Adaptive Plasticity in Perceiving Speech Sounds

Adaptive Plasticity in Perceiving Speech Sounds

Chapter · February 2022

DOI: 10.1007/978-3-030-81542-4_7

Milene Bonte
Maastricht University

Elia Formisano
Maastricht University

Jean Vroomen
Tilburg University

Chapter 7
Adaptive Plasticity in Perceiving Speech Sounds

Shruti Ullas, Milene Bonte, Elia Formisano, and Jean Vroomen

Abstract Listeners can rely on perceptual learning 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 experiments tend to use a targeted approach by embedding ambiguous phonemes into speech or speechlike items, and with continuous exposure, a learning effect can be induced in listeners, wherein disambiguating contextual information shifts the perceived identity of the same ambiguous sound. The following chapter will review current 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.

Keywords Recalibration · Perceptual learning · Speech perception · Phonetic processing · Lexical processing · Audiovisual speech · Speech-reading

7.1 Introduction

Speech perception is seemingly easy and automatic to the listener, and for healthy young listeners, 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

S. Ullas (✉) · M. Bonte · E. Formisano
Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands

Maastricht Brain Imaging Center, Maastricht University, Maastricht, The Netherlands
e-mail: shruti.ullas@maastrichtuniversity.nl; m.bonte@maastrichtuniversity.nl; e.formisano@maastrichtuniversity.nl

J. Vroomen
Department of Cognitive Neuropsychology, Tilburg University, Tilburg, The Netherlands
e-mail: j.vroomen@tilburguniversity.edu

L. Holt et al. (eds.), *Speech Perception*, Springer Handbook of Auditory Research 74, https://doi.org/10.1007/978-3-030-81542-4_7

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. The listener is able to easily resolve these inconsistencies and understand what is spoken. 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 speakers 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. 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 and lexically guided perceptual learning. Contextual information can guide the retuning process of phoneme category boundaries, after continuous exposure to speech or videos of speechlike 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 adjust a perceptual boundary between two phonemes by hearing ambiguous phonemes embedded into words. Similarly, Bertelson et al. (2003) identified a comparable effect as recalibration, where listeners utilized visual or speech-reading information to adjust the perceptual 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 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 aftereffect 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, recalibration and perceptual learning are typically measured in two phases, starting with an exposure phase and followed by a test phase (see Kraljic and Samuel 2009, for an overview). In the approach used to measure lexically guided perceptual learning, exposure stimuli are composed of audio recordings of words, whereas in audiovisual recalibration experiments, exposure stimuli comprise videos of 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, with words such as "half" and "paragraph" edited to remove the clear /f/ and replaced with the ambiguous version. Because "half" and "paragraph" are real words in English, whereas "halss" and "paragrass" are not, listeners tend to perceive the ambiguous token as an /f/. 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, as the exact same ambiguous tokens can be learned as different phonemes depending on the contextual cues. Perceptual retuning and recalibration studies (Bertelson et al. 2003; Norris et al. 2003; Krajlic and Samuel 2009) also reveal how flexible the units of speech are, and how they can be adapted depending on the surroundings of the listener. 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.

With the advancement of neuroimaging technologies, the ways in which the brain incorporates these perceptual shifts have been explored with greater detail and have revealed the areas of the brain likely to be involved in these processes. Techniques such as functional MRI (fMRI; see Table 7.1 for abbreviations) and electrocorticography (ECoG) recordings have proven especially useful in elucidating the potential neural mechanisms (Hickok and Poeppel 2007; Mesgarani et al. 2014). These findings, combined with existing theories of speech perception, are useful for understanding how the brain adapts to unclear speech and how the necessary changes may be implemented at the neural level.

This chapter will present an overview of the current literature regarding lexical (Sect. 7.2.1) and audiovisual influences (Sect. 7.3.1) on phoneme boundary recalibration, as well as some related works on selective speech adaptation (Sect. 7.3.2). Changes over time (Sect. 7.2.2), generalization over speakers and sounds (Sects. 7.2.3 and 7.3.3), and other features (Sect. 7.2.4) will also be discussed, as well as a comparison between lexical and audiovisual perceptual learning (Sect. 7.4).

