

Lexically-guided perceptual learning does generalize to new phonetic contexts

Scott Nelson^{a,*}, Karthik Durvasula^b

^a*Department of Linguistics, Stony Brook University, USA*

^b*Department of Linguistics & Germanic, Slavic, Asian and African Languages, Michigan State University, USA*

Abstract

Lexically-guided and visually-guided perceptual learning have been argued to tap into the same general perceptual mechanism. Using the visually-guided paradigm, some have argued that the resulting retuning effect is specific to the phonetic context in which it is learned; which in turn has been used to argue that such retuning targets context-dependent sub-lexical units. We use three new experiments to study the generalizable nature of lexically-guided perceptual learning of fricative consonants and how type variation in the training stimuli affects it. In contrast to visually-guided retuning, we show that lexical retuning does generalize to new phonetic contexts, particularly when listeners are trained with type variation. This suggests that there is an abstract context-independent representation that is used in speech perception and during lexical retuning. While the same generalization is not clearly observed when type variation is eliminated, the lack of a clear interaction effect between training types prevents us from inferring that lexically-guided perceptual learning *needs* type variation within the training stimuli to generalize to new phonetic contexts. Furthermore, we point out that some of these effects are subtle and are only observable if we take into account pre-training group difference between the control and test groups.

Keywords: speech perception, perceptual learning, lexical retuning, prelexical processing, type variation

1. Introduction

A theory of how listeners process ambiguous speech sounds is an important part of a larger theory of speech perception. For example, if the ultimate goal of a listener is to reverse infer some discrete underlying representation from the continuous speech signal (Gaskell & Marslen-Wilson, 1996, 1998; Gow, 2003;

*Corresponding author

Email address: `scott.nelson@stonybrook.edu` (Scott Nelson)

Mitterer et al., 2013a; Durvasula & Kahng, 2015, amongst others), receiving ambiguous input greatly increases the difficulty of this task. When the auditory input is sufficiently ambiguous, listeners often search outside the auditory domain in order to find cues that may help provide disambiguating information. Past research into these types of phenomena has shown that identification can be affected by lexical (Ganong, 1980), phonological (Moreton, 2002) and visual (McGurk & MacDonald, 1976) information. More recent experimental paradigms have probed how listeners use either lexical information (Norris et al., 2003; Eisner & McQueen, 2005; Kraljic & Samuel, 2006, amongst others) or visual information (Bertelson et al., 2003; Van Linden & Vroomen, 2007; Reinisch et al., 2014, amongst others) to “retune” the boundaries that make up their perceptual categories.

The lexical retuning paradigm was developed by Norris et al. (2003) in order to test the effect of lexical feedback on adapting to an unusual speaker. In this original set of experiments, Norris et al. (2003) showed that, because of the bias that gets induced through the “Ganong effect” (Ganong, 1980), listeners’ identification responses to an $f\sim s$ continuum varied with the types of words in which they had previously heard an ambiguous blend of [f] and [s] ($[?_{fs}]$).¹ More specifically, they observed that participants who heard the ambiguous segment in words normally containing /f/ gave more “f” responses in the follow-up phonetic categorization task than participants who heard the ambiguous segment in words normally containing /s/. This was especially true for the points on the continuum that were already ambiguous (i.e., those points in the middle of the continuum that typically elicit an identification response near 50/50).²

Based on these results, it was proposed that listeners are getting lexical feedback during the speech perception process not for (initial) on-line recognition but instead for learning (Norris et al., 2003, p. 233-234). This distinction between recognition and learning was made to comply with a strictly feed-forward model of speech perception previously proposed (Norris et al., 2000). Therefore, this paradigm and the resulting “retuning” effect have been frequently referred to as “perceptual learning” (Norris et al., 2003; Eisner & McQueen, 2005; Kraljic & Samuel, 2005; Eisner & McQueen, 2006; Kraljic & Samuel, 2006; Samuel & Kraljic, 2009). Since the original study, other segments beyond fricatives have been shown to be amenable to retuning within the paradigm: stops (Kraljic & Samuel, 2006), liquids (Scharenborg et al., 2011), lexical tones (Mitterer et al., 2011). Furthermore, phonotactic information has also been shown to elicit a retuning effect using the paradigm (Cutler et al., 2008), suggesting that the effect is a general adaptation response to ambiguous, but categorizable, speech (see Kleinschmidt & Jaeger (2015) for a computational model of this learning effect).

¹Hereafter, we will refer to a sound ambiguous between two sounds [X] and [Y] as $[?_{XY}]$.

²Importantly, in a second experiment, they also argued that their results cannot be explained by selective adaptation (Eimas & Corbit, 1973; Samuel, 1986).

During the same time-frame that the lexical retuning paradigm was studied, the visually-guided retuning paradigm was developed by Bertelson et al. (2003).³ Here, it was the interaction of auditory and visual information, similar to what happens with the McGurk effect (McGurk & MacDonald, 1976), that drove the retuning of perceptual categories. McGurk & MacDonald (1976) observed that when a listener gets visual (lipread) information that contradicts the auditory input, their identification of the auditory segment can be modulated by the visual (lipread) information. Using the general phenomenon of the interaction of audio-visual information, Bertelson et al. (2003) observed that the simultaneous presentation of unambiguous visual input with ambiguous audio caused the retuning of the participants' perceptual categories. More specifically, they found that speakers who saw the /d/ visual input were more likely to respond "d" afterwards than the speakers who saw /b/ visual input. Bertelson et al. (2003) interpreted these results fairly generally and concluded that when cross modal biases occur, retuning of some sort is possible.⁴

In order to directly compare lexical retuning and visually-guided retuning, Van Linden & Vroomen (2007) created a series of five experiments, where the same ambiguous segment $[?_{tp}]$, in the same syllabic position (coda), was used in the training session for both tasks. Their takeaway from all experiments was that lexical retuning and visually-guided retuning show similar aftereffects and therefore are likely to be a result of the same general perception mechanism, a point repeated by Samuel & Kraljic (2009) in their review article.

