Journal of Phonetics 104 (2024) 101313

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Spatial location does not consistently constrain perceptual learning in speech

Holger Mitterer[a,*], Eva Reinisch[b]

[a] University of Malta, Malta
[b] Acoustics Research Institute, Austrian Academy of Sciences, Austria

ARTICLE INFO

Article history:
Received 26 July 2023
Received in revised form 26 February 2024
Accepted 29 February 2024
Available online 23 March 2024

Keywords:
Phonetic recalibration
Speech perception
Location specificity
Generalization
Variation

ABSTRACT

Recent research showed selectivity of perceptual learning in speech to linguistic variables and non-linguistic variables. With regard to the latter Keetels et al. (2016) reported that perceptual learning for one spatial location does not fully generalize to another. This spatial selectivity has been suggested to indicate that learning may target non-linguistic representations. We test whether spatial selectivity is a general property of perceptual learning or whether it is related to specific design choices, such as using a single nonword throughout the study. Therefore, we aimed to replicate spatial selectivity with a paradigm that makes use of a larger set of word and nonword stimuli. However, in three experiments, one in-person and two web-based, no effect of spatial selectivity was observed. A Bayesian analysis suggests that the null hypothesis is better supported by the data than the alternative hypothesis based on the previously reported effect size. Repercussions for the debate about pre-lexical representations in speech processing are discussed.

1. Introduction

Even though phonetic categories of the native language are learned early in language acquisition (Pallier et al., 1997), it is now well established that they remain flexible throughout the life span, for instance, to adjust to different talkers and situations (Bertelson et al., 2003; Norris et al., 2003; Samuel & Kraljic, 2009). A common way to demonstrate this flexibility is by using exposure-test paradigms. It has been shown that during an exposure phase listeners use lexical and visual context information to interpret ambiguous phones. This contextual bias, in turn, influences the perception of phonetic categories at a subsequent test phase (for a review, see Samuel & Kraljic, 2009). This effect has become known as perceptual learning in speech or phonetic category recalibration and experiments typically target single phoneme contrasts. Learning in speech perception has also been shown in paradigms in which participants adapt more globally to a new accent rather than just one speech-sound contrast and show benefits in word recognition after adaptation (Bradlow & Bent, 2008; Clarke & Garrett, 2004). While certain parallels in the mechanisms of learning between adaptation to single sound contrasts and global accents have been demonstrated (Mitterer & McQueen, 2009; Reinisch, Weber, & Mitterer, 2013), in the current paper we focus on learning in exposure-test paradigms that target one speech-sound contrast.

One of the issues that has been discussed in the perceptual learning literature is to what extent learning is context-specific or generalizes to other contexts, such as syllable position (Jesse & McQueen, 2011; Mitterer et al., 2013; Nelson & Durvasula, 2021), phonetic context (Reinisch et al., 2014), phonetic features in the target (Mitterer et al., 2016a; Mitterer & Reinisch, 2017), allophones (Mitterer et al., 2013) to name but a few. While this work mostly tested generalization across different types of linguistic factors, Keetels et al. (Keetels et al., 2015, 2016) had shown that perceptual learning in speech may be constrained by the spatial location of the sound source, that is, whether the critical stimuli are perceived as being on the same versus different side of the listener during exposure and test. However, the evidence that we present in the present study suggests that this reported location specificity is not inherent to perceptual learning of speech but instead bound to specific forms of perceptual learning paradigms. As we will show, our findings have potential repercussions on the

[*] Corresponding author at: Holger Mitterer, Department of Cognitive Science, Faculty of Media and Knowledge Sciences, University of Malta, Msida MSD 2080, Malta.
E-mail address: holger.mitterer@um.edu.mt (H. Mitterer).

https://doi.org/10.1016/j.wocn.2024.101313
0095-4470/© 2024 The Authors. Published by Elsevier Ltd.

controversy regarding sublexical units in acoustic-to-lexical mappings (Bowers et al., 2016; Mitterer et al., 2018).

Our investigation starts with the realization that there are radically different forms of exposure-test paradigms, that differ most importantly in the amount of variation in the exposure materials. The design introduced by Norris et al. (2003) exposed participants to 20 different unusually pronounced target words hidden among a large number of fillers (>100). In this version, the bias typically is lexical, that is, based on the Ganong effect (Ganong, 1980), which refers to the finding that participants tend to interpret ambiguous speech sounds in a manner so that the utterance is perceived as a word. That is, an ambiguous sound between /s/ and /f/ is likely to be identified as /f/ in *gira*$[^{\mathrm{s}}/_{\mathrm{f}}]$*e* but as /s/ in *hor*$[^{\mathrm{s}}/_{\mathrm{f}}]$*e*. Norris et al. (2003) as well as studies following this example tended to embed 20 acoustically ambiguous target stimuli in a lexical decision task with 200 stimuli in total, which constitutes the exposure phase. During exposure, there typically is a between-participants manipulation that one group hears an ambiguous sound biased towards an /s/ interpretation and the other group hears an ambiguous sound biased towards an /f/ interpretation. Effects of this exposure are then measured in a test phase in which participants hear the sounds in an unbiased continuum (either nonword-nonword as in [ɛs]-[ɛf], or word-word [nais]-[naif], *nice-knife*) in a two-alternative forced choice task (2AFC). It has typically been found that the group exposed to ambiguous sounds biased towards /f/ (which we will call *amb2f* as a shorthand) provides more /f/ responses on these test continua than the group exposed to ambiguous sounds biased towards /s/. This is the effect of perceptual learning for speech or phonetic recalibration. Importantly, the paradigm typically uses one long, varied exposure phase[1] followed by one test phase, with typically 60 to 150 2AFC trials, although a few studies also used cross-modal priming during the test phase to reduce the potential influence of decision biases (McQueen et al., 2006; Sjerps & McQueen, 2010). Importantly, these latter studies show that the recalibration affects not only perceptual decisions but also spoken-word recognition, two aspects that can dissociate (Krieger-Redwood et al., 2013).

In another paradigm (Bertelson et al., 2003), a similar learning effect was observed at around the same time as Norris et al. (2003), in a design that has minimal variation in the exposure. The biasing information for the disambiguation of ambiguous speech sounds here is provided by visual context via lip-reading (i.e., relying on the McGurk effect, McGurk & MacDonald, 1976) rather than a lexical bias. For instance, participants are presented with an ambiguous speech sound between /b/ and /d/ in an /a_a/ context which does not form an existing word in their native language (i.e., in Dutch in the Bertelson et al. study). At the same time, participants view a video of a speaker either mouthing /aba/ or /ada/ providing the relevant disambiguating information through a visual bias. The exposure phase consists of repeated exposure to one of the exposure stimuli (usually presented eight times), followed by a short test phase with around six 2AFC test trials on ambiguous auditory stimuli from an /aba/-to-/ada/ continuum. Usually, there are between ten to twenty of these short exposure-test cycles, with bias manipulated within participants. Participants show a learning effect by labelling auditory-only stimuli as /b/ more often after exposure to an auditorily ambiguous stimuli biased to be perceived as /b/ than after exposure to an auditorily ambiguous stimuli biased to be perceived as /d/. This shows that exposure to the visually biased exposure stimuli changed the criteria for what is considered to be a /b/ or /d/ in the following test phase.

The vast majority of perceptual learning studies used either a large exposure set with the Ganong effect (lexical bias) to guide the interpretation of ambiguous stimuli, or a small stimulus set with the McGurk effect (visual bias) to guide the interpretation of ambiguous stimuli. However, there are two exceptions to this pattern. On the one hand, Van der Linden and Vroomen (2007) compared visual and lexical exposure biases with small stimulus sets using an experimental design with minimal variation and repeated exposure-test cycles. The study made use of eight Dutch words and nonwords that ended on either /t/ or /p/. The variability within these sets was limited since all words used in the experiment had the same stressed vowel before the word final stop. That is, these stimuli all rhymed. The exposure phase consisted of the presentation of four stimuli that gave rise to the same bias presented two times each, followed by a test phase. Exposure-Test cycles were repeated ten times. The studies revealed that exposure using a visual bias led to stronger learning than exposure using a lexical bias. Moreover, the presence of contrast stimuli (i.e., clear /t/ as a contrast stimulus to an unclear speech sound biased to be perceived as /p/) enhanced learning with both lexical and visual bias during exposure. Visual learning without contrast stimuli was of similar size as lexical learning with contrast stimuli. A comparison of the longevity of the two effects further indicated that learning from both types of exposure (i.e., visual and lexical) quickly dissipated during the test phase after about twenty trials, with no difference between them (see also Ullas et al., 2022).

On the other hand, Reinisch and Mitterer (2016) used an audiovisual bias in a study with a large amount of variation. In that study, the critical sounds (here: /t/ and /p/) were embedded in a variety of phonetically different words, sometimes also as part of a final consonant cluster. They found learning in such a set-up, however, in contrast to the minimal-variability studies in van der Linden and Vroomen (2007), the lexical-bias effect was stronger than the visual-bias effect. This indicates that the amount of variation may change the patterns of perceptual learning and suggests that the learning mechanism may be different depending on the amount of variation in the input.