Table 7.1 Table of abbreviations

Abbreviation	Full name
ECoG	Electrocorticography
EEG	Electroencephalogram
fMRI	Functional MRI
IFS	Inferior frontal sulcus
IPL	Inferior parietal lobe
ITS	Inferior temporal sulcus
MEG	Magnetoencephalogram
MTG	Medial temporal gyrus
PT	Planum temporale
STG	Superior temporal gyrus
STS	Superior temporal sulcus
SWS	Sine-wave speech

Theories and neuroimaging studies that may explain the underlying mechanisms of recalibration will also be reviewed (Sect. 7.5), followed by a final conclusion and summary (Sect. 7.6).

7.2 Lexical Knowledge and Auditory Perception

7.2.1 Introduction to Lexically Guided Perceptual Learning

As mentioned earlier in the introduction (Sect. 7.1), 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). Other researchers 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 a 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.

Extending further from 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. 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 the sounds as /s/. A third control group heard pseudo-words containing the ambiguous phoneme to test whether the absence of any lexical information could impact subsequent categorization. This group showed no bias toward either phoneme during the test phase. An example of the pattern of results is shown in Fig. 7.1.

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, and replicated the Ganong effect. 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 the said token. Thus, phoneme category 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. 7.2.3).

7.2.2 Perceptual Learning 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

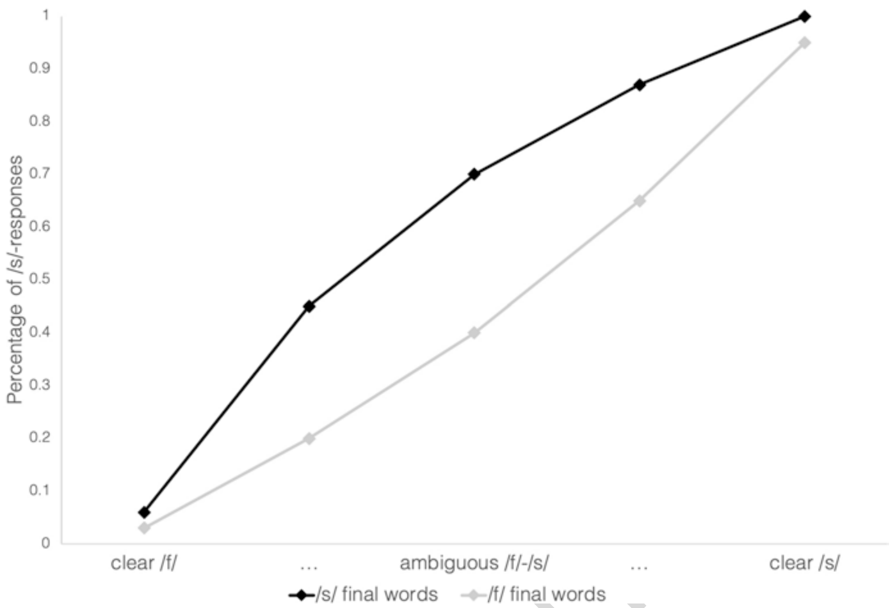

Fig. 7.1 Example graph of perceptual retuning results. After exposure to edited words, participants are presented with a continuum of sounds ranging from clear /f/ to clear /s/ in a test phase. Participants who hear words typically containing /f/ replaced with an ambiguous /f/-/s/ blend are likely to report hearing /f/ during the test phase (shown in gray), while participants who heard the same sound replacing /s/ in /s/-final words are likely to report hearing more /s (shown in black)

approach as Norris et al. (2003), testing native English speakers using words containing either /s/ or /ʃ/ (the "sh" sound in shoe), with items such as *eraser* and *publisher*. 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. Perceptual learning is seemingly unaffected by long gaps between exposure and test, which suggests that lexically guided perceptual learning is largely stable over the order of hours.

7.2.3 Generalization of Perceptual Retuning

Although lexically driven perceptual learning appears to be quite robust, other investigators have identified the limitations of such learning. For example, perceptual learning tends to be restricted by the stimuli, particularly by the speakers of the tokens. The shift in perception resulting from experience with one phoneme pair by one speaker may not apply to the same pair produced by a new speaker. 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 toward 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 toward 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.

