lobule (IPL) and the insula, but audiovisual exposure blocks specifically led to activation in the occipital cortex, between V1 and V2. Similarly, during test blocks, when listeners were undergoing recalibration or retuning resulting from the preceding audiovisual or lexical cues, activation was observed in HG and STG/STS in the temporal cortex, as well as IPL and insula. During audiovisual test blocks, significant activity was still evoked in the occipital cortex (V1/V2), even while no visual stimuli were presented. In addition, a number of regions defined by activity during the exposure blocks showed distinct differences in the degree of activation between high and low recalibration (i.e. more or fewer responses in the same direction as the bias contained in the prior exposure block). These regions included temporal, occipital, insular, and motor clusters, but only showed the high-low distinction for audiovisual test blocks, while no regions were significantly distinguishable for lexical test blocks. Overall, results showed that the areas of the brain involved in lexical retuning and audiovisual recalibration overlap in many respects especially within the auditory cortex, but audiovisual recalibration seems to trigger a specific

reactivation of the occipital cortex, which suggests the involvement of mental imagery (i.e. re-activation of visual representations from short-term memory) during shifts (Pearson, 2019). A network of regions across the brain also appears responsible for effectively shifting the category boundary, involving both low-level acoustic/phonetic processing, and higher-level cross-modal and semantic processing.

Taken together, the outcomes of these studies have clarified some of the similarities and differences between lexical retuning and audiovisual recalibration. In Chapter 2, retuning and recalibration were both found to be flexible, as listeners proved capable of switching between them, but lexical retuning can be limited in a design where blocks rapidly alternate between exposure and test, and between two different phonemes. However, this difference in effect size may represent differences in the typical applications of the processes, where audiovisual cues may be more suitable for short-term, situation specific learning (a noisy environment) whereas lexical cues may be more applicable to long-term, speaker-specific learning (unfamiliar accent, unusual pronunciations). Chapter 3 identified how retuning and recalibration seem to differ and do not additively combine to enhance aftereffects. It appears that lexical and audiovisual cues operate across different networks, and that there are domain-specific aspects of the phoneme categories that they tap into, which may prevent the cues from being utilized simultaneously. In addition, listeners do not seem to benefit from the combination of cues if one cue type is sufficiently informative; for example, the audiovisual cue may have already indicated to the listener what the ambiguous phoneme was most likely to be, then the lexical cue may not have provided any additional guidance. If two possible phoneme candidates are visually identical (such as /b/ and /p/), then lexical information may be more useful, but if two phonemes are visually different (such as /p/ and /t/), then audiovisual cues may be more helpful. Listeners most likely utilize whichever cue is fastest and most reliable in the given situation. Chapter 4 delineated the neural activity underlying retuning and recalibration, and both processes engaged areas across the temporal cortex that are known to be involved in rudimentary acoustic processing, such as HG, STG/STS, and PT. Both retuning and recalibration also showed patterns of reactivation between exposure and test,

as many of the same areas activated by the exposure blocks, when listeners where presented with either the audiovisual or lexical stimuli, were also activated by the test blocks, when only ambiguous phonemes were presented in a categorization task. However, the observed neural activity also points to modality-specific contributions, as audiovisual recalibration recruits the visual cortex, while lexical retuning largely relies on the speech network both within auditory cortex, and in other related areas such as IPL and insula.

Discussion

This dissertation sought a cohesive explanation of the various forms of perceptual adaptation, but a number of questions still remain unanswered and must be taken into consideration in order to bridge the gap in understanding between the two processes. The three studies revealed some of the limitations in perceptual adaptation studies, so future studies may benefit by circumventing these drawbacks accordingly. Many of these potential restrictions involved the stimulus construction, the study design, and the confines of an fMRI study. However, the findings across the three studies also elucidated some of the processes involved in speech perception, and how theories of speech perception may or may not be equipped to explain what perceptual adaptation entails.