Differences between experiments using these different paradigms are also found regarding the rate at which the learning effect dissipates. In Van der Linden and Vroomen (2007), it was found that the learning effects quickly dissipate during the test phase so that no learning remained after three blocks of six trials, which would be about a minute of testing. This contrasts with findings that the learning effect with a paradigm similar to Norris et al. (2003) can still be found 12 hours after exposure (Eisner & McQueen, 2006). Van der Linden and Vroomen (2007) pointed out an important confound in this comparison, that is, the presence of test trials. While the experiments using minimal variation during exposure tend to inter-

[1] Some studies used a shorter lexical exposure phase of 32 items and, like the Bertelson et al. (2003) study, had more than one repetition of exposure and test (Myers & Mesite, 2014; Saltzman & Myers, 2021).

sperse exposure and test trials, experiments with a large amount of variation during exposure typically consist of one long exposure followed by a test phase. It is often observed that the learning effect gets smaller over the course of the test phase (e.g., Cummings & Theodore, 2023; Liu & Jaeger, 2018; Tzeng et al., 2021), one possible explanation is that the learning effects may remain stable until the test phase starts. However, a reanalysis of the learning effect over the test phase of the experiment by Mitterer and Reinisch (2013) using a similar design as Norris et al. (2003) showed that the learning effect, though getting smaller over test trials, is still there after eighty test trials and even remained stable after eighty test trials (these new analyses are available in Mitterer & Reinisch, 2022). These findings show that the two paradigms—Mitterer and Reinisch (2013) with a long, varied exposure phase and Van der Linden and Vroomen (2007) with a minimal variation in the exposure set and repeated exposure-test cycles—may give rise to different kinds of learning or tap different aspects of the learning process.

Spatial selectivity has so far only been tested and reported with a paradigm with minimal-input variation using a visual bias. In Keetels et al. (2015), participants heard the same ambiguous sounds over headphones presented to either the left of the right ear during exposure while viewing a video of the speaker whose face was presented on the right or left half of the screen, coinciding with the auditory stimulus' location. During each short exposure block, stimuli were presented on the left and right side with a face presented on the left side of the screen and the (always ambiguous) sound coming from the left side through the headphones. Within one block, the visual stimulus on the right consistently supported an /aba/ percept and the visual stimuli on the left supported a /ada/ percept. In a subsequent short test phase, audio-only stimuli had to be categorized as /aba/ or /ada/ and were presented, over trials, on both the left or right side. Stimuli that were presented on the side that through exposure was associated with /ada/ were more often identified as such than stimuli presented on the side with an /aba/ bias. It could be shown that learning is influenced by location. However, since during each exposure block stimuli were presented in both spatial locations, it could not be shown whether a mismatch in location eliminated or merely reduced the learning effect.

This issue was tackled in a follow-up study in which Keetels et al. (2016) made use of loudspeakers that were on the left and right of the screen. One exposure block always used the same audio-visual stimuli (i.e., either /aba/ or /ada/ in a given block) on just one side, and test blocks presented auditory-only stimuli at the same and other side. In this way, it was possible to test whether there was at least some learning that generalizes over spatial location. Learning was still found at the other spatial location, so that after exposure to an ambiguous sound on the right accompanied by a visual /aba/, an ambiguous auditory-only stimulus on the left side was still more likely to be perceived as /aba/ than after exposure to visual /ada/on the right side. However, the learning effect was less than half as strong as when exposure and test location matched.

This finding dovetails well with findings in experimental psychology that learning often is context- and or location-specific (e.g., famously, Godden & Baddeley, 1975), and this may reflect a need to associate learning with location. For instance, in training a guard dog, one would want the dog to react hostile to strangers in the own home but friendly or neutral to strangers in a dog park. In fact, context-dependency is more a rule rather than an exception in learning (Heald et al., 2023). However, the role of context may depend on how relevant it is for the learning task at hand (Lucke et al., 2013) and in fact on the appraisal of the learner of how important a given context is (Heald et al., 2023). This line of reasoning therefore provides credence to the idea that spatial location may be important when learning is restricted to one nonword stimulus (as in Keetels et al., 2016) but may be less important when the exposure consists of a large set of existing words. It is hence an open question whether perceptual specificity can be observed in paradigm that is similar to that of Norris et al. (2023) with a large amount of variation, many fillers, and a single exposure and single test phase.

This question has further theoretical implications. Research using the the perceptual learning paradigm has been used to delineate the form and type of sub-lexical units in spoken-word recognition (Mitterer et al., 2013, 2016a; Mitterer & Reinisch, 2017; Nelson & Durvasula, 2021; Reinisch et al., 2014; Reinisch & Mitterer, 2016). One auxiliary hypothesis of this line of research is that the perceptual learning paradigm affects sub-lexical units that are used to achieve spoken-word recognition. While there was some evidence to support this assumption (Cutler et al., 2008; Mitterer & Reinisch, 2013; Sjerps & McQueen, 2010), the findings by Keetels and colleagues (2015, 2016) have been used to question this. Bowers et al. (2016) argued that, if perceptual learning is spatially selective, the results from such experiments are unlikely to reflect properties of sublexical units in speech perception, for which it would be counterproductive to be spatially selective. The spatial selectivity of perceptual learning in speech might hence reflect episodic learning outside the realm of spoken-word recognition: Listeners learn that this particular stimulus in that particular situation should be interpreted as, for instance, a labial speech sound. That is, rather than learning something about speech sounds, participants learn a specific stimulus–response pattern. While Bowers et al. (2016) do not settle on this interpretation—they also suggest that perceptual learning may focus on low-level representations that are in fact part of the processing chain in spoken-word recognition, with higher-level representations being phonemic (for further discussion, see Mitterer et al., 2018; Samuel, 2020)—the spatial-selectivity findings lead to lingering doubts about the usefulness of the perceptual-learning paradigm for the investigation of representations used in spoken-word recognition.

It is therefore interesting to investigate how general the finding of spatial selectivity in perceptual learning in speech is. The current experiments therefore test whether spatial selectivity is an inalienable property of the perceptual-learning paradigm or may depend on the specific implementation of exposure and test.

2. Experiment 1

In this experiment, we tested perceptual learning using the /s/-/f/ contrast in Maltese English. Participants first completed a picture-verification task with more than 90 trials, in which they were presented two pictures while listening to a word. Their task was to indicate which of the two pictures better fits the word they heard. Critically, there was a between-participant manipulation in that half of the participants heard words in which an underlying /s/ was replaced by an ambiguous fricative $[^{s}/_{f}]$ and the other half heard the same ambiguous fricative replacing /f/. Moreover, half of the participants heard these words coming from a speaker to the left of the computer screen, the other half heard them coming from a speaker to the right of the computer screen, leading to four groups in total. Test stimuli were then presented on either side. The critical question is whether learning is reduced if there is a mismatch between location of exposure and test stimuli.

2.1. Method[2]

2.1.1. Participants

37 students from the University of Malta participated in the study. This roughly doubles the sample size of the earlier studies (Keetels et al., 2015, 2016), that used 16 to 21 participants per experiment. They were native speakers of Maltese and Maltese English[3] and participated for a small monetary compensation. There were 22 female and 15 male participants, and they were aged between 18 and 29.

2.1.2. Stimuli and apparatus

Experiments were performed in a sound-attenuated booth at the Cognitive-Science lab of the University of Malta. Experiments were run on a standard PC using PsychoPy (version 1.84, Peirce, 2007). Sounds were presented using Logitech Z 150 speakers located on the right and left of a 22-inch monitor. The two speakers were about 70 centimeters apart, and the listeners had a viewing distance of about 60 cm to the screen, leading to a spatial separation of the two speakers of 60 degrees.

For the exposure phase, 22 words each containing /s/ and /f/ in various positions (see Appendix) were selected. Note that learning for fricatives generalizes over positions (Jesse & McQueen, 2011) and having no constraints on syllable positions allowed us to use more natural pairings of images and words. Additionally, 49 filler words were selected that contained neither /s/ or /f/ nor their voiced counterparts /z/ and /v/. These words plus three minimal pairs for the test phase (*knife-nice*, *rice-rife*, *lice-life*) were recorded by a male speaker of Maltese English. Additionally, the critical words were recorded with the “other” fricative (that is, *giraffe* was also recorded as *girasse* and *police* was recorded as *polife*). Based on these pairs, ambiguous tokens were generated. First of all, the fricatives were cut out from the recordings. The two remaining part words, that is the parts with the fricatives removed (e.g., *poli*... with formant transitions appropriate for either /s/ or/f/) were mixed using STRAIGHT (Kawahara et al., 1999) to minimize formant-transition cues to the place of articulation of the fricatives. For the fricative parts of the words, six continua (based on syllable position[4] and rounding of the vowel context, see Appendix) were generated. These fricative continua were then spliced back into the mixed part words containing no clear formant transition cues. All STRAIGHT resynthesis were carefully checked to not contain artefacts due to creaky voice (see McQueen et al., 2023, for the importance of such checks). These cross-spliced tokens of part word and fricative were then assessed by six native speakers of Maltese English to find the most ambiguous tokens. Based on these judgements, one ambiguous token was selected for the main experiment. Additionally, for each word, a matching picture and an unrelated distractor picture were selected by a Google image search. These pictures were used for the picture-matching task. Three additional pictures related to existing items were selected for a short old/new picture task that introduced a short delay between exposure and test of about 2 minutes.