Despite the observed similarity between lexical retuning and visually-guided retuning, there are reasons to perhaps be skeptical that they are tapping into the same general speech perception mechanism. First, the Ganong effect is typically observed with ambiguous speech, while the McGurk effect can alter the perception of even clear speech (McGurk & MacDonald, 1976). This suggests that visual input is specific to low-level information, while auditory input can tap into higher-level information. Samuel & Lieblich (2014) argue for this type of view by dividing the speech signal into a perceptual object and a linguistic object. Their claim is that visual information can only affect the perceptual object (low-level) while auditory information can affect both the perceptual object and the linguistic object (high-level). Additionally, Ullas et al. (2020) present experimental results showing that a combined retuning task (i.e., training involving both visual and auditory stimuli) does not enhance the level of perceptual learning. If both visual and lexical information were being used for the same task, then an additive effect should be expected. One hypothesis that Ullas et al. (2020) offer for explaining these data is that there may be "differences in their underlying structures and networks," between the mechanisms driving

³This has also been called "audio-visual recalibration".

⁴Interestingly, the McGurk effect is more prevalent when using a non-labial segment for the visual presentation accompanied by a labial auditory segment than the opposite scenario. This is briefly noted by Bertelson et al. (2003), but because they were interested in the "after-effects" (i.e., learning/retuning/recalibration rather than initial response), they did not explore this issue further.

the audiovisual and lexical retuning effects.

One area, crucial to this manuscript, where there continue to be conflicting results across (and sometimes within) the two paradigms is in regards to generalization to new stimuli within perceptual learning experiments.⁵ For example, Jesse & McQueen (2011) observed, using a lexical retuning paradigm, that when Dutch listeners were trained with the ambiguous segment in coda position, they transferred the effect when tested on segments from a [f \emptyset]~[s \emptyset] continuum where the crucial segment was in an onset position. These results suggest a generalization to a position-independent representation. Furthermore, generalization on a featural level has been shown for both stops (Kraljic & Samuel, 2006) and fricatives (Schuhmann, 2015; Durvasula & Nelson, 2018). Kraljic & Samuel (2006) trained listeners with a [?_{dt}] segment and then tested listener’s categorization of either the same d~t continuum, or one that varied in its place feature (b~p), and found that the effect did generalize to the previously unheard continuum. Similarly, using an [?_{fs}] training segment, Schuhmann (2015) found that fricatives could also generalize over a voice feature to a v~z continuum, but could not generalize to a p~t continuum where they differed in manner features. Certain vowel features have also been shown to generalize as Chládková et al. (2017) found that listeners are able to generalize from an i~e continuum to an u~o continuum in Greek.

In contrast, the lexical retuning effect has been argued to not generalize from one allophonic variant to another, *i.e.*, the acoustic target of a specific allophonic variant seems to be retuned, and not the phonemic category (Mitterer et al., 2013b). This suggests that while the retuning effect may be operating over a more abstract representation than just auditory features, it is still bounded by certain phonetic dimensions. Mitterer et al. (2016) show that some level of phonetic similarity is necessary for generalization to occur based on the perceptual learning of Korean stops. Echoing this sentiment, Reinisch & Mitterer (2016) observed that generalization does not seem to happen across manner features within the lexical retuning paradigm. Specifically, they observed that retuning with a p~t continuum does not transfer to an m~n continuum. Furthermore, Mitterer & Reinisch (2017) find that retuning guided by de-voiced final stops in German does not transfer to voiced or voiceless word-medial stops. These results suggest that perceptual learning requires not only phonetic similarity, but position-specificity as well.⁶

⁵A second area, not probed in this paper, in which empirical results for each paradigm have conflicted is stability or how long each effect lasts. Vroomen et al. (2004) observed that the retuning effect induced by visually-guided information was relatively short lasting. In contrast, for lexical retuning, independent results suggest that the effect was relatively long lasting; Kraljic & Samuel (2005) observed that it could last up to 25 minutes, and Eisner & McQueen (2006) found the effect to last for 12 hours regardless of whether the participant slept or not. The long-lasting nature of the lexical retuning effect could be due to it affecting a higher-level representation, while the visually-guided information may simply alter the “perceptual object”.

⁶Note, results from the related effect of selective adaptation are similarly diverse (Bowers et al., 2016; Mitterer et al., 2018; Llompert & Reinisch, 2018).

In work that is immediately relevant to the current paper, Reinisch et al. (2014) used a visually-guided retuning paradigm and observed no evidence of generalization across phonetic environments. Results from a series of three experiments suggest that the phonetic environments needed to be identical between training and testing in order to induce the perceptual retuning. In all their experiments, they used stimuli of the form [VCV]. Experiment 1 had the same phoneme cued by different vowels (/aba/-/ada/ and /ibi/-/idi/). Experiment 2 had different phonemes cued by the same vowel (/aba/-/ada/ and /ama/-/ana/). Experiment 3 had the same phonemes cued by different vowels, but this time with more distinct acoustic contexts (/aba/-/ada/ and /ubu/-/udu/). In all experiments, retuning was found for the control cases (same training and testing environments), but no retuning was found for the generalization cases (different training and testing environments). They interpret their results to mean that perceptual retuning in general is specific to phonemes, acoustic cues, and the phonetic context, but state that, “This conclusion rests, however, on the assumption that the visually-guided retuning reflects a general speech-perception mechanism (an assumption empirically supported by Van Linden & Vroomen (2007)” (p. 104).⁷ The general speech perception mechanism alluded to in this passage is one in which visually-guided retuning is the same as lexical retuning.

It is worth pointing out that, in many cases, it is difficult to directly compare the results of lexical retuning and visually-guided retuning as there are typical design differences between the two paradigms (as observed by Reinisch & Mitterer (2016)). Some examples of this include the presence/absence of the opposing phoneme during training blocks, the exposure to both phonemes as the ambiguous segment throughout the experiment, and the presence/absence of variability in training stimuli. Crucially, as is typical with visually-guided retuning experiments, Reinisch et al. (2014) presented the *same* training token with the ambiguous segment throughout the training block. Since the participants only heard the ambiguous auditory segment in one string/word context, it is possible that participants may have considered the ambiguity to be unique to that string.

If we look outside of the domain of perceptual learning, it has been shown that the learning of linguistic generalizations is aided by type experience. For example, it was found that a large amount of type variation in training data resulted in better identification of the /l/ ~ /r/ contrast for Japanese L2 English speakers (Logan et al., 1991; Lively et al., 1993). In their experiments, they trained Japanese speakers on the contrast by using a high number of minimal pairs (40+ in both cases). Their results showed a clear increase in correct identification of new words containing /l/ or /r/ from before training to after training. It is also noteworthy that they adapted the experimental paradigm of Strange & Dittmann (1984), which crucially lacked type variation and found minimal

⁷While we will use the term “phonetic context” throughout, it has also been called “phonological context”.

evidence of the ability to generalize the learned pattern. It was therefore argued that, in order to learn a phonological contrast, it was necessary for the training set to contain variable information. We find even more examples of this type of argument beyond the speech perception literature. Gerken & Boltt (2008) present results that suggest that 9-month old infants are able to generalize a constraint that says heavy syllables should be stressed to novel stimuli when presented with three unique training items, but failed to do so when presented with one unique training item presented multiple times. Denby et al. (2018) draw a similar conclusion using artificial language learning experiments. They looked at listeners' abilities to learn gradient phonotactic patterns and found that contextual variability (type frequency) clearly affected learning while the number of times an exemplar was repeated (token frequency) had no such clear effect. Furthermore, the learning of syntactic patterns (Gomez, 2002), morphological patterns (Endress & Hauser, 2011), and even visual patterns (Posner & Keele, 1968; Quinn & Bhatt, 2010) have all been observed to be strengthened by type variation.