Stimulus construction & design

The three studies used largely similar approaches to measure perceptual shifts, with alternating blocks of exposure and test, containing only six or eight stimuli, and with the phoneme bias also changing throughout the experimental session. This design, derived from a previous study (van Linden & Vroomen, 2007), allowed us to compared retuning and recalibration under the same constraints, as well as efficiently testing two forms of perceptual learning in two directions within the same session (Figure 1).

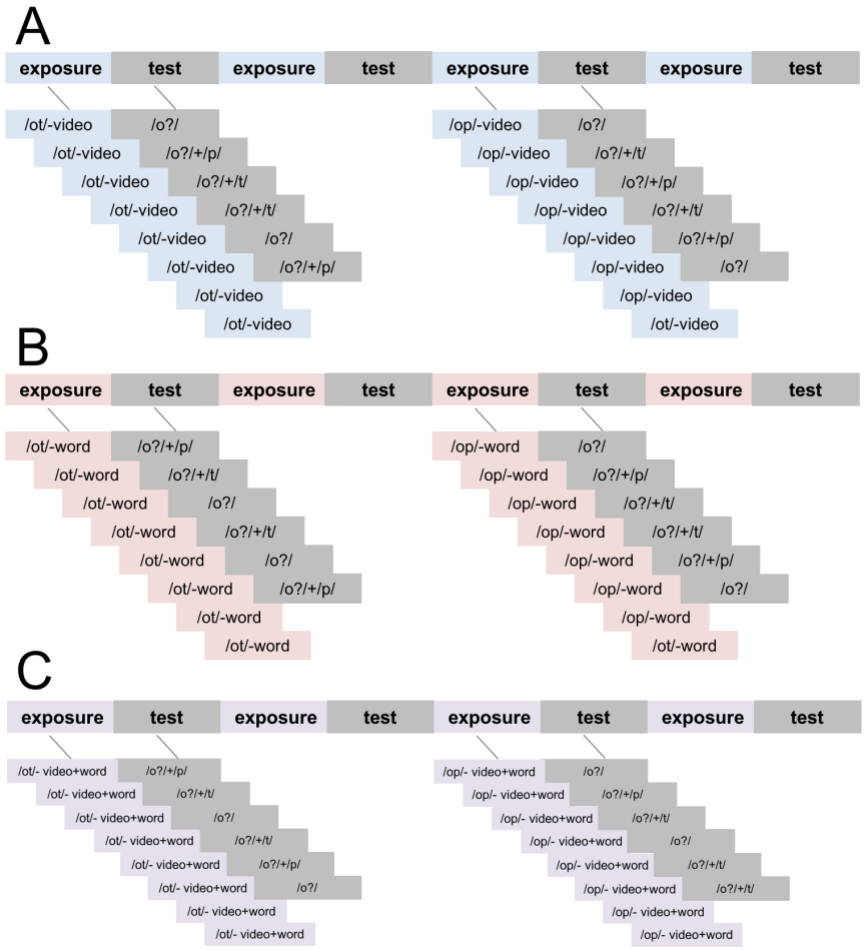


Figure 1. Experimental design used in all studies. Listeners received alternating blocks of exposure and test, where exposure blocks contained eight biasing stimuli towards one phoneme (/p/ or /t/), followed by test blocks that presented an ambiguous phoneme without context, and listeners were asked to respond with what they heard. In Chapter 2, listeners received A, B, or alternating A & B. In Chapter 3, listeners underwent A, B, or C. In Chapter 4, listeners were given A & B in every run (order counterbalanced).

However, both recalibration and retuning are sensitive to differences in the experimental designs, and while the approach we used presents an advantage in its flexibility, it also may have led to a reduction in lexical retuning effects, compared to previous studies. Therefore, it would be useful for future studies to explore how retuning and recalibration may fare under other designs, with longer or shorter lengths of exposure and test phases, changing the phoneme bias in exposure blocks

less or more often, or multiple sessions for more robust learning. Moreover, the designs used by previous studies with both lexical and audiovisual techniques, as well as by the present studies, are not truly representative of real-life listening scenarios, and future studies may also benefit by attempting to more closely emulate realistic listening, by embedding stimuli into sentences or conversations with multiple speakers, much like Eisner & McQueen (2006) who embedded critical exposure items into a story.