For the test phase, the STRAIGHT algorithm (Kawahara et al., 1999) was used to generate an eleven-step continuum between the natural utterances in steps of 10% for the three minimal pairs (*knife-nice*, *rice-rife*, *lice-life*). Continua were again informally tested for their ambiguous range and levels 4 to 8 were used for the test phase.

2.1.3. Procedure

Experiments started with the instruction for the exposure phase. All instructions were presented as text on the computer screen. Participants were instructed that they would see two pictures and hear a word. They would have to indicate via a keyboard-button press which of the two pictures better fitted the word, using the right and left arrow key.

Based on the order of testing, participants were assigned to either the /s/-biased exposure (i.e., ambiguous fricatives in words with /s/, e.g., *hor*$[^{s}/_{f}]$*e*, and unambiguous [f] in word with underlying /f/, short: amb2s for ambiguous sound to s for participants with an odd number) or /f/-biased exposure (i.e., ambiguous fricatives in words with /f/, e.g., *gira*$[^{s}/_{f}]$*e*, and unambiguous [s] in word with underlying /s/, short: amb2f for participants with an even number). Moreover, the first two participants in a set of 4 were presented exposure stimuli from the right speaker, the other two to stimuli from the left speaker.

Trials during exposure had the following structure. One picture each appeared on the right and one on left side of the screen. After 400 ms, an auditory stimulus was played via one of the speakers. Participants were instructed to press the left or right arrow key, depending on which picture better matched the sound. Each exposure phase started with five practice trials followed by a mix of the 44 critical trials and 44 fillers. Order of presentation was randomized individually, with the constraint that no two tokens with an ambiguous fricative were presented in direct succession.

After the exposure, participants received written instructions to an old/new picture recognition task. During this task they saw one picture per trial and had to indicate with the arrow keys whether the picture had been seen before during the

[2] All materials, data, and analyses files are available here: https://osf.io/6sx5y/.

[3] Malta is officially bilingual (Maltese and English), but the majority of the speakers are more confident in Maltese than in English. Note, however, that the /s/-/f/ contrast occurs in Maltese and English. Moreover, courses at the University of Malta are taught in English, so that students need to have a good command of English. This, in turn, means that the perceptual learning of this contrast should not be a problem see e.g., (Reinisch et al., 2013, for recalibration of sounds in a non-native language that are shared with the native language).

[4] Note that syllable position is not critical for learning about fricatives (Jesse & McQueen, 2011), as long as the biasing cue, in our case, the picture, is available when the fricative is heard.

experiment or not. Six pictures were presented, three of which were new.

After this short picture-recognition task, the test phase started. It consisted of a two-alternative forced choice (2AFC) task using the minimal pairs *knife-nice*, *rife-rice*, *lice-life*. Instructions stated that participants would hear a word that might be difficult to identify. They would see the two words of a minimal pair written on the screen and decide which of the two words better matched the sound. Each of the 15 stimuli (three continua times 5 steps) was presented at both locations (left/right) totaling in 30 different stimuli during a test block. This means that, even though learning condition is manipulated between participants, location match is manipulated within participant (see Table 1). Each participant completed five of those blocks without a break between blocks. Order of presentation was randomized within each block.

2.2. Results and discussion

During the exposure phase, participants nearly always chose the intended picture (about 99% correct responses). We tested whether all participants chose the "correct" pictures even for the ambiguous items as intended (>85%), since earlier research has shown that rejection of ambiguous items lowers the amount of learning (Sjerps and Reinisch, 2015). This was the case, and all participants were retained for the analyses of the test phase. Fig. 1 presents the results of the test phase as the proportion of trials on which participants responded with /s/, depending on exposure and location match. Exposure here refers to the bias induced during exposure (amb2s vs. amb2f) and Location match refers to whether a test stimulus during the test phase was presented on the same side as the exposure stimuli (note that location of exposure stimuli was varied between participants, but the location of test stimuli varied within participants).

Fig. 1 shows that there is a clear difference between the exposure conditions, but no clear effect of having a location match between exposure and test sounds. These observations were borne out by a generalized linear mixed effect model with a logistic linking function in R (v 4.1.3, R Core Team, 2022) using the lmerTest package (Kuznetsova et al., 2015). The dependent variable was the likelihood of an /s/-response, and the predictors were contrast coded (Exposure Condition: amb2f = -0.5, amb2s = 0.5, Location Match: 0.5 = same, −0.5 = opposite, continuum step ranging from −2 to 2). The model specified an interaction of Exposure Condition and Location Match. Step was only used as a control variable and was not allowed to interact with the other factors. The random-effects structure included a random effect for Participants with random slopes for Continuum Step and Location Match. This is maximally converging random-effect structure with correlations between random slopes restricted to zero.[5]

With this coding, all expected regression weights should be positive. For instance, the learning effect (i.e., effect of Exposure Condition) should be reflected in a positive regression weight, since the likelihood of an /s/-response is expected to be higher in the amb2s group than in the amb2f group. The results presented in Table 2 show that this is the case. The critical question is whether this learning effect is larger when there is a match in location between exposure and test phase. This issue is tested by the interaction between Exposure Condition and Location Match, which is not significant. In fact, the estimate is opposite to the expected direction for such an enhancement.

The absence of a modulation of the learning effect by spatial location was further investigated with a Bayesian analysis. To this end, a "Location Match" effect was calculated for each participant. This effect was the difference in logOdds of /s/-responses for trials in which the location of the sound at test matched the exposure side and for trials in which the location mismatched the exposure side. This difference was calculated in such a way that the effect of location on perceptual learning should be reflected in a positive value and therefore depends on the exposure condition. For the amb2s group, the logOdds of /s/-responses with non-matching locations were subtracted from the logOdds of /s/-responses for matching locations, reflecting the expectation that the bias toward /s/ should be larger when locations match. That is, if there is spatial selectivity of learning (with more learning at the location used during exposure), this subtraction should lead to a positive number. For the amb2f group, the expectation is that there should be fewer /s/-responses if the locations of exposure and test match, because the exposure induces an /f/-bias. Therefore, the logOdds of /s/ responses during test at the same side of exposure was subtracted from the logOdds of /s/ responses at the opposite side, again leading to the expectation of a positive number if there is spatial selectivity of learning. This matching effect can be related to the learning effect when exposure and test stimuli were from the same location, which is 0.732 logOdds units. This allows us to estimate a prior for the effect of matching sides, since Keetels et al. (2016) reported a reduction of 59% of the original learning effect, so that we should expect a reduction of the learning effect by 0.432 logit units (0.732 logit units * 59% reduction) between the location match and mismatch conditions. In the case of such a clear prior, Dienes (2014) suggests an alternative hypothesis with a normal likelihood distribution with a mean of the expected effect size (0.432 logit units) and a standard deviation of half that mean, so that negative outcomes are considered unlikely. We used these priors with the Bayes calculator provided by Dienes in its R implementation (Baguley & Kaye, 2010). The observed effect of location match was −0.001 logit units (SE = 0.214), which leads to a BF of 0.245. This indicates that the null hypothesis (no difference depending on location match) is about four times more likely than the alternative hypothesis (there is a difference depending on location match roughly of the same size as observed by Keetels et al., 2016).

In summary, the current experiment gave rise to a clear perceptual learning effect such that participants in the amb2s group interpreted more sounds at test as /s/ relative to participants in the amb2f group. This effect was not modulated by the manipulation of spatial location, that is match versus mismatch in spatial location of the stimuli between exposure and test. Importantly, the Bayes Factor is in the range [1/10, 1/3] which is considered substantial evidence (here: for the null hypothe-

[5] The formula in R was: choice.lmer = glmer(resp ∼ Condition *LocationMatch + step + (step+LocationMatch||participant), data = test, family = "binomial"). Note that a random slope for Exposure Condition over participants would not be meaningful since it is manipulated between participants.

Table 1
The allocation of participants to condition, showing how learning condition is manipulated between participants but critically, location match is varied within participants.