These results taken together suggest that type experience is beneficial when generalizing a learned pattern. If this is indeed the case, then whether or not retuning can generalize to unobserved phonetic contexts may be predicated on proper type variation being provided to the listener. In this paper, we will present three experiments that more carefully test the generalizable nature of lexical retuning and probe what effect type vs. token experience has on it.

2. Experiment 1: Reproducing generalization in lexical retuning experiments

In Experiment 1 the goal was two-fold: (a) establish the typical finding that lexical retuning follows from the standard variation found in the training stimuli of such experiments (Norris et al., 2003), (b) propose a new way to analyze the relevant phonetic categorization results to allow for new insights into test and control group differences.

This experiment looked at training conditions with multiple unique vowels adjacent to the crucial consonant, as well as variation in syllabic position. We used voiceless fricatives as our test segments since they have been hypothesized to be more likely to transfer a learning effect across syllabic position than other consonant types (Mitterer et al., 2018). The testing block of the experiment always had the target fricative in onset position. The training block had it in both onset and coda positions. While there is clear evidence of lexical retuning transferring from training in coda position to testing in onset position (Jesse & McQueen, 2011), the status of lexical retuning when the training and testing segment are both in onset position is less clear. Jesse & McQueen (2011) also found a non-significant, but directionally expected, difference when participants were both trained and tested with the crucial segment in onset position. However, given the shift in categorization from retuning seen in their experiment was in the expected direction, the non-significance suggests a possibility of the

study being under-powered. Furthermore, as we point out in the results section of the current experiment, a retuning effect can sometimes be masked by differences in pre-training group baselines.

2.1. Method

2.1.1. Participants

Ninety-four undergraduate students from Michigan State University (Mean Age = 21.4; 70 female, 5 unreported gender) received either course credit or a small monetary reimbursement (\$10) for participating in the study. All participants identified as American English speakers and did not report any hearing problems.

2.1.2. Materials

The LDT used a list of 150 words containing 75 real English words and 75 phonotactically licit English nonce words. Thirty-four of the English words were training items, while the remaining 41 English words and all 75 nonce words were used as fillers. The list of training words can be found in Appendix A. The 34 training tokens contained either an [f] or an [s] (each segment distributed equally) and did not form a minimal pair when replaced with the opposing segment (e.g., *beef*, *sing*). All the training words used were monosyllabic and contained one of nine different vowels. Nine of the 17 words for each segment had the crucial segment in the word-initial onset position. Test words were controlled for frequency using the SUBTLEX-US corpus (Brysbaert & New, 2009). While the [f]-words were less frequent (Mean=12.85/million; MeanLog=2.6) than the [s]-words (Mean=20.77/million; MeanLog=2.46), a statistically unclear difference between the log frequencies was found [$t(28.3) = 0.48$, $p = 0.64$].⁸ The remaining 114 filler words contained no instances of [f s v z]. The real word fillers ranged in syllable count from 1-3 while the nonce word fillers ranged from 1-4.

The 150 words for the LDT were spoken by a female native American English speaker from Michigan. Each word was read aloud into a Logitech 980186-0403 microphone (frequency response 100Hz–16kHz; -67dBV/ubar, -47dBV/Pascal \pm 4dB) in a quiet room and recorded directly into Praat (Boersma & Weenink, 2016) at a sampling rate of 44.1 KHz. The speaker also recorded tokens of [fi] and [si]. These were used to make a fi~si continuum. The stimulus creation process was as follows: first, recordings of the selected tokens of [fi] and [si] were manually annotated in Praat (Boersma & Weenink, 2016) to mark the fricative and vowel portions of the token. From here, the entire process was automated using Praat scripting. To make each continuum, equal amounts of the fricative portion of each token was spliced out (165 ms; normalized to 50dB SPL). The amplitudes of the f/s pairs were then blended in 41 equal steps (e.g., step 1 was

⁸A note on terminology: we follow Dushoff et al. (2019) in discussing high p-values as being evidentially “unclear”, and low p-values (below $\alpha=0.05$) as clear evidence of differences. Further note, a lack of clear evidence does not automatically mean that there is no difference.

245 100% [f] and 0% [s]; step 2 was 97.5% [f] and 2.5% [s]; ... step 41 was 0% [f] and
 100% [s]). The continuum was then re-spliced back onto the vowel portion of the
 original [fi] token. It is well known that that formant transition information in
 the vowel acoustics is a cue for place of articulation (Delattre et al., 1955, 1962),
 therefore this method may introduce a slight bias towards “f” responses. Despite
 250 this limitation, this method has been shown to elicit normal response functions
 in previous studies (Norris et al., 2003; McQueen et al., 2006b; Durvasula &
 Nelson, 2018).

The continuum was then used in a pre-test to find the most ambiguous step.
 Of the 41 total steps, 14 were chosen to use for phonetic categorization. These
 255 included steps 1 & 41 which were the 100% [f] and 100% [s] tokens, respectively,
 as well as every other step between steps 7-29. This meant that the majority of
 the tokens that participants categorized were from the more ambiguous portion
 of each continuum. Thirteen American English speakers (Mean Age = 20.9
 years; 8 female, 1 unreported gender) from Michigan State University separate
 260 from those in the main experiment participated in the pre-test for partial class
 credit.