All three of the present studies also used the same phoneme contrast – a pair of voiceless plosive/stop consonants (/p/ and /t/). Other phoneme pairs should be investigated to see whether the patterns of effects remain the same or if they differ. Previous studies have also found differences in effects due to the phoneme pair, particularly the degree to which the garnered effects generalize to other speakers (Kraljic & Samuel, 2007; Mitchel, Gerfen, & Weiss, 2016; van der Zande, Jesse, & Cutler, 2014).

The three studies have also relied on ambiguous phonemes in order to measure recalibration and retuning effects, but the ways in which these stimuli are created can vary. Some previous studies have aimed to maximize the physical ambiguity of the stimuli by using the same ambiguous token for all participants based on a group average (studies by Vroomen and colleagues). In contrast, the present studies attempted to maximize perceptual ambiguity by creating multiple stimuli sets that were individually tailored per participant. Both approaches result in similar effects on average (Bruggeman & Cutler, 2019), but conversely, a few previous studies have found selective speech adaptation effects when using ambiguous lexical stimuli in a blocked design such as ours (Samuel, 2001; Samuel & Frost, 2016). Notably, Samuel & Frost (2016) found that participants who underwent exposure to ambiguous lexical stimuli that still contained co-articulatory cues showed selective speech adaptation effects. Selective speech adaptation effects are generally in the opposite direction to perceptual learning, where consistent exposure to a clear phoneme or syllable leads to a reduction in reports of hearing that phoneme (i.e. hearing /p/ repeatedly leads to a reduction in /p/ responses during a categorization task on ambiguous /p/-like sounds). In our studies, we aimed to eliminate any co-articulatory cues prior to the final critical phoneme in

the stimuli which could have contained enough phonemic information for listeners to pick up on. It is therefore important to take into account how the stimuli are constructed, so as to ensure that the participants undergo the desired effect.

fMRI limitations

Chapter 4 presented an fMRI study wherein we compared lexical retuning and audiovisual recalibration, and identified commonalities in neural activity between the two processes. While this illuminated many of the neural underpinnings of retuning and recalibration, a number of possibilities remain unexplored due to some limitations in both the experimental design and the requirements imposed by an fMRI study. In a pilot study, we attempted to use a slow-event related design during the test blocks in order to separate the neural response of each individual trial (i.e. each time the participant heard an ambiguous token and responded with what they heard). To implement this, each trial ranged from 15 to 18 seconds, to allow for enough time for the participant to hear the sound, and to separate the button press response from the perceptual event. Several previous studies have successfully used this design and applied multivariate pattern analysis (MVPA) to decode what participants were perceiving (i.e. decoding whether participants perceived /p/ or /t/ while being presented the same ambiguous token across the test trials; Bonte, Correia, Keetels, Vroomen, & Formisano, 2017; Kilian-Hütten et al., 2011; Lüttke, Ekman, Van Gerven, & De Lange, 2016). However, in the present study, this proved unsuccessful, and participants no longer showed perceptual learning effects.

This outcome may have several origins. For instance, it may have been due to the long trials which could have erased the perceptual bias induced by the prior exposure block. Furthermore, the prior studies that have used this approach have repeated the same stimulus (a single syllable) during the prior exposure block (i.e. pairing ambiguous audio with video of /aba/ eight times), while the present study used a mixture of stimuli during the exposure blocks. The greater variety during the exposure blocks in the present study prevents listeners from potentially using strategies and guessing as to what is expected during the subsequent test block, but

may have also reduced the strength of the response, so that the effect was thereby lost during the long trials. In addition, we used more complex stimuli (words and pseudo-words) compared to the previous studies, and this may have also led to less specificity in the obtained effects. To avoid these timing-related issues, the study in Chapter 4 used a faster blocked design for the test blocks as well, with shorter trials that could not be individually decoded but still led to retuning and recalibration. Future studies that may continue to pursue exploring retuning and recalibration using fMRI, as well as pattern analysis to decode what listeners perceived, may benefit from using repeated stimuli during exposure blocks, with shorter and less complex stimuli, as well as multiple sessions to accumulate enough trials. Distractor blocks may also help to prevent participants from forming response strategies.