Participant ID number	Exposure type	Exposure side	Presentation at test	
			right	left
1,5,7,...	amb2s	right	+match	-match
2,6,10,...	amb2f	right	+match	-match
3,7,11,...	amb2s	left	-match	+match
4,8,12,...	amb2f	left	-match	+match

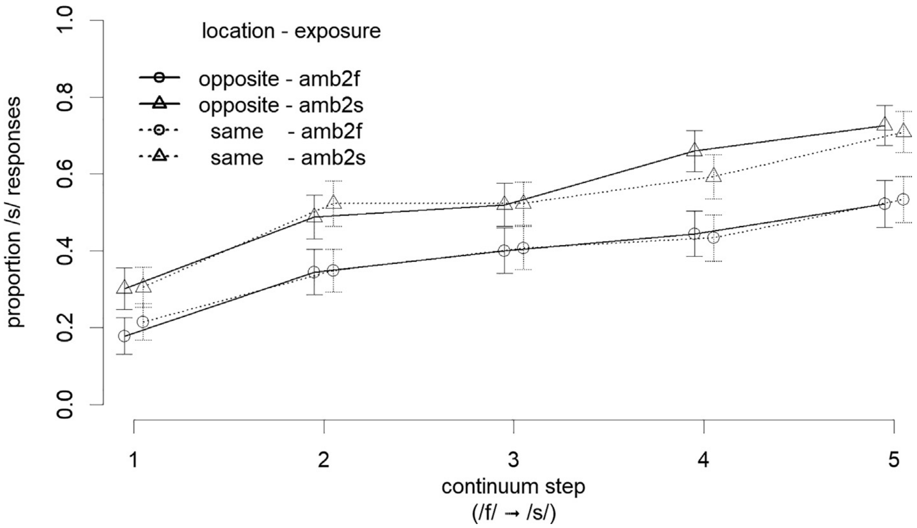

Fig. 1. Mean proportion of /s/ responses for the two exposure conditions over the continuum in Experiment 1. Point symbols indicate the exposure condition (circles = amb2f, triangles = amb2s). Location match is coded by line type (dashed lines= "same location in exposure and test", solid lines = "opposite location in exposure and test") and with dodged positions to increase readability. Error bars show the confidence interval estimated through the function summarySEwithin from the R package Rmisc (Hope, 2012/2014).

Table 2
Results from the generalized linear mixed-effects model for the likelihood of /s/-responses during the test phase of Experiment 1.

	Estimate (SE)	z	p
Intercept	−0.213 (0.135)	−1.569	0.117
Exposure Condition	0.799 (0.271)	2.949	0.003
Location Match	−0.019 (0.227)	−0.086	0.932
Step	0.453 (0.031)	14.43	<0.001
Exposure Condition * Location Match	−0.042 (0.454)	−0.092	0.927

sis) though not strong evidence (Dienes, 2014; Jeffreys, 1961). Therefore, we set out to replicate the results of Experiment 1 with a different set of materials that already have been shown to give rise to a large learning effect. With an even larger learning effect, it should be easier to find a modulation of the learning effect.

3. Experiment 2

This experiment made use of the Dutch materials used in Mitterer and Reinisch (2013). It was run in a web-based setting, which makes it easier to test a larger number of participants. Since we expected participants in a web experiment to wear headphones, we made use of virtual spatial locations by varying the timing and amplitude of the left and right channels (following the parameters in Palomäki et al., 2005). To check that participants were able to differentiate these virtual spatial locations, we tested this prior to the main experiment.

3.1. Method

3.1.1. Participants

Eighty-three participants, all native speakers of Dutch, were recruited via the Prolfic.co site. They were required to be between 18 and 40 years old. The average age of those who participated was 28.8 (sd = 5.6). 56 declared their gender as female, 23 as male. Constraints for participation were that they were aged between 18 and 40 years and Dutch was one of their native languages. Moreover, within Prolific, the experiment was made accessible only from a desktop computer, not from a tablet or mobile phone.

3.1.2. Stimuli

The stimuli were taken from the study by Mitterer and Reinisch (2013), where strong learning effects of more than two logit units difference (which is around 40% difference around a 50% baseline) had been found with these materials in lab-based studies. Materials (available at: https://osf.io/v2unz/) compromised 200 Dutch stimuli for the exposure phase which made use of a lexical-decision task. There were 20 critical stimuli that ended in /s/ and 20 stimuli that ended in /f/. Whether the /s/- or /f/-final words carried an ambiguous fricative was manipulated between participants. The ambiguous fricatives were based on morphs, generated with the STRAIGHT algorithm (Kawahara & Irino, 2005) between the word and a nonword utterance with the other fricative (e.g., for the word *radijs,* Engl,. *radish*, the nonword *radijf*) and the ambiguous step was based on a pretest (see Mitterer &

Reinisch, 2013). Additionally, there were 60 words and 100 nonwords that contained neither /s/ or /f/. For the test phase, stimuli were based on four minimal pairs (*doos-doof*, Engl. 'box'-'deaf', *kuis-kuif* Engl. 'chaste'- 'crest', *les-lef*, Engl. 'lesson'-'courage', *roos-roof*, Engl. 'rose'-'robbery'). For each minimal pair, four stimuli from the ambiguous range of a morph-series with 10% steps had been selected in Mitterer and Reinisch (2013) and we used the same steps here.

For the present study, for each of the stimuli used in Mitterer and Reinisch (2013), we generated a version that, over headphones, would appear to come from the front left or right (i.e., ±45°) using the parameters for amplitude and time delay as reported in Palomäki et al., (2005, ITD = 0.39 ms, ILD = 11.2 7 dB). We preferred using this form over spatialization above panning fully to the left and right channel, because the latter is an ecologically invalid situation that only arises with headphones which were not available throughout most of human evolution.

3.1.3. Procedure

The procedure was similar to Experiment 1, with an exposure phase and a test phase, but there were two differences. First of all, we tested whether participants were able to identify the virtual locations, which might be difficult if they ignored the instructions to wear headphones. Second, the exposure task was a lexical decision task rather than a picture verification task. For the localization test, ten filler words from the exposure phase were used and presented from left or right virtual locations. Participants had to indicate where the sound was coming from and were provided with feedback on whether their choice was correct or not. Note that some participants may wear their headphones the wrong way round, and therefore consistently be wrong (see Results for details). Each word was presented twice, leading to 20 trials. The experiment started with this localization task, followed by the exposure phase (200 trials) followed by 160 test trials, in which the participants decided whether the words ended on /s/ or /f/ through a 2AFC task using whole-word prompts. That is, the participants decided which word they heard rather than on which fricative it ended (see Krieger-Redwood et al., 2013, for why the former task may be more natural). To dissociate "left" and "right" from the response options, participants were instructed to use the up- and down arrow keys for exposure (up = word, down = nonword) and test (up = /s/, down = /f/).

3.2. Results and discussion

For the analysis of the localization task in which participants had to indicate the virtual location of the sounds, we focused on the last 15 trials, since some participants were consistently wrong at the beginning, indicating that they were wearing their headphones the wrong way round. Therefore, we disregarded the first five trials and retained participants if they had at least 14 out of 15 trials correct on the remaining trials. This led to the rejection of six participants. For the remaining participants we tested whether they accepted more than 17 of the critical items as real words in the lexical decision task, because participants who reject many of the ambiguous items tend to show a lower or even no learning effect (Sjerps & Reinisch, 2015). This led to the rejection of one additional participant, so that the final sample contained data from 76 participants.

Fig. 2 shows the results from the test phase. As the figure shows, participants in the amb2s exposure condition (who had heard the ambiguous sound in words where it replaced /s/) gave more /s/ responses at test than participants in the amb2f exposure condition. The perceptual learning effect was larger when tested on the same side as exposure for the first two steps—as shown by the larger separation between the dotted and the solid lines at steps one and two. However, for steps three and four, the perceptual-learning effect (i.e., the separation of the amb2s and amb2f lines) was similar for the same-location and opposite-location conditions.

The statistical analysis followed the same pattern as in Experiment 1, with contrast-coded predictors for Step, Exposure Condition, and Location Match. The outcome (see Table 3) showed that we replicate the learning effect observed in Mitterer and Reinisch (2013), but no significant interaction between Location Match and Exposure Condition could be found. As in Experiment 1, we performed an additional Bayesian analysis to see whether the data are more supportive of an effect that is similar to the one reported by Keetels et al. (2016) or more likely under a null hypothesis. We calculated a perceptual learning effect over participants of 0.891 logit units, which predicts an effect size for the reduction by spatial location of 0.526 logit units. The actual reduction effect (calculated in the same way as in Experiment 1) is 0.049 (SE = 0.044), leading to a Bayes factor of 0.062. This means that the null hypothesis is fifteen times more likely than the alternative hypothesis.