Similarly, generalization to new speakers may also be dependent on the accent of the speaker. Kraljic et al. (2008a) compared effects of speaker characteristics on perceptual learning, with an idiosyncratic pronunciation versus an accent commonly known to the participants. The idiosyncrasy, or speaker-specific version, was designed by placing an ambiguous /s/-/ʃ/ sound before any 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 some regional American accents. Phoneme boundary retuning was not successful in the latter group that was exposed to the tokens typical of the accented speech, but was detected in the non-accented group. Knowledge of reasonable and unrealistic deviations, which may be implicit or explicit, also seem to impact perceptual learning. In contrast, 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), and to a lesser extent to speakers whose voices were acoustically more distant. The discrepancy in findings between Xie and Myers (2017) and Krajlic and Samuel (2008a) may once again reflect differences in learning effects due to the phoneme pair used. AU1

Just as speaker specificity of perceptual learning is tied to the type of phoneme pairs, the same applies to generalization across phoneme pairs within a single speaker. Kraljic and Samuel (2006) saw that perceptual learning 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 and a /b/-/p/ continuum. Participants were able to extend retuning to the /b/-/p/ continuum in the same direction of voicing, or the point in time at which the vocal folds vibrate, where /b/ and /d/ are voiced, whereas /d/ and /t/ are unvoiced. Participants who heard words with an ambiguous /b/ were more likely to report a greater amount of both /b/ along the /b/-/p/ continuum, as well as more /d/ during an additional test phase on a continuum of /d/-/t/. Mitterer et al. (2013) also explored phoneme specificity by creating exposure stimuli using Dutch words ending in an approximant /r/ (the /r/ in red) or 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/ (not in American English phonology but similar to the t-sound in *better*) or a light /l/ (the /l/ in *leaf*). Once again, the specificity of recalibration seems to be dependent on the acoustic features of the phoneme pair being learned.

Overall, it appears that retuning is often phoneme- and speaker-specific, but 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.

7.2.4 Other Attributes of Perceptual Learning

Most studies of the lexically guided perceptual learning studies described throughout Sect. 7.2 are twofold. 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 measurable 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 2015), suggesting that recalibration cannot be performed automatically and requires attentional resources. But attention alone is also not enough to induce retuning, as can listeners 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/, in place of the oft-mentioned ambiguous phoneme (Sjerps and McQueen 2010). Again, the acoustic or perceptual similarity 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 Sect. 7.2, 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 these 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. 2015). 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).

7.2.5 Summary of Lexically Driven Perceptual Learning

Section 7.2 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.

7.3 Audiovisual Information and Speech

7.3.1 Overview of 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. 7.2, 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 (Fig. 7.2).

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 Sect. 7.3.2.

7.3.2 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

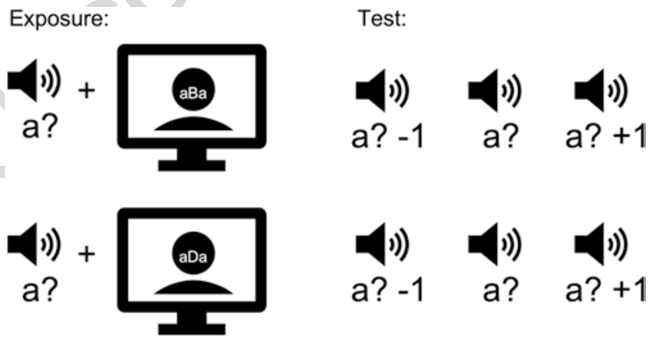

Fig. 7.2 A typical audiovisual recalibration procedure. Exposure phases pair ambiguous phoneme blends (such as an /aba/-/ada/ blend) with video of a speaker pronouncing one of the two phonemes (/aba/ or /ada/). Following exposure to these videos, listeners are then presented with the auditory items (the ambiguous /aba/-/ada/ blend, along with other similar sounds) and asked to respond with what they hear

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, 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. 7.3.1.

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, 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 buildup 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 audiovisual recalibration effects are short-lived and can be present only up until approximately 6 test trials, while selective speech adaptation effect can be continuously sustained for up to 60 test items (Vroomen et al. 2004).