Retuning, recalibration, and current theories of speech perception

Despite some of the limitations described in the studies, we established several conclusions regarding retuning and recalibration, and how listeners exploit regularities in the speech signal to adjust representations of phoneme categories. Theories of speech comprehension can be helpful in elucidating how retuning and recalibration may operate, but are generally geared towards understanding word and speech recognition overall, rather than the process of perceptual learning itself. Modular feedforward models, such as Cohort or Merge, suggest that no top-down information, such as lexical knowledge, is necessary during perception. According to the Cohort model (Gaskell & Marslen-Wilson, 1997), as listeners encounter each segment of a word, a set of possibilities are activated, then narrowed down as the listener continues to parse the remaining segments of the word, and until only one possible match remains. Similarly, the proponents of the Merge model (originally proposed as Shortlist, and later updated to Merge; Norris, McQueen, & Cutler, 2000) have argued that bottom-up, feed-forward connections are sufficient to drive speech perception, and top-down feedback is not necessary during word recognition, as it may not improve accuracy nor increase speed of processing. Based on the incoming auditory input, a subset of word candidates is created, and

inhibitory connections between the candidates (driven by degree of fit with the input) determine which word is chosen.

Conversely, the interactive connectionist model TRACE (McClelland & Elman, 1986) proposes that speech perception encompasses several layers (features, phonemes, and words) between which there are connections that are activated by the incoming auditory signal, and the strength of activity between these connections determines what the listener perceives. Unlike either connectionist or modular theories, according to the fuzzy-logical model of perception (Massaro, 1987; Oden & Massaro, 1978), listeners can piece together the acoustic features of a word (or item), plus any other available cues, and use this process of featural integration to identify what they are most likely hearing by guessing the likelihood of the item belonging to a particular category. Features containing ambiguity are weighed less compared to clear features, and thereby exercise less influence upon the final item selection.

All of these theories share similar concepts in that words are retrieved based on their constituent properties, but depending on the theory, may explain the influence of contextual information by changing the weights between connections, by adding or strengthening certain connections between layers, or by computing a likelihood estimate using all of the incoming information (acoustic, visual, or any other source of information) (see Weber & Scharenborg, 2012). It remains unclear as to the exact point in time in which contextual influences affect speech comprehension, either while phonemes are heard or at a later decision-making stage. With regard to the research presented in this dissertation, Chapter 2 established some of the bounds in flexibility of retuning and recalibration, while Chapter 3 explored whether lexical and audiovisual cues could cooperatively boost perceptual learning effects, but these studies were inconclusive as to the point in time in which contextual cues affect listeners' perception and comprehension. However, the results presented in Chapter 4 point towards the possibility that the contextual influences affect phoneme perception and not the decision alone, as we found significant engagement of HG, PT, STG, and STS during the categorization test blocks, regions which are known to be responsible for elementary acoustic and phonetic processing (Binder, 2000; Mesgarani, Cheung, Johnson, & Chang, 2014;

Mesgarani, David, Fritz, & Shamma, 2008; Yi, Leonard, & Chang, 2019). However, fMRI results alone are not enough to categorically define whether acoustic processing is separated from contextual influences, or if they do indeed overlap, as the timing of activity remains uncertain.

The manner in which higher-level information impacts perception is still debated between researchers, as sources such as lexical knowledge or audiovisual lip-reading may play a role at a later point in time and not necessarily during the reconfiguration process itself. On the other hand, it may be likely that top-down information is needed while listeners to interpret ambiguous acoustic signal, and then apply this knowledge towards shifting the category boundary. However, these sources may not influence what is heard, but rather, how it is interpreted. The guidance of higher-level contextual knowledge combined with the recognition of the degraded acoustic signal may be what ultimately directs retuning and recalibration, but the timing of when this knowledge is relied upon remains disputed.