An additional detail that is worth mentioning at this stage is that the perceptual learning effects here were less than half as strong as in the lab-based study of Mitterer and Reinisch (2013). While many studies (Gould et al., 2015; Kim et al., 2019; Reinisch & Penney, 2019) reported good data quality for web-based experiments, we find a strongly reduced perceptual learning effect in the web-based setting compared to the laboratory results obtained by Mitterer and Reinisch (2013). Following Cooke and García Lecumberri (2021), we tested whether such effects are related to the self-estimated quality of the headphones (see the online repository from footnote 1 for this analysis). This was not the case. Moreover, another aspect of our data supports the validity of web-based data acquisition. The effect of Continuum Step (i.e., the steepness of the identification function) is quite similar in the present study compared to Mitterer and Reinisch (2013; i.e., around 1.3 logit units per step). If participants were doing the task less diligently online, we would expect them to press buttons randomly more often, which would lead to shallower identification functions. The lack of a difference in the steepness of the identification functions makes it, in turn, more surprising that the learning effect is clearly smaller in the web-based setting. Note, however, that for perceptual learning, participants need to encode the "unusualness" of the ambiguous stimuli during the whole exposure phase. This suggests that for single-trial responses, the quality of web-based data may match that of lab-based settings, but effects that rely on partic-

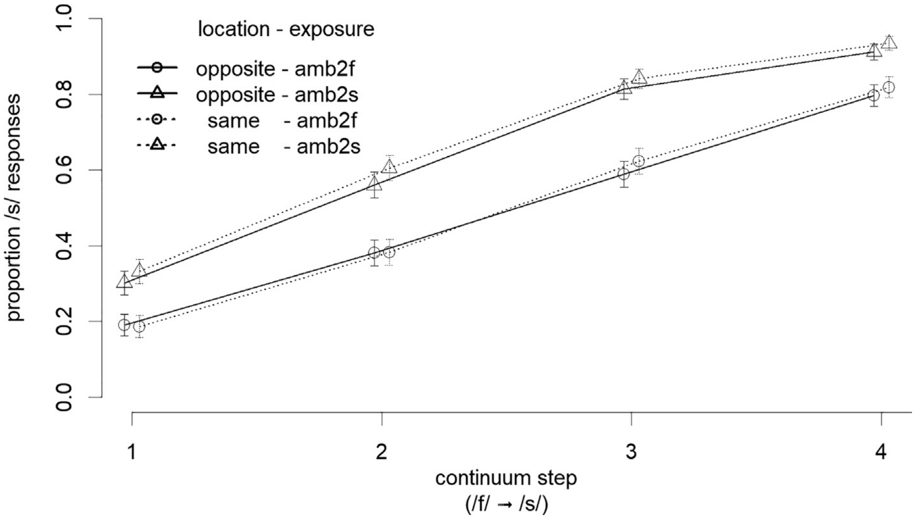

Fig. 2. Mean proportion of /s/ responses for the two exposure conditions over the continuum in Experiment 2. Point symbols indicate the exposure condition (circles = amb2f, triangles = amb2s). Location match is coded by line type (dashed lines= "same location in exposure and test", solid lines = "opposite location in exposure and test") and with dodged positions to increase readability. Error bars show the confidence interval estimated through the function summarySEwithin from the R package Rmisc (Hope, 2012/2014).

Table 3
Results from the generalized linear mixed-effects model for the likelihood of /s/-responses during the test phase of Experiment 2.

	Estimate (SE)	z	p
Intercept	−2.765 (0.17)	−16.261	<0.001
Exposure Condition	0.856 (0.334)	2.563	0.01
Location Match	0.154 (0.059)	2.591	0.01
Step	1.346 (0.053)	25.352	<0.001
Exposure Condition * Location Match	0.145 (0.119)	1.225	0.221

ipants encoding the study's history may have to deal with smaller effect sizes in an online setting. This does not mean that such effects cannot be observed, after all, a perceptual-learning effect was found here and by others (Papoutsi, Zimianiti, Bosker, & Frost, 2023; Tzeng, Nygaard, & Theodore, 2021) but our data suggest that they may be weaker in an online setting.

At this juncture, our data indicate that the strong reduction in learning in speech perception due to spatial location does not necessarily occur in a paradigm using a lexical bias with a single exposure phase with a large amount of variation and fillers. However, this design differs in many factors from that of Keetels et al. (2016). Therefore, it is difficult to ascribe the differences in results to one factor. The two most salient factors are the use of the visual modality and the amount of variation within the exposure set. The use of the visual modality may be important because location is processed with greater precision in vision than in audition. Therefore, in Experiment 3, we make use a stimulus set in which learning is induced by lip-reading, but based on a stimulus set of more than 100 stimuli (Reinisch & Mitterer, 2016) and without repetition of ambiguous training stimuli as in Keetels et al. (2016). Visual stimuli were presented either on the right or left side of the screen and the audio was fully panned to the right or left speaker (as in Keetels et al., 2015)—despite the previously discussed low ecological validity of this stimulus set-up—to maximize the possibility to find an effect of spatial location on perceptual learning.

4. Experiment 3

4.1. Method

4.1.1. Participants

We aimed at recruiting 60 participants with valid data. Acquiring 60 data sets that passed data-quality screening required 62 participants, since two participants failed initial data quality checks (see Procedure for details). Participants were recruited from the Prolific.co platform with the constraints that their native language was German and that they were aged between 18 and 40. After an additional data check, one of the initially rejected participants could be included after all (see below for details), so that the final sample contained 61 participants, the mean and median age were 30, with a range from 20 to 40. Nine were female and 52 were male. Moreover, within Prolific, the experiment was only accessible from a desktop computer, not from a tablet or phone.

4.1.2. Stimuli

Stimuli were taken from the study by Reinisch and Mitterer (2016), which focused on the potential recalibration of the stop consonants /t/ and /p/ in German. They video-taped a female native speaker of German uttering 77 nonwords and 121 words. Of these 121 words, 77 were filler words that, as the nonwords, did not contain the critical stops nor their voiced counterparts /b/ and /d/. The remaining 44 words were 22 minimal pairs with the stops in word-final position (e.g., /ɑlt/ - /ɑlp/, Engl, 'old'-'alp'). For these, the audio tracks were extracted from the video recordings and continua were generated using STRAIGHT (Kawahara & Irino, 2005). An ambiguous step was found through a pretest (see Reinisch & Mitterer, 2016) and used to replace the original sounds in the videos. This procedure leads to 44 videos each with an ambiguous audio track that is disambiguated by the visual speech gestures towards an alveolar or labial stop percept (i.e., /t/ vs. /p/ respectively).

Moreover, four videos containing filler words and nonwords were additionally edited to contain a green dot that appeared between the upper and lower lip of the speaker while she was uttering the word. These were used for an attention check (see below). For the current project, we used all these videos and generated two versions with the audio panned fully to the right or left channel. For the test phase, we used the same continuum from [ʔap] to [ʔat] as Reinisch and Mitterer (2016), but used different steps (i.e., steps 3, 6, 8, 10, and 12, of the 20-step continuum rather than 0, 2, 4, 6, 8, and 10). This was because the original data showed a small /p/-bias and, moreover, we aimed for a slightly wider range of the continuum for web-based testing, to make sure that most participants perceive a difference between the steps along the continuum (at least the presented endpoints).

4.1.3. Procedure

Participants first performed the same headphone test as in Experiment 2. Then they were provided with the instruction for the Exposure phase, which explained that they would see a speaker uttering a word and their task was to press the "arrow-up" button the word was an existing word in German but the "arrow-down" button if it was not. Moreover, they were told that, sometimes, a green dot would appear on the videos, and if they see the dot, their task was to press the space bar.

The exposure consisted of 144 trials in total of which 22 trials were critical exposure trials. Half of these trials presented an ambiguous audio track but a clear visible speech gesture. The other half of these trials presented "congruent" AV stimuli with the audio using a continuum endpoint presented with the congruent visual speech gesture (i.e., [p] audio presented with a labial closing gesture). Without such contrast stimuli, participants might simply learn that the speaker does not produce clear stops rather than learning that either the alveolar or labial stop only is produced in a somewhat unclear fashion.

Within these critical exposure trials, the main exposure condition and exposure side was implemented between-participants, so that half of the participants were presented with clear [p] and ambiguous sounds biased towards [t] (i.e., ambiguous = [t], "amb2t") while the other half were presented with the opposite (i.e., ambiguous = [p], "amb2p"). These 22 exposure trials were embedded within 132 filler trials (77 nonwords and 55 words), eight of which contained a green dot appearing on the speaker's face. For the online experiment, forty different random orders of these 154 trials were generated, twenty each with a /p/-bias and twenty with a /t/-bias, half of which had the exposure stimuli presented on the left or right.

The test phase was audio-only and presented the stimuli for all participants through either the right or left headphone speaker (on different trials). The five steps of the continuum were presented seven times each on the right and left side. While this is a relatively short test phase, the earlier study (Reinisch & Mitterer, 2016) had shown that the perceptual-learning effect dissipates more quickly with audio-visual than lexical disambiguation of the critical stimuli.

The experiments were controlled with jsPsych (de Leeuw, 2015). After the headphone check, the instruction of the exposure phase presented one example of the catch trials with the green dot twice, as to fully alert participants about the attention check. For the exposure phase, the "video-keyboard-response" plugin of jsPsych was used as a template. This template was slightly changed so that the screen contained a table that spanned the whole screen, and the video was aligned to the left or right position using the "float: left/right" style attribute of the video. The altered plugin is available in the online repository. On the bottom of the screen, the instructions were repeated (arrow-up" = word, "arrow-down" = nonword).

The test phase was using the orthographic transcriptions < aap > and < aat> (both nonwords in German) as prompts stacked on top of each other, and (as in Experiment 2) participants responded with an arrow-up or -down key.

The experiment was initially opened for 50 participants. After 50 participants had participated, it was checked how many valid data sets were available in each of the four possible combinations of exposure and exposure side. Additional checks were made after 56 and 60 participants to balance the number of exposure conditions. This was achieved by slightly changing the script so that combinations that already were completed by 15 participants were made unavailable in the randomization procedure. Of the 62 participants who participated, 61 could be retained.[6] Thirty were exposed to an amb2t exposure list and thirty-one to amb2p exposure list.