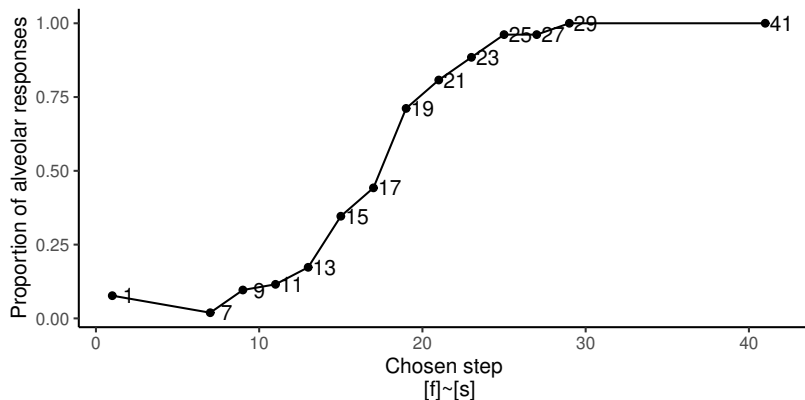


Figure 1: Categorization results for the fi~si continuum pre-test

Participants were tested using PsychoPy (Peirce et al., 2019). Each partic-
 ipant heard the 14 steps from the continuum four times each in random run-
 ning order. A two alternative forced choice paradigm was used. After hearing
 265 a sound, the participant was instructed to use a computer mouse to indicate
 whether the sound they had just heard contained either “f” or “s”. The mean
 response for each step was calculated to create an identification response
 function. This resulted in the sigmoidal function expected for consonant seg-
 ments (Liberman et al., 1957). The point on the response function closest to a 50%
 270 response rate was then interpreted as the most ambiguous point on the con-
 tinuum. For the fi~si continuum, the 50% response rate lay in between steps 17
 and 19 and therefore the fricative portion of step 18 was used to create [?_{fs}].

This was then used as a replacement for all of the [f] sounds in the LDT for the *FISI_{test}* training group.

275 Praat was once again used to manipulate audio stimuli, this time to create an altered version of the LDT wordlist. The *FISI_{test}* version of the list had all of the [f] tokens in the LDT replaced with [ʔ_{fs}]. The *FISI_{control}* wordlist was unaltered and therefore had no instances of an ambiguous token. To make the altered wordlist, the fricative portion of each word containing [f] was manually
280 annotated and marked at points of zero crossing. A Praat script was then used to automatically remove the original frication portion of the word and replace it with [ʔ_{fs}]. The returned wordlist was then manually checked for naturalness by the authors.

2.1.3. Procedure

285 Up to 12 participants were tested in the lab simultaneously. Prior to the experiment, participants were verbally instructed that they would be doing three tasks on the computer using various response mechanisms - a phonetic categorization (first and third tasks) requires the clicking of a mouse while the LDT (second task) would require them to use the keyboard to give a yes/no response.
290 Participants were also verbally instructed to answer as quickly and accurately as possible and to remain seated and quiet until everyone in the room had completed all three tasks. Participants wore over-the-ear headphones (Plantronics Audio 355; 20Hz-20kHz response) throughout the experiment, and specific instructions were visually presented to them on the screen before and during each
295 task.

The phonetic categorization tasks were the same format as the pre-test described above; both the phonetic categorization tasks contained the same 14 steps played four time each in random running order. Each participant was assigned to one of two groups: *FISI_{test}*, *FISI_{control}*. Both groups were played
300 the same fi~si continuum during the phonetic categorization tasks. As before, participants were given a two alternative forced choice task (“f” or “s”) and instructed to use a computer mouse to click which sound they heard.

Upon completion of the first phonetic categorization task (“Before”), participants did the LDT. They were instructed that they would hear a series of words,
305 one at a time, and would have to decide whether or not the word they heard was a real English word. Additionally, they were instructed to use the ‘a’ and ‘l’ keys on their computer keyboard to answer “no” or “yes” to the question, “Is this an English word?” An ‘a’ response corresponded to “no” and an ‘l’ response corresponded to “yes”. This information was constantly on screen as reference
310 for participants.

If a participant did not respond within 3.5 seconds of the onset of the sound, a new sound was played and no response was recorded. Each word was presented in random running order. Depending on the group that an individual participant was assigned to, they were presented the corresponding LDT list
315 as described above. After completion of the LDT, participants were given the second phonetic categorization task (“After”), which was identical to the first

phonetic categorization task. The only difference from the previous phonetic categorization was that a new random running order of the stimuli was used.

2.2. Results

320 In the original lexical retuning experiments, Norris et al. (2003) set a cri-
terion of 50% accuracy rate of the ambiguous segment in the LDT in order to
keep participants' data for analysis. This was to ensure that participants were
recognizing the words in the LDT as quality exemplars. In the current experi-
ment, participants were required to score 50% or higher in accuracy separately
325 in recognizing both words with [s], and those with either [f] or [$?_{fs}$], as well
as have an overall score of 50% or higher. Three participants ended up being
removed from the analysis. Overall, both groups had accuracy rates of 90%.

Our analysis will focus on a three-step window around the step used as [$?_{fs}$].
Since step 18 was determined to be the most ambiguous spot along the contin-
uum, this is where we should expect to see the most amount of change. The
poles of the continuum should remain relatively stable due to their unambigu-
ous statuses. It is the middle of the continuum, along the boundary between
the two segments, where recognition is most likely to fluctuate. By using these
three steps, it also more closely matches the area that Reinisch et al. (2014)
330 tested in their experiments. The primary difference is that their three steps
were individualized to each participant's most ambiguous step ± 1 step, while
our three steps are taken from the sample's most ambiguous step. Recall that
step 18 was used as [$?_{fs}$] in our experiments, but not presented in the phonetic
categorization tasks. Therefore, we set step 17 as the center of our window for
335 analysis, and define the 3-step window as steps 15, 17, and 19.

All data analysis and plotting were performed using R (R Development Core
Team, 2014) and relied heavily on the `tidyverse` set of packages (Wickham,
2017).⁹ The mixed effects logistic regression analyses were done using the `lme4`
package (Bates et al., 2015). Throughout, we modeled the counts of alveolar
340 responses as the dependent variable. We expect the $FISI_{test}$ group to give fewer
alveolar responses after training due to the ambiguous token [$?_{fs}$] replacing [f] in
the LDT. This is because the ambiguous segment should cause their boundary
for [f]-like segments to shift closer to the [s] side of the continuum, thus making
segments previously categorized as "s" to be more likely categorized as "f".