A model of perceptual learning

A full-fledged model of perceptual learning for phoneme categories is still to be achieved. The aforementioned theories of speech perception have proven useful in understanding how perceptual learning fits into speech perception at large, but mostly do not contain specifics of how perceptual learning is implemented or its outcomes. However, Kleinschmidt & Jaeger (2015) have proposed a more comprehensive account of phonetic adaptation, a Bayesian model of audiovisual recalibration and selective speech adaptation as two endpoints along a continuum of exposure length. Audiovisual recalibration is a result of short exposure, resulting in a bias towards the stimuli presented (perceiving more /p/ after /p/-biased exposure) whereas selective speech adaptation builds over a longer period of time and results in fatigue after exposure (less perceived /p/ after lengthy /p/ exposure). Perhaps a model such as this could be extended to describe the various other forms of perceptual learning, to include the possible cue types, exposure lengths, phoneme types, degree of generalization. This would require a

comprehensive explanation of the phoneme category adjustments resulting from audiovisual and lexical cues, such as many of the studies discussed thus far, and even visual phonetic categories (i.e. visual representations of a speaker pronouncing a phoneme), which can also undergo shifts after exposure to lexical information (van der Zande, Jesse, & Cutler, 2013).

A newer line of research has explored text-based recalibration, or perceptual shifts as a result of exposure to text coupled with ambiguous phonemes (Bonte et al., 2017; Keetels, Schakel, Bonte, & Vroomen, 2016; Romanovska, Janssen, & Bonte, 2019). Similar to lexical retuning, text-based shifts may reflect a top-down influence on phonemes, from a higher level than lexical knowledge or audiovisual cues, similar to previous findings wherein lexical information has been proven capable of guiding letter perception (Norris, Butterfield, McQueen, & Cutler, 2006). Other contextual sources, such as phonotactic information (valid phoneme combinations; i.e. in English, /b/ can be followed by /r/ but not by /n/) or hand gestures can guide speech comprehension (Cutler, McQueen, Butterfield, & Norris, 2008; Drijvers & Özyürek, 2017; Idemaru & Holt, 2011), but may also be capable of guiding phoneme boundary adjustments.

A more complete model may also delve into individual differences, to uncover why some listeners undergo perceptual learning to a greater degree than others. This difference may reflect listeners' general listening abilities (i.e. auditory acuity), or how they are able to adapt to new speakers, or even how they may learn a second-language and acquire new speech sounds. Previous studies have also explored differences in lexical retuning between native and non-native speakers of a language, and non-native speakers can show category shifts to a similar degree as native speakers (Bruggeman & Cutler, 2019; Reinisch & Holt, 2014; Reinisch, Weber, & Mitterer, 2013), but this can be modulated by the proficiency in the second language (Samuel & Frost, 2016). Accordingly, strengthening mappings between degraded speech and phoneme categories could enable non-native speakers to gain proficiency in a new language, and thereby demonstrate shifts in accordance with a given speaker or situation.

Conclusion

The studies described in this dissertation explored two forms of perceptual adaptation under similar constraints in order to compare and contrast their various properties. We presented a paradigm under which both lexical retuning and audiovisual recalibration could be tested, which was then extended into an fMRI study, and allowed us to identify the neural substrates of the two processes. We discovered that retuning and recalibration share some characteristics, such as their ability to be flexibly induced in a short amount of time, and that they both primarily rely on the core areas of the speech network, such as the temporal cortex. However, the two forms of learning also differ, in that they do not appear to be fully independent of the contextual cues themselves, as seen by the lack of additive effects and the significant recruitment of the visual cortex during audiovisual recalibration. We succeeded in unraveling how retuning and recalibration are elicited under the same circumstances, but a number of questions still remain unexplored in order to build a cohesive model of perceptual learning.