4.2. Results and discussion

The data from one participant was excluded from further analysis because all catch-trials were responded to as lexical-decision trials, indicating no or at least limited attention to the visual stimuli. For the remaining 61 participants, at least six out of eight catch trials (mean correct: 91.3%) were responded to correctly and lexical decision performance was 93.7% correct overall with the lowest accuracy of a participant at 78.8%. Note that lexical-decision on critical trials cannot be used to test whether participants perceived the stop as intended, because both /ɑlp/ and /ɑlt/ lead to a "yes-response".

Fig. 3 shows the performance during the test trials. The results show a perceptual learning effect with more /t/ responses for the amb2t group than the amb2p group (i.e., triangles above squares in Fig. 3). This effect is not obviously altered by the match in spatial location between exposure and test. While there is a larger separation (i.e., a larger learning effect) with matching spatial location (i.e., a larger separation for the dashed than the solid lines) for the second continuum step, no such difference or a slight difference in the opposite direction is observed at the other continuum steps. This is also evident in the statistical analysis with a generalized linear mixed-effects model using a binomial linking function, contrast-coded predictors, and the maximally converging random-effects structure (which contained only a slope for match). This analysis (reported in Table 4) revealed a significant learning effect, but no moderation of this effect by a match in spatial location between exposure and test.

[6] During data collection, we performed a semi-automated fast data-quality check to determine whether a data set was usable (i.e., most catch trials were responded to correctly). Two data sets failed the initial check, leading us to collect data from 62 participants. Upon checking after data collection, it turned out that one participant failed the semi-automatic check, because, just for this participants, *true* and *false* in the output were written in capitals but the semi-automatic data check expected the critical variable value to be case-sensitive. Therefore, the semi-automatic data check wrongly marked this participant as inattentive, even though they answered correctly on all catch trials. This is why we re-entered this participant's data to the final dataset, leading to a total of 61 participants.

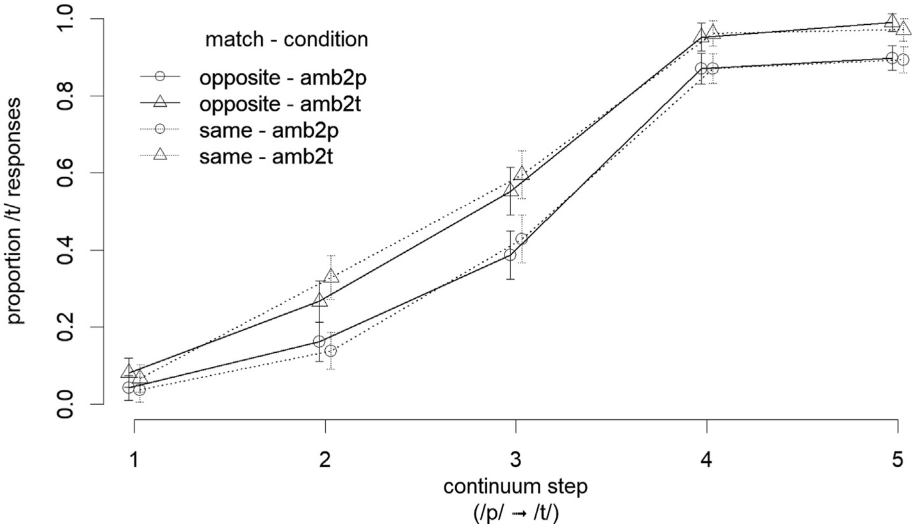

Fig. 3. Mean proportion of /t/ responses for the two exposure conditions over the continuum in Experiment 2. P Point symbols indicate the exposure condition (circles = amb2p, triangles = amb2t). Location match is coded by line type (dashed lines= "same location in exposure and test", solid lines = "opposite location in exposure and test") and with dodged positions to increase readability. Error bars show the confidence interval estimated through the function summarySEwithin from the R package Rmisc (Hope, 2012/2014).

Table 4
Results from the generalized linear mixed-effects model for the likelihood of /s/-responses during the test phase of Experiment 2.

	Estimate (SE)	z	p
Intercept	0.192 (0.229)	0.841	0.401
Exposure Condition	1.128 (0.459)	2.46	0.014
Location Match	0.094 (0.153)	0.614	0.539
Step	8.62 (0.27)	31.91	<0.001
Exposure Condition * Location Match	0.173 (0.307)	0.563	0.573

As in the previous experiments, we also calculated a Bayes Factor. The alternative hypothesis was based on the results of Keetels et al. (2016) with a 59% reduction of learning; and this hypothesis was compared to a null hypothesis. As before, we calculated for each participant how their identification of the test stimuli differed in the location match and mismatch conditions, taking into account the expected directionality (more /s/ responses for the amb2s group in the match condition, but fewer /s/ responses for the amb2f group in the match condition, that is, a larger learning effect in the match condition). Based on the mean and standard error of this measure, the resulting Bayes Factor is 0.137, representing substantial evidence for the null hypothesis.

Finally, a reviewer suggested that the discrepancy between our data and that of Keetels et al. (2016) is due to the fact that we did not include a fixation point in the visual display, so that listeners would look at the speaker and thereby annihilate any spatial disparity within a retinotopic coding. While this is a possible explanation, we deem this unlikely. According to the Procedure described in Keetels et al. (2016), participants in that study were neither instructed to keep looking at the fixation point nor were any data reported that they did so. It is therefore unlikely that participants kept looking at the fixation point, given the well-documented effects of visual movement on visual attention (for a review, see van der Heijden, 1992), which would lead to the automatic response to look at the moving face. This makes it unlikely that the presence of a fixation point in the display is a likely explanation for the different results.

Note, however, that the preceding BF analyses that this does not mean that the current data make it unlikely that there is *any* reduction due to spatial location. To assess the possibility for *some* reduction of the learning effect due to position mismatch, we used weaker versions of an alternative hypothesis, in which the assumed reduction of learning by a mismatch in location is assumed to be less than the nearly 60% observed in Keetels et al. (2016). Therefore, we calculated the BF for a reduction of 50% to 10% in steps of 10%, first for each experiment individually and then for the whole data set. Given such "point" hypothesis (i.e., the reduction is 10%, 20%, etc.), the cumulative BF can be found as the product of the individual BFs (see Dienes, 2014, supplementary materials, note 4 on meta-analysis). As Table 5 shows, the analysis of 10% steps suggests that even smaller amounts of reduction are unlikely given the data. Only for the smallest assumed reduction of 10% is the BF slightly favoring the alternative hypothesis. This suggests that it is unlikely that perceptual learning is strongly constrained by spatial location when tested with a paradigm that uses a long exposure phase with variation and many fillers.

5. General discussion

The purpose of this study was to evaluate the scope of location specificity in perceptual learning in speech. The results indicate that spatial selectivity is likely not a general property of perceptual learning in speech. Three experiments with an exposure phase containing many different stimuli failed to find any evidence for spatial selectivity. As for the amount of spatial separation, Experiment 1 used a separation of about 60 degrees in a lab-based setting, Experiment 2 used a virtual separation of 90 degrees, and Experiment 3 presented the sounds on the right and left speaker of headphones, hence used a full separation. A Bayesian analyses showed that the results are more supportive of a null effect than of the effect observed in Keetels et al. (2016). The combined evidence of these studies indicates that spatial selectivity is not an inalien-

Table 5
Bayes Factors comparing the null hypothesis versus different alternative hypothesis in terms of reduction of learning due to spatial location.

Reduction by	BF (Exp1)	BF (Exp2)	BF (Exp3)	Cumulative BF	Interpretation
50%	0.306	0.078	0.160	0.003	Very Strong for H0
40%	0.411	0.109	0.232	0.048	Strong for H0
30%	0.558	0.171	0.384	0.037	Strong for H0
20%	0.742	0.353	0.775	0.203	Substantial for H0
10%	0.917	1.062	1.711	1.667	Barely worth mentioning for H1

Note: Interpretations are based on Jeffreys (1961).

able property of perceptual learning in speech. Instead, spatial selectivity of perceptual learning has only been observed when tested in a paradigm with minimal variation and repeated exposure-test cycles (Keetels et al., 2015, 2016) but not in paradigms with a large amount of variation during exposure as were used in the present study.