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 (Remez et al. 1981). These stimuli are often unintelligible unless listeners are explicitly told that the sounds have been extracted from actual speech. Vroomen and Baart (2009b) also compared recalibration and selective speech adaptation in groups that viewed audiovisual SWS tokens as speechlike versus non-speechlike. 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 including and 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, which 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 measurable 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 on perception and listeners still adapted accordingly. The awareness of speechlike 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. 7.5.

7.3.3 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. While recalibration is robust enough to not depend on working memory (Baart and Vroomen 2010), 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. Therefore, audiovisual recalibration appears to be constrained by the acoustics features, as learning could not extend to other phonemes, or even to the same phonemes paired with different vowels. The ear itself can limit recalibration (Keetels et al. 2016a, b), as 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. 7.5). Notably, listeners also have the capacity to recalibrate each ear in opposite directions using the same ambiguous sounds, e.g., one ear recalibrated toward /aba/, the other toward /ada/, with test sounds presented into the corresponding ears of the exposure phase (Keetels et al. 2015). Thus, phoneme representations may not be completely abstracted from the input received and can retain speaker- and context-specific details. 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 overgeneralize.

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). 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. Similarly, audiovisual recalibration is most often tested with consonant contrasts, but Franken et al. (2017) have found that recalibration is possible with a vowel contrast pair of /e/-/ø/. In addition, recalibration with a vowel pair and multiple speakers has also been observed, wherein the gender identity of the speakers combined with the visual cue indicated by the speech-reading information influenced listeners' categorization responses (Burgering et al. 2020).

The majority of the studies described 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 8 years old. Van Linden and Vroomen (2008) measured recalibration effects in two groups of children and determined that children at 8 years old could recalibrate with audiovisual stimuli, but children at 5 years old could not, so the ability may be developed within this window of 3 years. Dyslexia does not restrict the effect 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. Even children with dyslexia are capable of undergoing recalibration driven by text (Romanovska et al. 2019), even though children with dyslexia often experience difficulties in speech-reading and letter-speech sound mappings (Snowling 1980; van Laarhoven et al. 2018).

7.3.4 Summary of Audiovisual Recalibration

Section 7.3 described audiovisual recalibration, originally described by Bertelson et al. (2003), and its various attributes. Later studies by Vroomen and colleagues have established the general buildup 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 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.

7.4 Comparison of Audiovisual Recalibration and Lexical Retuning

Sections 7.2 and 7.3 have discussed audiovisual recalibration and lexical retuning separately, but the two processes also share many common attributes. 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. Audiovisual effects were similar irrespective of faster and slower response times, while lexical information showed a weaker effect overall and was associated with 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 *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 toward /p/, and the remaining half were biased toward /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 toward one phoneme. Theories of top-down and bottom-up processing will be discussed in more depth in Sect. 7.5.

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. 7.2), such as /f/ and /s/ (Kraljic and Samuel 2007). Overall, lexical and audiovisual recalibrations seem to be markedly similar, although the pathways supporting them may not be identical, and may only overlap.

The two types of retuning 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. 7.2, 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.

More recently, studies comparing audiovisual recalibration and lexical retuning within both a single session and the same participants have found that the resulting effects were similar between the two, with similar patterns of dissipation as well (Ullas et al. 2020a). The simultaneous presentation of both audiovisual and lexical information within exposure (i.e., listeners presented with videos of words edited to contain an ambiguous final phoneme) also showed effects comparable to audiovisual recalibration alone, suggesting that the combination leads to no benefit in subsequent phoneme boundary retuning as a result of differences in the pathways involved in the two forms of perceptual learning (Ullas et al. 2020b). Overall, 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.

7.5 Theoretical and Neural Explanations of Recalibration

7.5.1 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, so recognition and categorization operate at a pre-lexical level. Feedback during categorization could be time-consuming or lead to misinterpretations of the input, 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 do 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.

While most classical theories of speech perception have not accounted for the role of visual information, more recently, Kleinschmidt and Jaeger (2011) have put forth a belief-updating model based on Bayesian inference, by using data from previous studies of recalibration and selective speech adaptation to calculate probabilities of outcomes. This model, called the Ideal Adaptor Framework, is tailored to explain audiovisual recalibration and selective speech adaptation. As described in Sect. 7.3.2, audiovisual recalibration and selective speech adaptation are two forms of perceptual learning, but their response profiles are in direct contrast. According to 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. 7.2.2, 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.