References

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

Binder, J. R. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, *10*(5), 512–528. https://doi.org/10.1093/cercor/10.5.512

Bonte, M., Correia, J. M., Keetels, M., Vroomen, J., & Formisano, E. (2017). Reading-induced shifts of perceptual speech representations in auditory cortex. *Scientific Reports*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-05356-3

Bruggeman, L., & Cutler, A. (2019). No L1 privilege in talker adaptation. *Cambridge University Press*. https://doi.org/https://doi.org/10.1017/S1366728919000646

Buchsbaum, B. R., Hickok, G., & Humphries, C. (2001). A multidisciplinary role of left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cognitive Science*, *25*(5), 663–678. https://doi.org/10.1207/s15516709c0g2505_2

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. *Laboratory Phonology 10*, 91–111. https://doi.org/10.1017/CBO9781107415324.004

Cutler, A., McQueen, J. M., Butterfield, S., & Norris, D. (2008). Prelexically-driven perceptual retuning of phoneme boundaries. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2008)* (p 2056) Red Hook, NY: Interspeech.

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. (2006). Perceptual learning in speech: stability over time. *The Journal of the Acoustical Society of America, 119*(4), 1950-1953.

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science, 322*(5903), 970-973. https://doi.org/10.1126/science.1164318

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12*(5-6), 613-656. https://doi.org/10.1080/016909697386646

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance, 37*(6), 1939–1956. https://doi.org/10.1037/a0025641

Jäncke, L., Wüstenberg, T., Scheich, H., & Heinze, H. J. (2002). Phonetic perception and the temporal cortex. *NeuroImage, 15*(4), 733–746. https://doi.org/10.1006/nimg.2001.1027

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

Kilian-Hütten, N., Valente, G., Vroomen, J., & Formisano, E. (2011). Auditory cortex encodes the perceptual interpretation of ambiguous sound. *The Journal of Neuroscience, 31*(5), 1715–1720. https://doi.org/10.1523/jneurosci.4572-10.2011

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

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56*(1), 1–15. https://doi.org/10.1016/j.jml.2006.07.010

Lüttke, C. S., Ekman, M., Van Gerven, M. A. J., & De Lange, F. P. (2016). McGurk illusion recalibrates subsequent auditory perception. *Scientific Reports, 6*(32891), 1–7. https://doi.org/10.1038/srep32891

Massaro, D. W. (1987). *Speech Perception By Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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)900150

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1010. https://doi.org/10.1126/science.1245994

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *The Journal of the Acoustical Society of America*, *123*(2), 899–909. https://doi.org/10.1121/1.2816572

Mitchel, A. D., Gerfen, C., & Weiss, D. J. (2016). Audiovisual perceptual learning with multiple speakers. *Journal of Phonetics*, *56*, 66–74. https://doi.org/10.1016/j.wocn.2016.02.003

Norris, D., Butterfield, S., McQueen, J. M., & Cutler, A. (2006). Lexically guided retuning of letter perception. *Quarterly Journal of Experimental Psychology*, *59*(9), 1505–1515. https://doi.org/10.1080/17470210600739494

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *The Behavioral and Brain Sciences*, *23*(3), 299–325. https://doi.org/10.1017/S0140525X00003241

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

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85*(3), 172-191. https://doi.org/10.1037/0033-295X.85.3.172

Pearson, J. (2019). The human imagination: the cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience, 20,* 624-634. https://doi.org/10.1038/s41583-019-0202-9

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 539–555. https://doi.org/10.1037/a0034409

Reinisch, E., Weber, A., & Mitterer, H. (2013). Listeners retune phoneme boundaries across languages. *Journal of Experimental Psychology: Humam*

Perception and Performance, *39*(1), 75–86. https://doi.org/10.1121/1.3655312

Romanovska, L., Janssen, R., & Bonte, M. (2019). Reading-induced shifts in speech perception in dyslexic and typically reading children. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.00221

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, *12*(4), 348–351. https://doi.org/10.1111/1467-9280.00364

Samuel, A. G., & Frost, R. (2015). Lexical support for phonetic perception during nonnative spoken word recognition. *Psychonomic Bulletin & Review, 36*(5), 1746–1752. https://doi.org/10.1002/jmri.23741.Proton