350 The categorization functions for these comparisons can be seen in Figure 2.

⁹All original source files, including R code, are publicly available at the following permanent
link: https://bitbucket.org/snelson89/retuning_generalization/

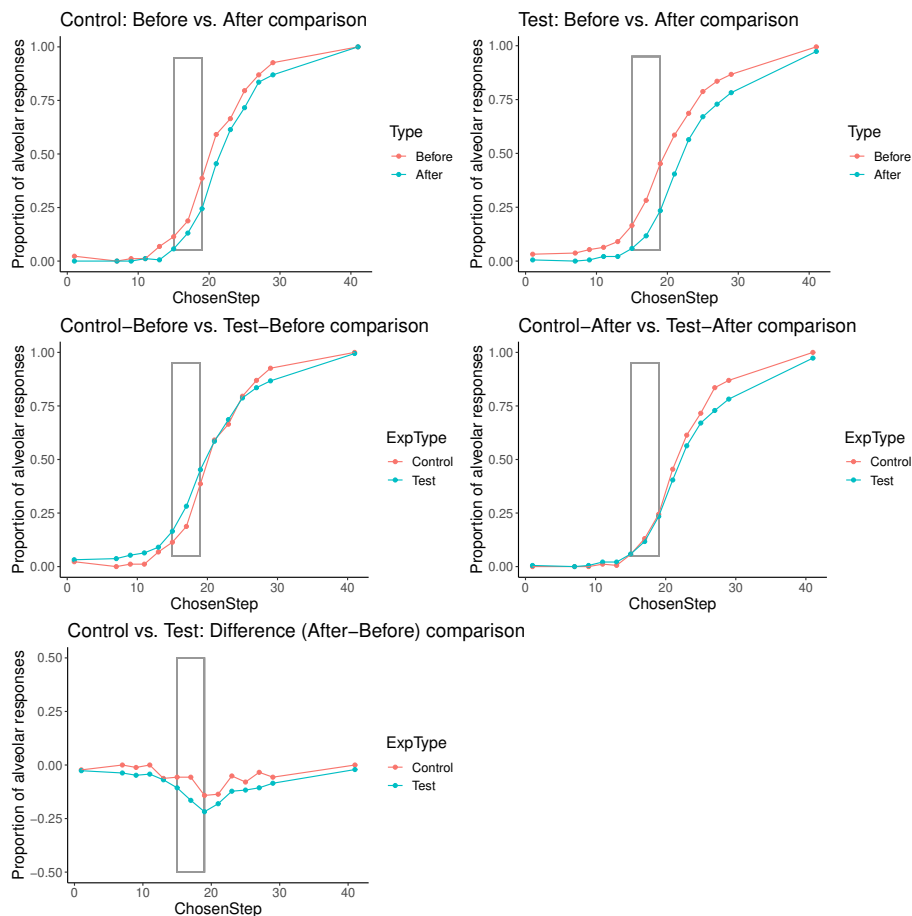


Figure 2: **Categorization results for Experiment 1.** First row, left column is $FISI_{control}$ before vs after comparison; First row, right column is $FISI_{test}$ before vs after comparison; Second row, left column is the before responses of $FISI_{control}$ vs the before responses $FISI_{test}$; Second row, right column is the after responses of $FISI_{control}$ vs the after responses $FISI_{test}$; Bottom row is the difference between after and before responses between the two groups. Boxed area in each graph indicates the analysis window.

We looked at 4 different pairwise statistical comparisons: (a) a within-participants “Before-After” comparison for the control group (b) a within-participants “Before-After” comparison for the test group; (c) a between-participants Test-Control comparison for the “Before” responses; (d) a between-participants Test-Control comparison for the “After” responses;

For comparisons (a-b), the models had a random intercept of participant and ChosenStep, and a by-participant random slope of `CONDITIONTYPE` (After vs.

Before; Before was the baseline).¹⁰ For comparisons (c-d), the models had a random intercept of participant and ChosenStep, and a by-participant random slope of EXPERIMENTTYPE (Control vs. Test; Control was the baseline).

The results of the statistical comparisons are in Table 1. As can be seen, there is no statistically clear evidence of a difference in the alveolar responses between the Test and Control groups before the LDT exposure. The traditional comparison that is made in lexical retuning experiments is between the responses of the “After” phonetic categorization task for both groups; for this standard After-After comparison, the results are statistically unclear too. This result is not entirely surprising as previously it has been observed that there was no clear difference between control and test groups (Norris et al., 2003). If one were to look at only the traditional analysis, one might have wrongly concluded that there is no clear evidence of retuning. However, there is clear evidence of a shift in the alveolar responses (After-Before) due to the LDT exposure in both the test and control groups. This suggests that the results could be confounded by the fact that both the control group and the test group showed shifts in the categorization function due to training.

	Comparison	Coefficient	SE	z-value	p ($> z $)
Control Group	After vs. Before	-0.83	0.27	-3.0	<0.01
Test Group	After vs. Before	-1.43	0.001	-1606.8	<0.00001
“Before” Responses	Test vs. Control	0.30	0.37	0.8	0.42
“After” Responses	Test vs. Control	-0.30	0.40	-0.75	0.45

Table 1: The mixed effects logistic results for the different pairwise comparisons in Exp. 1

Given the experiment design, we are able to ask the important question of whether the test group saw a bigger shift in alveolar responses than the control group. This comparison looks for shifts in the categorization function due to retuning while controlling for differences in baselines between the two groups and other task related effects. If one were to find such an interaction, then it would constitute evidence of a retuning effect. For this analysis, we used mixed effects logistic regression to model the counts of alveolar responses as the dependent variable. Given that there are two different variables, namely, CONDITIONTYPE and EXPERIMENTTYPE, we first tried to identify the best model through model comparison using the model AIC values, which balance model fit while controlling for over-parameterization, thereby decreasing the possibility of over-fitting (Akaike, 1974; Burnham & Anderson, 2002) (Note: lower AIC values indicate more support for a model.) The minimal model considered was one with just an

¹⁰Following the suggestions from Barr et al. (2013), the specific random effects structures used throughout this article were based on the largest ones that converged for the equivalent comparisons throughout the article. We did this to ensure that the comparisons in the three experiments were comparable.

intercept, and the maximal model considered was one with three independent variables: EXPERIMENTTYPE (Control vs. Test; Control was the baseline) as a between-subjects factor, CONDITIONTYPE (After vs. Before; Before was the baseline) as a within-subject factor, and an interaction of the two. All models that formed proper subsets of the maximal model were also considered.

As can be seen in Table 2, based on the AIC values, the best logistic regression model identified was the maximal model considered (*i.e.*, the model with both main effects and an interaction effect). The same inference can also be made based on the Chi-squared test of the log-likelihood ratios.