7.5.2 Neural Basis of Recalibration and Perceptual Learning

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 toward connecting phonological and lexical representations, while regions along a dorsal pathway, including parietal-temporal, (pre)motor, and inferior frontal regions, are geared toward connecting phonological with sensorimotor and articulatory representations. Adank and Devlin (2010) also explored how listeners adjust to recordings of unclear sentences and found activation patterns consistent with the Hickok and Poeppel (2007) model. Jäncke et al. (2002) also identified structures of the brain in the planum temporale (PT) and middle superior temporal gyrus (STG) that are specific to phoneme perception. 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 inferior frontal gyrus (Rogers and Davis 2017) and the supramarginal gyrus (Raizada and Poldrack 2007; see Davis and Johnsrude 2007 for a review).

While these studies paved the way toward 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 based on prior neuroimaging studies of speech perception, reminiscent of the Merge model (with similarities to TRACE as well, but this model focuses on word recognition and not on abstraction). 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. Similarly, Holdgraf et al. (2016) have found evidence for acoustic updating, using spectro-temporal receptive field mapping on ECoG recordings of the auditory cortex. Responses of cortical populations were observed to have increased sensitivity to speechlike 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 fMRI studies of recalibration and retuning. 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), superior temporal sulcus and superior temporal gyrus (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 audio-visual recalibration experiment, solely using the activation patterns. Active clusters were found in and around left PT and left Heschl's gyrus and sulcus, which are typically viewed as low-level auditory areas, but they may have been influenced by information other than rudimentary acoustics features as they effectively predicted the percepts that were driven by the visual cue and not the auditory information alone.

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 SWS 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 nonphonemic, 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 (Liebenthal et al. 2010). AU5

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 toward /s/ for one group, and toward /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, the results of these two studies of lexical context imply that perceptual learning involves areas responsible for both lower and higher levels of information processing in resolving the perception of these sounds. However, it remains unclear as to whether the flow of information is simply feed-forward or not, as the exact timing as to when each region is engaged is not yet understood. The authors suggest that initial processing of the unclear sounds relies on higher-level executive regions, but once the listener undergoes sufficient training and has shifted the perceptual boundary, then regions responsible for lower levels of processing, such as STG, can be activated in response to the ambiguous sound.

Combined magnetoencephalogram (MEG) and electroencephalogram (EEG) data have also confirmed that activity in STG reduced over time, as participants learned to improve in identification of degraded speech sounds combined with matching text (Sohoglu and Davis 2016). Furthermore, the results were framed within a model of predictive coding, not unlike Bayesian inference, such that the listener learns to reduce prediction errors as a consequence of learning. STG is proposed to process acoustic features and receives 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 (Guediche et al. 2015). Listeners learned to identify words distorted by noise vocoding, and consequently, cerebellar regions showed changes, as well as functional connections to cortical language and auditory regions. Stemming in part from this finding, another model of speech adaptation has been proposed, also relying on a predictive coding mechanism, but supervised by the cerebellum (see Guediche et al. 2014, for a complete review). In contrast, some areas of the brain may be uniquely engaged by either recalibration or retuning. When compared directly using fMRI within the same participants, audiovisual recalibration and lexical retuning showed largely similar areas of activation, over temporal, parietal, and motor cortex areas, although audiovisual recalibration specifically seems to retrigger activation within areas of the visual cortex, despite the lack of visual stimuli during the recalibration test trials (Ullas et al. 2020).

7.5.3 Summary of Theories of Speech Perception

Section 7.5 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 2011), 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 remains unclear and may require more advanced neuroscientific methods.

7.6 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 7.2 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 7.3 discussed audiovisual recalibration, first described by Bertelson et al. (2003) and expanded upon by others.

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. 2016a, b), 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. 2016a, b). 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 PT 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 are 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.

Compliance with Ethics Requirements Shruti Ullas declares no conflict of interest.
Milene Bonte declares no conflict of interest.
Elia Formisano declares no conflict of interest.
Jean Vroomen declares no conflict of interest.