Sjerps, M. J., & McQueen, J. M. (2010). The bounds on flexibility in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(1), 195–211. https://doi.org/10.1037/a0016803

Van der Zande, P., Jesse, A., & Cutler, A. (2013). Lexically guided retuning of visual phonetic categories. *The Journal of the Acoustical Society of America*, *134*(1), 562–571. https://doi.org/10.1121/1.4807814

Van der Zande, P., Jesse, A., & Cutler, A. (2014). Cross-speaker generalisation in two phoneme-level perceptual adaptation processes. *Journal of Phonetics*, *43*, 38–46. https://doi.org/10.1016/j.wocn.2014.01.003

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(3), 387–401. https://doi.org/10.1002/wcs.1178

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron*, *102*(6), 1096–1110. https://doi.org/10.1016/j.neuron.2019.04.023

6

Knowledge valorization

Valorization

Speech is essential for human interaction, but it is not accessible or experienced by everyone in the same manner. 6.1% of the world's population is estimated to have hearing loss (World Health Organization, 2020), and many rely on hearing devices or cochlear implants to be able to listen and communicate. However, these devices do not always operate optimally; they may amplify background noise or other sounds irrelevant to the listener, so users of these devices cannot solely rely on the now-amplified auditory signal in order to understand speech. Consequently, users of hearing devices, as well as others with hearing impairments who do not use such devices, may rely on information other than the acoustic signal itself to guide speech perception. Such populations may utilize lip-reading (also known as speech-reading) to support speech comprehension when the available acoustic signal is inadequate.

The studies presented in this dissertation have touched upon lip-reading, and specifically addressed various ways in which listeners can use contextual information to guide perceptual shifts of phonetic categories, particularly through knowledge of the lexicon and by attending to lip-reading cues. This line of inquiry has highlighted the importance of the non-acoustic contextual cues contained in speech, and how they can reshape what a listener hears and lead to shifts in internal representations of phoneme categories. The results of these studies hold implications for improving and refining educational strategies for lip-reading. Lip-reading can support speech comprehension, and while most listeners use lip-reading cues to some extent (and uniquely evidenced by the McGurk effect), for listeners with hearing impairments, lip-reading may supplement or even replace the auditory signal. Training in lip-reading involves conscious concentration on lip-movements being produced by the speaker in order to enhance recognition. Listeners thereby learn to build stronger links between singular and/or sequences of lip-movements with phonemes, syllables, and words. However, lip-movements alone may not convey enough information for the listener to interpret the speaker, as multiple phonemes map onto the same viseme (i.e. /pa/ and /ba/ are visually identical). Therefore, lexical knowledge also plays an in important role in lip-

reading and can be an additional source of clarification. Training and educational strategies that incorporate both components may be more useful than either on their own, as each cue individually may be insufficient. Lexical knowledge and semantic context can be useful for the listener, so as to narrow the possible items of what the speaker is most likely to be saying, such as a word rather than a non-word (*bottle* versus *pottle*), the word most probable depending on the remainder of the sentence or phrase (baseball *bat* versus *pat),* or based on word frequency within a language (*pear* versus *bear*). Building strong links between visemes and sequences of lip-movements, along with their respective words may make lip-reading faster and more efficient. Lip-reading education already incorporates both lexical and audiovisual aspects, but potential advancements in lip-reading should place further emphasis on strengthening the mappings between phonemes, visemes, syllables and the lexicon. A multimodal approach to lip-reading and speech recognition featuring salient, non-acoustic contextual cues is more likely to benefit listeners struggling to comprehend speech, than strategies focused entirely on learning lip-movements and visemes themselves.

In conclusion, it is important to consider combining contextual cues when training listeners in lip-reading, as the combination of multiple contextual sources may be more useful to listeners who cannot rely on the auditory signal alone, and each source individually might not be a sufficient source of guidance. Investigating speech perception is not only essential for understanding a fundamental human experience, but is also necessary in order to make improvements upon technological devices designed for speech and communication purposes.