Model	AIC	Chi.Sq.	Chi.Sq. df	p-value
Only Intercept	1861			
Int. + CONTYPE	1863	0.29	1	0.59
Int. + EXPTYPE	1792			
Int. + CONTYPE + EXPTYPE	1793	0.29	1	0.59
Int. + CONTYPE*EXPTYPE	1788	6.91	1	0.009

Table 2: Comparison of different logistic regression models of all the data in Exp. 1. The best model is boldfaced. Each p-value represents the comparison of the model on that line with that on the preceding line. Note, Chi-squared tests are possible only for nested models, hence not all cells have Chi-squared values and associated p-values.

Given that the best model included the interaction of EXPERIMENTTYPE and CONDITIONTYPE, we present the model below (Table 3). As observed by the statistically clear interaction, the *change in alveolar responses* from Before to After in the test group was larger than in the control group. We take this as clear evidence of retuning in the test group.

Effect	Coefficient	SE	z-value	p ($> z $)
Intercept	-1.69	0.52	-3.2	<0.01
Test	0.43	0.33	1.3	0.2
After	-0.72	0.18	-4.1	<0.0001
Test:After	-0.68	0.25	-2.7	<0.01

Table 3: The best model for the categorization results in Experiment 1.

2.3. Discussion

The overall takeaway from Experiment 1 is that there is some evidence for generalization within lexical retuning experiments, provided one makes the appropriate comparison. That is, a comparison of the response curves of the two groups (Test/Control) *change* from “Before” to “After” the LDT. The general observation is a replication of prior work that retuning occurs in such experiments. The strongest support for generalization in these results comes from the fact that the set of training items only had one word that contained the vowel

410 [i]. Even if we include words with the phonetically similar [ɪ] vowel, this still
only accounts for five out of the seventeen training items. For a majority of the
training items, there is simply no overlap in training and testing environments.
This is suggestive of some type of generalization. To strengthen the claim that
415 fricatives can generalize to new phonetic contexts, stricter training and testing
conditions are required. In Experiment 2, we further probed the generalizable
nature of lexical retuning by seeing whether the learning effect can generalize
to a phonetic context that is completely absent from the training block.

3. Experiment 2: Generalization with non-identical training and testing conditions

420 Experiment 1 confirmed the basic findings from previous research, and showed
that the lexical retuning effect could be learned from a training set containing
many phonetic environments and transferred to one specific testing environment;
however, there was some overlap in training/testing vowel environments. It is
possible that the subset of training words where the vowel context was similar to
425 the vowel context used in the testing condition may have been enough to induce
the shift in categorization seen in Experiment 1. Therefore, while the results
are evidence of retuning, they are not clear evidence for context-independent
generalization of the retuning.

Recall that context dependency was more strictly tested using the visually-
430 guided retuning paradigm, and it was observed that the perceptual learning
effect appeared to be phonetic context-dependent. One of the findings from
Reinisch et al.'s (2014) study is that when participants are trained in the envi-
ronment of an [i] vowel, they are unable to generalize the effect when tested in
the environment of an [a] vowel (i.e., /ibi/ or /idi/ for training did not induce
435 the perceptual retuning shift for an [aba]~[ada] continuum). There has been no
study using the lexical retuning paradigm that has imposed as stringent training
and testing environmental restrictions. For this reason, Experiment 2 uses the
lexical retuning paradigm to see whether participants trained with an ambigu-
ous segment in the environment of an [i] or [ɪ] vowel will show lexical retuning
440 effects when tested on a [fa]~[sa] continuum. Since there is no overlap in vowel
context between the training and testing conditions, a shift in the categoriza-
tion by the test group would more explicitly support the context-independence
of the generalization.

3.1. Method

445 3.1.1. Participants

One hundred and thirty two undergraduate students from Michigan State
University (Mean Age = 19.7; 90 female, 1 gender fluid, 4 unreported gender)
received either course credit or a small monetary reimbursement (\$10) for par-

450 participating in the study.¹¹ All speakers identified as American English speakers and did not report any hearing problems.

3.1.2. Materials

The wordlist for the LDT once again contained 150 words split into 75 English words and 75 phonotactically licit English nonce words. The 116 filler words (41 real/75 nonce) from Experiment 1 were used in the Experiment 2
455 list. Since the goal of this experiment is to observe how the lexical retuning effect behaves when the phonetic context for the training and testing are non-overlapping, all the training items have the crucial segments next to an [i] or [ɪ] vowel. Expanding the criteria beyond just the [i] vowel was necessary in order to obtain a large enough training set. The addition of [ɪ] was due to it being the
460 most phonetically similar segment to [i] in American English (Hillenbrand et al., 1995). We used the [ɑ] vowel in the categorization task since it is maximally different from [i, ɪ] along the front and back dimensions; furthermore, the vowels chosen matched those in Reinisch et al. (2014).

All 34 training tokens once again contained either an [f] or an [s] (each segment distributed equally) and do not form a minimal pair when replaced with the opposing segment. The list of training words can be found in Appendix A. For both segments, 13 of the tokens had the crucial segment in onset position. It appeared in coda position for the remaining four. Each segment had 6 disyllabic tokens and 11 monosyllabic tokens. All disyllabic tokens had the crucial segment in onset position. Eight of the words for each segment contained [ɪ] and
470 were all monosyllabic (6 onset, 2 coda). The training tokens were controlled for frequency using the SUBTLEX-US corpus Brysbaert & New (2009). Due to the limited number of words matching the criteria, there was a slight mismatch in frequencies. The [s]-words were more frequent (Mean=47.56/million; MeanLog=2.68) than the [f]-words (Mean=19.39/million; MeanLog=2.58), but a statistically unclear difference was found between the log frequencies of the
475 two groups [$t(28.67) = -0.35, p = 0.73$].

The 34 words for the LDT were spoken by the same female native American English speaker from Michigan, as in Experiment 1. Each word was read aloud
480 into a Logitech 980186-0403 microphone (frequency response 100Hz–16kHz; -67dBV/ubar, -47dBV/Pascal \pm -4dB) in a quiet room and recorded directly into Praat (Boersma & Weenink, 2016) at a sampling rate of 44.1 kHz. The speaker also recorded tokens of [fa], and [sa], which were used to make the fa~sa continuum. Stimulus creation was done the same way as Experiment 1. The
485 created fa~sa continuum was used in a new pre-test to find the most ambiguous step. Eight American English speakers (Mean Age = 20.5 years; 2 female) from Michigan State University separate from those in the main experiment participated in the pre-test for partial class credit. Participants heard the same

¹¹When we first submitted the paper for review, we had 80 participants (78 for analysis); however, to increase power, we added another 51 participants to this experiment, and 3 participants did not make the threshold LDT criterion.