References

AU6

Adank P, Devlin JT (2010) On-line plasticity in spoken sentence comprehension: adapting to time-compressed speech. NeuroImage 49(1):1124–1132. https://doi.org/10.1016/j.neuroimage.2009.07.032

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. https://doi.org/10.1016/j.jml.2015.06.008

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

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

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

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. https://doi.org/10.1523/JNEUROSCI.4339-13.2014

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. https://doi.org/10.1038/s41598-017-05356-3

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

Brancazio L (2004) Lexical influences in audiovisual speech perception. J Exp Psychol Hum Percept Perform 30(3):445–463. https://doi.org/10.1037/0096-1523.30.3.445

Burgering M, van Laarhoven T, Baart M, Vroomen J (2020) Fluidity in the perception of auditory speech: cross-modal recalibration of voice gender and vowel identity by a talking face. Q J Exp Psychol (Hove) 73(6):957–967. https://doi.org/10.1177/1747021819900884

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

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

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

Davis MH, Johnsrude IS (2007) Hearing speech sounds: top-down influences on the interface between audition and speech perception. Hear Res 229(1–2):132–147. https://doi.org/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. https://doi.org/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. https://doi.org/10.1016/j.neuroimage.2004.09.039

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

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

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

Eisner F, McQueen JM (2006) Perceptual learning in speech: stability over time. J Acoust Soc Am 119:1950–1953. https://doi.org/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. https://doi.org/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, pp 655–658. https://doi.org/10.21437/Interspeech.2017-122

Ganong WF (1980) Phonetic categorization in auditory word perception. J Exp Psychol Hum Percept Perform 6(1):110–125. https://doi.org/10.1037/0096-1523.6.1.110

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

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. https://doi.org/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. https://doi.org/10.1093/cercor/bht428

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402. https://doi.org/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. https://doi.org/10.1038/ncomms13654

Holt LL, Lotto AJ (2008) Speech perception within an auditory cognitive science framework. Curr Dir Psychol Sci 17(1):42–46. https://doi.org/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. https://doi.org/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. https://doi.org/10.1016/j.cognition.2015.04.019

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

Keetels MN, Stekelenburg JJ, Vroomen J (2016b) A spatial gradient in phonetic recalibration by lipread speech. J Phon 56:124–130. https://doi.org/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. https://doi.org/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. https://doi.org/10.1016/j.neuroimage.2011.05.043

Kleinschmidt DF, Jaeger TF (2011) Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. Psychol Rev 122(2):148–203. https://doi.org/10.1037/a0038695

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

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

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

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

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

Kraljic T, Samuel AG, Brennan SE (2008b) First impressions and last resorts: how listeners adjust to speaker variability. Psychol Sci 19:332–338. https://doi.org/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. https://doi.org/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. https://doi.org/10.1093/cercor/bhq045

Lüttke C, Ekman M, van Gerven M, de Lange FP (2016) McGurk illusion recalibrates subsequent auditory perception. Sci Rep 6:32891. https://doi.org/10.1038/srep32891

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

McClelland JL, Elman JL (1986) The TRACE model of speech perception. Cogn Psychol 18:1–86. https://doi.org/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. https://doi.org/10.1016/j.tics.2006.06.007

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

McQueen JM, Cutler A, Norris D (2006a) Phonological abstraction in the mental lexicon. Cogn Sci 30:1113–1126. https://doi.org/10.1207/s15516709cog0000_79

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

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. https://doi.org/10.1121/1.2816572

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

Mitterer H, Scharenborg O, McQueen JM (2013) Phonological abstraction without phonemes in speech perception. Cognition 129:356–261. https://doi.org/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. https://doi.org/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. https://doi.org/10.1093/cercor/bhm053

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

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

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

Norris D, McQueen JM, Cutler A (2016) Prediction, Bayesian inference and feedback in speech recognition. Lang Cogn Neurosci 31(1):4–18. https://doi.org/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. https://doi.org/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. https://doi.org/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 Hum Percept Perform 40(2):539–555. https://doi.org/10.1037/a0034409

Reinisch E, Weber A, Mitterer H (2013) Listeners retune phoneme categories across languages. J Exp Psychol Hum Percept Perform 39:75–86. https://doi.org/10.1037/a0027979