Acknowledgments

First, thank you to my supervisors, Elia and Anne, for giving me this wonderful opportunity. Anne, thank you so much for your guidance and encouragement, I admire your enthusiasm and aspire to have the same attitude. Elia, thank you so much for your support and patience, and for giving me the freedom to work independently. I feel very lucky to have been able to learn from researchers as knowledgeable as the both of you. And thanks for taking a chance on a seemingly strange person who was willing to leave California for Maastricht (I didn't realize it could actually rain this much). Thank you also to Lars, for being both a great officemate and a great mentor/co-supervisor, the fMRI study is all thanks to you.

Thank you to the secretaries of the CN department – Christl, Eva, José, and Riny, for always being so helpful and kind. Thank you to Language in Interaction for funding this project and to the other members of the consortium for your helpful advice throughout this project.

To all my other colleagues and friends, thank you for making this experience so special, I think this department is truly one-of-a-kind. To the Auditory group, thanks for creating such a nice and welcoming atmosphere during meetings and trips, I enjoyed the time we spent together. Linda, I'm so glad I discovered that you're a fellow Beyhive member and Drag Race fan. And to the best lunch group ever, Amaia, Anita, & Laurien, (aka WC, but currently missing Hannah & Tabea), you all made this experience so memorable and fun. I wouldn't have made it without the (noice) laughs, advice, and (occasional) venting sessions.

Muchisimas gracias a Iñigo, eres *literally!* (en la voz de Chris) lo mejor. And finally, to my parents for your all your love and support over the years, for trusting and believing in me, thanks for everything.

Curriculum vitae

Shruti Ullas was born on November 11th, 1992 in Calcutta (India). She attended Monta Vista High School and De Anza College in Cupertino (United States) until 2010, then received her bachelor's degree in Cognitive Science from the University of California, Los Angeles (UCLA) in 2012. In 2013, she received her master's degree in Human Cognitive Neuropsychology from the University of Edinburgh. In 2014, she started her PhD at Maastricht University in the Faculty of Psychology and Neuroscience (Department of Cognitive Neuroscience) under the supervision of Prof. Elia Formisano, Dr. Frank Eisner, and Prof. Anne Cutler.

Publications

Peer-reviewed journals:

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

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

Haas, B.W., Barnea-Goraly, N., Sheau, K.E., Yamagata, B, Ullas, S., & Reiss, A.L. (2013). Altered microstructure within social-cognitive brain networks during childhood in Williams syndrome. *Cerebral Cortex*, 24(10), 2796-2806. doi:10.1093/cercor/bht135.

Ullas, S., Hausfeld, L., Cutler, A., Eisner, F., & Formisano, E. (under review). Neural correlates of phonetic adaptation as induced by lexical and audiovisual context. *Journal of Cognitive Neuroscience*.

Book chapter:

Ullas, S., Bonte, M., Formisano, E., & Vroomen, J. (2020, forthcoming). Adaptive Plasticity in Perceiving Speech Sounds. *Springer Handbook of Auditory Research: Auditory Cognitive Neuroscience.*

Conference contributions:

Ullas, S., Hausfeld, L., Eisner, F., Cutler, A., & Formisano, E. Lexical and audiovisual information as sources of phoneme boundary recalibration. *Auditory Cortex*, August 2017, Banff, Canada.

Ullas, S., Eisner, F., Cutler, A., & Formisano, E. Recalibration of phonetic categories using lexical and audiovisual information. *Donders Discussions*, November 2016, Nijmegen, Netherlands.

Ullas, S., Eisner, F., Cutler, A., & Formisano, E. Lexical and lip-reading information as sources of phoneme boundary recalibration. *Society for Neurobiology of Language*, August 2016, London, United Kindgom.

Haas, B.W., Barnea-Goraly, N., Sheau, K.E., Yamagata, B, Ullas, S., & Reiss, A.L. Genetic effects on the micro-structure of the amygdala, fusiform, hippocampus in humans. *Cognitive Neuroscience Society Annual Meeting*, March 2012, Chicago, USA.