The difference in sensitivity to spatial location adds to the differences between these experimental paradigms that were introduced in the introduction. Experiments using one long and varied exposure phase tend to provide different results than those using minimal-variation and often multiple exposure-test cycles. The former are more effective with lexical biases than visual biases (Reinisch & Mitterer, 2016) while the reverse is observed with the latter (van der Linden & Vroomen, 2007). The former paradigm with variation during exposure also leads to learning effects that are more resistant to testing than the latter (see the new analyses of the data of Mitterer & Reinisch, 2013, in Mitterer & Reinisch, 2022). The two paradigms are similar in that they manage to change the perception of an otherwise ambiguous stimulus based on an exposure phase (Kleinschmidt & Jaeger, 2015). Despite this superficial similarity, it is conceivable that they rely on different mechanisms. After all, both, showing an image of a male or female speaker (Johnson et al., 1999) and presenting a rounded versus unrounded vowel (Mitterer, 2006) shift the perceived/reported category boundary between /s/ and /ʃ/ but few would argue that those two effects rely on the same mechanism, even though they lead to the same effect. Moreover, to our knowledge, there have been no studies that critically appraised whether the two perceptual learning paradigms discussed here really reflect a similar mechanism. For instance, with the paradigm using a single, long, and varied exposure phase, it has been found that learning generalizes over syllable position and vowel context (Jesse & McQueen, 2011; Nelson & Durvasula, 2021), generalizes strongly to a new set of words (Mitterer et al., 2011) and generalizes when the perceived speaker is different (Eisner & McQueen, 2005; Reinisch & Holt, 2014). To our knowledge, none of these generalizations have been shown in the paradigm with minimal variation and repeated exposure-text cycles, and we predict that few and possibly none of them can be observed.

Regarding the issue of spatial selectivity of perceptual learning, there are at least two possible conceptualizations to best characterize the difference in outcomes between the two learning paradigms in previous studies and the present experiments; that is, paradigms with relatively generalizable learning that have a single, long, and varied exposure phase versus highly context-specific learning in the paradigm with minimal variation and repeated exposure-test cycles. Firstly, Heald et al. (2023) argue that whether context, or a specific aspect of the context, plays a role in learning or not depends on how important a given context is to the perceiver. Clearly, with an impoverished stimulus set, perceivers may weigh non-linguistic context as relatively important but when there is a large set of words, spatial location becomes an unimportant feature for the perceiver. Moreover, the fact that spatial location varied during exposure in the studies by Keetels et al., (2015,2016) but not in the current studies might have highlighted the importance of spatial location.[7] Note that under this account, the learning mechanism for both paradigms would be the same; the only difference would be the salience of the spatial location given the amount of variation in the linguistic input.

A second account refers to another dissociation in learning that has been observed by Gaskell and Dumay (2003). They found fast learning of new words and slow integration of that learning into the linguistic system. A similar dissociation may be possible with regard to perceptual learning, where repeated exposure to minimal variation leads to episodic learning while exposure to different examples embedded in large set of fillers may lead to more generalized perceptual learning. In this conceptualization, the different types of paradigms may give rise to different types of learning.

While our data is compatible with both accounts, one previous finding makes the latter explanation more likely. Perceptual learning has often been categorized as speaker-specific, but this overlooks a nuance in the actual findings (Creel et al., 2008; Kleinschmidt & Jaeger, 2015). Eisner and McQueen (2005, Exp 3) found that perceptual learning did not necessarily generalize from one speaker to another. This could easily be interpreted in the framework of context-dependent encoding, so that listeners take the context of a given speaker as crucial in applying the perceptual learning. This explanation, however, fails to account for another finding of the same study (Eisner & McQueen, 2005, Exp 2). Here, the VC tokens used during test consisted of vowels from a different speaker to which the fricatives from the exposure speaker were spliced onto. In a post-experiment questionnaire, more than two thirds of the participants indicated that they had noticed a change in speaker between exposure and test phase with these “hybrid” tokens. However, the learning effect was of a similar size as in the same-speaker condition. That is, the *perceived* identity of the speaker is not relevant for perceptual learning to generalize but rather the *acoustic similarity* of the ambiguous segment is crucial (see also Reinisch and Holt, 2014). These findings cannot be explained by assuming that listeners use *speaker identity* as a relevant context in percep-

[7] Note that this would have little repercussions outside the lab; in normal conversations the use of eye gaze in natural conversations (Rossano et al., 2009) is conventionalized so that relative spatial location is relatively constant in natural interactions.

tual learning because that would predict that a difference in perceived speaker identity should lower the transfer. This reasoning about the role of context then suggests that perceptual learning in the paradigm with a long and varied exposure phase is of a fundamentally different nature than that in the paradigm with minimal variation and repeated exposure-test cycles. Consequently, the two paradigms cannot be used interchangeably. However, this issue may need to be reappraised if the types of generalization found with one paradigm (see above) can also be observed with the other.

In the current context, however, the data at the very least show that perceptual learning is not necessarily constrained by spatial location - at least with one long and varied exposure phase. It seems that, when participants encounter variation during the exposure phase, spatial location of these tokens is not important for perceptual learning. This has two repercussions for the discussion about the nature of pre-lexical representations in spoken-word recognition. First of all, it means that findings with the paradigm using one long and varied exposure phase *cannot* easily be discounted as evidence for the shape of these units, because the perceptual-learning task is not necessarily sensitive to low-level features that would place it outside the domain of linguistic processing. However, the flip side of this argument is that paradigm with minimal variation may not be well suited to delineate the pre-lexical representations in spoken-word recognition. In this paradigm, learning is spatially selective, and we agree with Bowers et al. (2016) that this makes it unlikely that this paradigm is revealing properties of linguistic processing, which is generally considered to be spatially unspecific. When a design with minimal variation and repeated exposure-test cycles was used to delineate linguistic units by Reinisch et al., 2014, they found surprisingly specific learning. Learning was found consistently only for the trained contrast and did not generalize to other continua differing in manner but sharing the same place of articulation (/a[$^{b}/_{d}$]a/ → /a[$^{m}/_{n}$]a/). Similarly, no generalization was found to the same phoneme contrast cued by different acoustic cues (/a[$^{b}/_{d}$]a/ →/i[$^{b}/_{d}$]i/ where in the former place of articulation is mostly cued by formant transitions and in the latter it is mostly cued by the spectrum of the burst release). Even generalization to the same phoneme contrast when cued both by formant transitions (/a[$^{b}/_{d}$]a/ →/u[$^{b}/_{d}$]u/) was not found. Given that findings obtained with a design using minimal variation may not generalize to designs using one long and varied exposure phase, we must ask whether these results can be taken as evidence for highly specific prelexical representations. Fortunately, most of these results have in the meantime been replicated with a paradigm using one long and varied exposure phase. The non-generalization of learning across manner of articulation (/a[$^{b}/_{d}$]a/ → /a[$^{m}/_{n}$]a/) have been reported in different labs using such a paradigm (Mitterer et al., 2016b; Mitterer, Cho, & Kim, 2016; Reinisch & Mitterer, 2016; Schumann, 2014). Regarding the non-generalization to differently cued versions of the same phoneme (e.g., from /a[$^{b}/_{d}$]a/ to /i[$^{b}/_{d}$]i/), similar results have also been observed by two studies using one, long and varied exposure phase (Mitterer et al., 2013; Mitterer & Reinisch, 2017). Only the last finding by Reinisch et al. (2014), a lack of generalization across phonological context cued by similar acoustic cues /a[$^{b}/_{d}$]a/ →/u[$^{b}/_{d}$] u/), has not yet been replicated with another paradigm. It is conceivable, that learning might generalize in this condition where more variable input is presented. This is an avenue for future research.

In summary, the current data indicate that perceptual learning for speech perception in exposure-test paradigms is not always spatially selective. Spatial selectivity only seems to be involved when tested in a paradigm with minimal stimulus variation and short exposure-test block alterations. Our data indicate that the use of a visual bias versus a lexical bias does not seem to be a crucial variable in this respect. What is important is the amount of variation in the exposure set. At this stage, it is unclear whether spatial selectivity is only observed when exposure and test stimuli are identical, or whether it also could be observed with minimal variation as in the study of Van der Linden and Vroomen (2007). While this is an avenue for further investigation, the current data, viewed together with those of Keetels et al., (2015,2016), indicate that studies on perceptual learning of linguistic representations requires a sufficiently sized stimulus set. This is, from an experimental point of view, not good news, because this makes implementing studies not only more laborious but also makes it more difficult to control phonetic detail. The flip side of this point is that studies based on a varied, well curated, phonetically well-motivated input are likely to provide ecologically valid results.

Author note

The project was partly supported by a University of Malta Research Grant (CGSRP01-18) to the first author and a grant from the German Research Foundation (DFG; grant nr. RE 3047/1–1) to the second author, as well as by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A5A2A03042969).

CRediT authorship contribution statement

Holger Mitterer: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Eva Reinisch:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Conflicts of interest

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Appendix

Table A1: Critical word in the exposure phase of Experiment 1 and the ambiguous fricative used for the ambiguous stimulus.

word	ambiguous token source
roof	coda_rounded
knife	coda_unrounded
giraffe	coda_unrounded
laugh	coda_unrounded
cliff	coda_unrounded
handkerchief	coda_unrounded
sheriff	coda_unrounded
handcuff	coda_unrounded
earmuff	coda_unrounded
cough	coda_unrounded
dolphin	medial_unrounded
buffalo	medial_unrounded
butterfly	medial_unrounded
elephant	medial_unrounded
telephone	medial_unrounded
jellyfish	medial_unrounded
foot	onset_rounded
fork	onset_rounded
fan	onset_unrounded
fish	onset_unrounded
feather	onset_unrounded
finger	onset_unrounded
horse	coda_rounded
mouse	coda_rounded
greenhouse	coda_rounded
moose	coda_rounded
platypus	coda_unrounded
police	coda_unrounded
ice	coda_unrounded
price	coda_unrounded
airbus	coda_unrounded
cactus	coda_unrounded
compass	coda_unrounded
humus	coda_unrounded
lettuce	coda_unrounded
lotus	coda_unrounded
minus	coda_unrounded
necklace	coda_unrounded
palace	coda_unrounded
octopus	coda_unrounded
dinosaur	medial_rounded
seal	onset_unrounded
sun	onset_unrounded
saddle	onset_unrounded

References

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science, 14*, 592–597. https://doi.org/10.1046/j.0956-7976.2003.psci_1470.x.