Reinisch E, Wozny D, Mitterer H, Holt LL (2014) Phonetic category recalibration: what are the categories? J Phon 45:91–105. https://doi.org/10.1016/j.wocn.2014.04.002

Remez RE, Rubin PE, Pisoni DB, Carell TD (1981) Speech perception without traditional speech cues. Science 212:947–950

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

Rogers JC, Davis MH (2017) Inferior frontal cortex contributions to the recognition of spoken words and their constituent speech sounds. J Cogn Neurosci 29(5):919–936. https://doi.org/10.1162/jocn_a_01096

Romanovska L, Janssen R, Bonte M (2019) Reading-induced shifts in speech perception in dyslexic and typically reading children. Front Psychol 10:221. https://doi.org/10.3389/fpsyg.2019.00221

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

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

Snowling MJ (1980) The development of grapheme-phoneme correspondence in normal and dyslexic readers. J Exp Child Psychol 29:294–305. https://doi.org/10.1016/0022-0965(80)90021-1

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

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

Ullas S, Hausfeld L, Cutler A, Eisner F, Formisano E (2020) Neural correlates of phonetic adaptation as induced by lexical and audiovisual context. J Cogn Neurosci:1–14. https://doi.org/10.1162/jocn_a_01608

Ullas S, Formisano E, Eisner F, Cutler A (2020a) Interleaved lexical and audiovisual information can retune phoneme boundaries. Atten Percept Psychophys 82:2018–2026. 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. Psychon Bull Rev 27:707–715. https://doi.org/10.3758/s13423-020-01728-5

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. https://doi.org/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. Dev Sci 21(1):e12504. https://doi.org/10.1111/desc.12504

Van Linden S, Vroomen J (2007) Recalibration of phonetic categories by lipread speech versus lexical information. J Exp Psychol Hum Percept Perform 33(6):1483–1494. https://doi.org/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. https://doi.org/10.1017/S0305000908008817

Vroomen J, Baart M (2009a) Phonetic recalibration only occurs in speech mode. Cognition 110(2):254–259. https://doi.org/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. https://doi.org/10.1177/0023830909103178

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. https://doi.org/10.1016/j.specom.2004.03.009

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. https://doi.org/10.1016/j.neuropsychologia.2006.01.031

Winn M (2018) Speech: it's not as acoustic as you think. Acoust 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. https://doi.org/10.1016/j.jml.2017.07.005

Zhang X, Samuel AG (2015) Perceptual learning of speech under optimal and adverse condition. J Exp Psychol Hum Percept Perform 40(1):200–217. https://doi.org/10.1037/a0033182

Author Queries

Chapter No.: 7   0005197830

Queries	Details Required	Author's Response
AU1	"Krajlic and Samuel (2008a), Samuel (2016), Eimas and Corbitt (1973), Vroomen et al. (2006), Scott and Johnsrude (2003), Rauschecker and Scott (2009)" is cited in the body but its bibliographic information is missing. Kindly provide its bibliographic information. Otherwise, please delete it from the text/body.	
AU2	The citation "Keetels et al. (2016)" has been changed to "Keetels et al. (2016a, b)" to match the author name/date in the reference list. Please check if the change is fine in this occurrence and modify the subsequent occurrences, if necessary.	
AU3	The citation "Gaskell and Marslen-Wilson 1987" has been changed to "Gaskell and Marslen-Wilson 1997" to match the author name/date in the reference list. Please check if the change is fine in this occurrence and modify the subsequent occurrences, if necessary.	
AU4	The citation "Kleinschmidt and Jaeger (2015)" has been changed to "Kleinschmidt and Jaeger (2011)" to match the author name/date in the reference list. Please check if the change is fine in this occurrence and modify the subsequent occurrences, if necessary.	
AU5	"Sect. 1.1" is not available in this chapter. Please check and provide alternate citation	
AU6	References "Baart & Samuel (2015), Norris et al. (2006), Reinisch & Holt (2014), Roberts & Summerfield (1981), Vroomen & Baart (2009a)" were not cited anywhere in the text. Please provide in text citation or delete the reference from the reference list.	