14 steps as in Experiment 1 (1,7,9,11,13,15,17,19,21,23,25,27,29,41). The mean
 490 response for each step was once again calculated to create the identification
 response function. Figure 3 shows these results below.

For the fa~sa continuum, step 16 was chosen as the most ambiguous segment.
 According to the pre-test results, the 50% response rate is between steps 13 and
 15, but there was also a steep rise from steps 13 to 15 and then a drop from 15
 495 to 17. Because of step 15's idiosyncratic behavior and the risk of outliers when
 using a small number of participants, we decided to listen to all steps between
 13 and 17 and choose the one which the first author considered to be maximally
 ambiguous. Ultimately, step 16 was chosen as the step to be used to create the
 [ʔ_{fs}] sound that would replace all of the [f] sounds in the LDT for the *FASA_{test}*
 500 group.

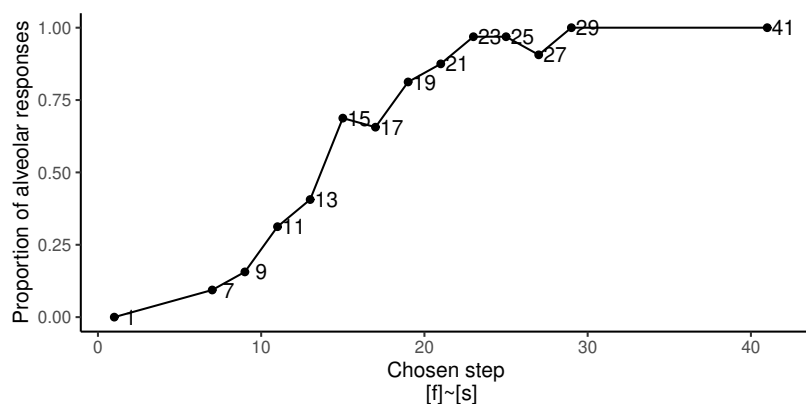


Figure 3: Categorization results for the fa~sa continuum pre-test

Praat was once again used to create an altered version of the LDT wordlist.
 The *FASA_{test}* version of the list had all of the [f] tokens in the LDT replaced
 with [ʔ_{fs}]. The *FASA_{control}* wordlist was unaltered and therefore had no in-
 stances of an ambiguous token. To make the altered wordlist, the same method
 505 was used as Experiment 1. The resulting wordlist was then manually checked
 for naturalness by the authors. While this method is potentially worrisome
 due to there being meaningful acoustic information about a following vowel in
 the frication portion of the signal (Yeni-Komshian & Soli, 1981; Soli, 1981; Mc-
 Murray & Jongman, 2016), previous experiments have used a similar method
 510 with no reported problem (Norris et al., 2003; Eisner & McQueen, 2005, 2006;
 McQueen et al., 2006a,b; Durvasula & Nelson, 2018).

3.1.3. Procedure

The general procedure is the same as outlined in Experiment 1 above. The
 few differences are outlined here. Participants were randomly assigned into one
 515 of two groups: *FASA_{test}* or *FASA_{control}*. For the phonetic categorization tasks,

both groups categorized the fa~sa continuum. For the LDT, participants heard the list that corresponded to their assigned group as described above. PsychoPy was once again used and all other procedural methods were exactly the same as Experiment 1.

520 *3.2. Results*

The same criteria as Experiment 1 were used to exclude any participants from analysis. Six participants failed to identify target segments accurately and were therefore removed. The *FISI_{control}* group had an overall response accuracy rate of 90%, while the *FASA_{test}* group's accuracy was slightly lower at 88%. The
525 categorization functions can be seen Figure 4.

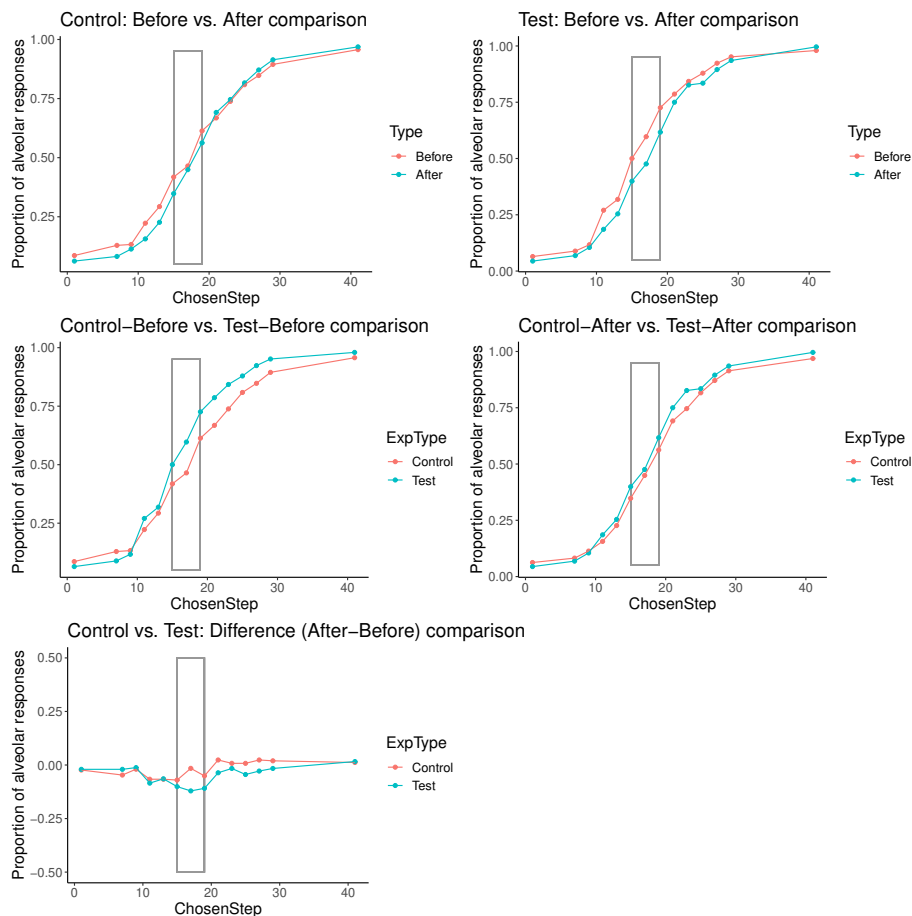


Figure 4: **Categorization results for Experiment 2.** First row, left column is $FASA_{control}$ before vs after comparison; First row, right column is $FASA_{test}$ before vs after comparison; Second row, left column is the before responses of $FASA_{control}$ vs the before responses $FASA_{test}$; Second row, right column is the after responses of $FASA_{control}$ vs the after responses $FASA_{test}$; Bottom row is the difference between after and before responses between the two groups. Boxed area in each graph indicates the analysis window.