Bowers, J. S., Kazanina, N., & Andermane, N. (2016). Spoken word identification involves accessing position invariant phoneme representations. *Journal of Memory and Language, 87*, 71–83. https://doi.org/10.1016/j.jml.2015.11.002.

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

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America, 116*, 3647–3658.

Cooke, M., & García Lecumberri, M. L. (2021). How reliable are online speech intelligibility studies with known listener cohorts? *The Journal of the Acoustical Society of America, 150*(2), 1390–1401. https://doi.org/10.1121/10.0005880.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition, 106*(2), 633–664. https://doi.org/10.1016/j.cognition.2007.03.013.

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2008). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory Phonology 10* (pp. 91–111). Mouton de Gruyter.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods, 47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00781.

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. 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*, 110–125.

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition, 89*, 105–132. https://doi.org/10.1016/S0010-0277(03)00070-2.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66*, 325–331. https://doi.org/10.1111/j.2044-8295.1975.tb01468.x.

Gould, S. J. J., Cox, A. L., Brumby, D. P., & Wiseman, S. (2015). Home is where the lab is: a comparison of online and lab data from a time-sensitive study of interruption. *Human Computation, 2*(1), Article 1. https://doi.org/10.15346/hc.v2i1.4.

Heald, J. B., Lengyel, M., & Wolpert, D. M. (2023). Contextual inference in learning and memory. *Trends in Cognitive Sciences, 27*(1), 43–64. https://doi.org/10.1016/j.tics.2022.10.004.

Jeffreys, S. H. (1961). *The Theory of Probability* (Third Edition). Clarendon Press.

Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review, 18*, 943–950. https://doi.org/10.3758/s13423-011-0129-2.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics, 27*, 359–384.

Kawahara, H., & Irino, T. (2005). Underlying Principles of a High-quality Speech Manipulation System STRAIGHT and Its Application to Speech Segregation. In P. Divenyi (Ed.), *Speech Separation by Humans and Machines* (pp. 167–180). US: Springer. https://doi.org/10.1007/0-387-22794-6_11.

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. *Speech Communication, 27*, 187–207. https://doi.org/10.1016/S0167-6393(98)00085-5.

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.

Kim, J., Gabriel, U., & Gygax, P. (2019). Testing the effectiveness of the Internet-based instrument PsyToolkit: A comparison between web-based (PsyToolkit) and lab-based (E-Prime 3.0) measurements of response choice and response time in a complex psycholinguistic task. *PloS One, 14*(9), e0221802.

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

Krieger-Redwood, K., Gaskell, M. G., Lindsay, S., & Jefferies, E. (2013). The selective role of premotor cortex in speech perception: A contribution to phoneme judgements but not speech comprehension. *Journal of Cognitive Neuroscience, 25*(12), 2179–2188. https://doi.org/10.1162/jocn_a_00463.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmerTest'. *R Package Version, 2*.

Lucke, S., Lachnit, H., Koenig, S., & Uengoer, M. (2013). The informational value of contexts affects context-dependent learning. *Learning & Behavior, 41*(3), 285–297. https://doi.org/10.3758/s13420-013-0104-z.

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

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological Abstraction in the Mental Lexicon. *Cognitive Science, 30*(6), 1113–1126. https://doi.org/10.1207/s15516709cog0000_79.

McQueen, J. M., Jesse, A., & Mitterer, H. (2023). Lexically mediated compensation for coarticulation still as elusive as a white christmash. *Cognitive Science, 47*(9), e13342.

Mitterer, H. (2006). On the causes of compensation for coarticulation: Evidence for phonological mediation. *Perception & Psychophysics, 68*(7), 1227–1240. https://doi.org/10.3758/BF03193723.

Mitterer, H., Chen, Y., & Zhou, X. (2011). Phonological abstraction in processing lexical-tone variation: Evidence from a learning paradigm. *Cognitive Science, 35*, 184–197. https://doi.org/10.1111/j.1551-6709.2010.01140.x.

Mitterer, H., Cho, T., & Kim, S. (2016). What are the letters of speech? Testing the role of phonological specification and phonetic similarity in perceptual learning. *Journal of Phonetics, 56*, 110–123. https://doi.org/10.1016/j.wocn.2016.03.001.

Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PLoS ONE, 4*(11), e7785.

Mitterer, H., & Reinisch, E. (2013). No delays in application of perceptual learning in speech recognition: Evidence from eye tracking. *Journal of Memory and Language, 69*(4), 527–545. https://doi.org/10.1016/j.jml.2013.07.002.

Mitterer, H., & Reinisch, E. (2017). Surface forms trump underlying representations in functional generalisations in speech perception: The case of German devoiced stops. *Language, Cognition and Neuroscience, 32*(9), 1133–1147. https://doi.org/10.1080/23273798.2017.1286361.

Mitterer, H., & Reinisch, E. (2022). *No delays in application of perceptual learning in speech recognition: Evidence from eye tracking*. https://osf.io/v2unz/.

Mitterer, H., Reinisch, E., & McQueen, J. M. (2018). Allophones, not phonemes in spoken-word recognition. *Journal of Memory and Language, 98*(Supplement C), 77–92. https://doi.org/10.1016/j.jml.2017.09.005.

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

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.

Nelson, S., & Durvasula, K. (2021). Lexically-guided perceptual learning does generalize to new phonetic contexts. *Journal of Phonetics, 84*. https://doi.org/10.1016/j.wocn.2020.101019 101019.

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

Pallier, C., Bosch, L., & Sebastian-Gallés, N. (1997). A limit on behavioral plasticity in speech perception. *Cognition, 64*(3), B9–B17. https://doi.org/10.1016/S0010-0277(97)00030-9.

Palomäki, K. J., Tiitinen, H., Mäkinen, V., May, P. J. C., & Alku, P. (2005). Spatial processing in human auditory cortex: The effects of 3D, ITD, and ILD stimulation techniques. *Brain Research. Cognitive Brain Research, 24*(3), 364–379. https://doi.org/10.1016/j.cogbrainres.2005.02.013.

Papoutsi, C., Zimianiti, E., Bosker, H. R., & Frost, R. L. A. (2023). Statistical learning at a virtual cocktail party. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-023-02384-1.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

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., & Mitterer, H. (2016). Exposure modality, input variability and the categories of perceptual recalibration. *Journal of Phonetics, 55*, 96–108. https://doi.org/10.1016/j.wocn.2015.12.004.

Reinisch, E., & Penney, J. (2019). The role of vowel length and glottalization in German learners' perception of the English coda stop voicing contrast. *Laboratory Phonology, 10*(1). https://doi.org/10.5334/labphon.176 1.

Reinisch, E., Weber, A., & Mitterer, H. (2013). Listeners retune phoneme categories across languages. *Journal of Experimental Psychology: Human Perception and Performance, 39*(1), 75–86. https://doi.org/10.1037/a0027979.

Reinisch, E., Wozny, D. R., Mitterer, H., & Holt, L. L. (2014). Phonetic category recalibration: What are the categories? *Journal of Phonetics, 45*, 91–105. https://doi.org/10.1016/j.wocn.2014.04.002.

Rossano, F., Brown, P., & Levinson, S. C. (2009). Gaze, questioning and culture. In J. Sidnell (Ed.), *Conversation analysis: Comparative perspectives* (pp. 187–249). Cambridge University Press.

Saltzman, D., & Myers, E. (2021). Listeners are initially flexible in updating phonetic beliefs over time. *Psychonomic Bulletin & Review, 28*(4), 1354–1364. https://doi.org/10.3758/s13423-021-01885-1.

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language, 111*. https://doi.org/10.1016/j.jml.2019.104070 104070.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics, 71*, 1207–1218. https://doi.org/10.3758/APP.71.6.1207.

Schumann, K. (2014). *Perceptual Learning in Second Language Learners [dissertation]*. Stony Brook University.

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

Sjerps, M. J., & Reinisch, E. (2015). Divide and conquer: How perceptual contrast sensitivity and perceptual learning cooperate in reducing input variation in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 41*, 710–722. https://doi.org/10.1037/a0039028.

Tzeng, C. Y., Nygaard, L. C., & Theodore, R. M. (2021). A second chance for a first impression: Sensitivity to cumulative input statistics for lexically guided perceptual learning. *Psychonomic Bulletin & Review, 28*(3), 1003–1014. https://doi.org/10.3758/s13423-020-01840-6.

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

van der Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 1483–1494.