A three-step window analysis was used once again. Because step 16 was used as $[?_{fs}]$, step 17 was chosen as the center of the window due to it having a response closer to 50% in the pre-test phonetic categorization task. The three-step analysis window therefore included steps 15, 17, and 19.

530 The statistical analyses for the relevant pairwise comparisons were identical to those in Experiment 1 (Table 4); the corresponding mixed effects logistic regression models had the same fixed effects and random effects. As with Experiment 1, the traditional “After-After” comparison does not result in a statistically clear result. However, this comparison, as with Experiment 1, could have been

535 confounded by different baselines for the test and control groups. As can be seen
in Table 4, there is a statistically clear difference between the alveolar responses
of the test and control groups, *before* the LDT exposure.¹² Some evidence for
retuning comes from the within-subjects “Before-After” comparison for the Test
group, which shows a statistically clear difference in the expected direction.

Comparison		Coefficient	SE	z	p ($> z $)
Control Group	After vs. Before	-0.23	0.16	-1.4	0.16
Test Group	After vs. Before	-0.64	0.18	-3.6	<0.001
“Before” Responses	Test vs. Control	0.67	0.30	2.2	0.026
“After” Responses	Test vs. Control	0.22	0.30	0.7	0.47

Table 4: The mixed effects logistic results for the different pairwise comparisons in Exp. 2

540 However, as a reminder, the better test according to us is whether the test
group saw a bigger *change* in alveolar responses than the control group. We
followed the statistical analysis strategy in Experiment 1. Based on the AIC
values, the best logistic regression model identified was the maximal model
considered, *i.e.*, with both main effects and an interaction effect (Table 2).
545 Again, the same inference can be made based on the Chi-squared test of the
log-likelihood ratios.

Model	AIC	Chi.Sq.	Chi.Sq. df	p-value
Only Intercept	3488			
Int. + CONTYPE	3488	2.643	1	0.11
Int. + EXPTYPE	3464			
Int. + CONTYPE + EXPTYPE	3464	2.642	1	0.11
Int. + CONTYPE*EXPTYPE	3461	4.177	1	0.04

Table 5: Comparison of different logistic regression models of all the data in Exp. 2. The best
model is boldfaced. Each p-value represents the comparison of the model on that line with
that on the preceding line. Note, Chi-squared tests are possible only for nested models, hence
not all cells have Chi-squared values and associated p-values.

The best model included the interaction of EXPERIMENTTYPE and CONDI-
TIONTYPE and is presented below (Table 6). As observed by the statistically
clear interaction, the *change in alveolar responses* in the test group was larger
550 than in the control group (mirroring the results in Experiment 1).¹³ Again, we

¹²Note, this was also true for the original set of 78 participants. This suggests that expecting
pre-training categorization functions to be identical might not be reasonable even with large
numbers of participants. This further reinforces our point that the “After-After” comparisons
are likely incomplete windows into the retuning effect.

¹³Note, though not presented here, we also ran a fully bayesian logistic mixed effects

take this as clear evidence of retuning in the test group.

Effect	Coefficient	SE	z-value	p ($> z $)
Intercept	-0.02	0.35	-0.05	0.96
Test	0.62	0.28	2.19	0.02
After	-0.26	0.12	-2.16	0.03
Test:After	-0.36	0.17	-2.06	0.03

Table 6: The best model for the categorization results in Experiment 2.

3.3. Discussion

Experiment 2 shows that lexically-induced perceptual learning persisted under strict training and testing conditions and was able to generalize from one phonetic environment to another. As with Experiment 1, the crucial addition of the “Before” phonetic categorization in the experimental design allowed us to observe a clear retuning effect for the test group, while there was no such clear evidence with the standard comparison.

The current results also bear on an issue that was raised in the introductory discussion to Experiment 1, namely the syllabic position of the training segments. Previous results have suggested that the retuning effect is clearly observed only when the training segment is in coda position (Jesse & McQueen, 2011). If this is the case, then it follows from the results of this experiment that only four critical training items are necessary to generalize the learning from one phonetic context to another; which is noticeably fewer than the 10 training items proposed to be necessary for the general retuning effect to occur (Poellmann et al., 2011). Given the results of this experiment, it could be that some amount of retuning does persist from onset-training to onset-testing but only passes a significant threshold when combined with retuning from coda-training to onset-testing. Regardless of whether or not the retuning effect was aided by the onset training segments, the observed generalization in this experiment resulted from more than one unique training item, suggesting that generalization in retuning experiments is clearly observable when this is the case.

Overall, the results from Experiment 2 confirm that the lexical retuning effect is able to generalize to new phonetic contexts even when the training context is strictly reduced and is non-overlapping with the testing context. Comparing these results to the results presented in Reinisch et al. (2014), they could be interpreted as showing that lexical information is potentially privileged in a way that auditory information is not (at least, when it comes to perceptual learning). There is still an alternative explanation for the difference in results. The current lexical retuning experiment was designed in a way to provide type experience

model incorporating the full fixed effect structure with default priors using the package `brms` (Bürkner, 2018). The two-sided 95% credible interval did not include zero, suggesting again that *the change in alveolar responses* in the test group was larger than in the control group.

to listeners during the LDT training phase, while visually-guided retuning experiments provide only token experience during the training phase. In order to better evaluate whether the generalizable nature is modality-specific or not, the
585 type/token distinction needs to be further tested.

4. Experiment 3: Reducing Stimulus Variation in Lexical Retuning

The effect of stimulus variation in the training data has seen little attention in the perceptual learning literature. Visually-guided retuning experiments normally present the same string (typically, some sort of VCV nonce word) continuously in order to induce the perceptual shift. In contrast, lexical retuning
590 experiments present multiple, unique words throughout the LDT. In the former case, there is no type or even token variation, while the latter includes within-experiment stimulus type and token variation. Experiment 3 tests what happens if one removes the within-experiment stimulus variation from a lexical
595 retuning experiment, and makes it as similar to the visually-guided paradigm as possible. As with Experiment 2, it will test whether training in the context of an [i] vowel will lead to retuning effects in a fa~sa continuum. The difference here is that now instead of 17 different training words, the LDT will contain 1 training word that gets repeated 17 times.

600 4.1. Method

4.1.1. Participants

One hundred and twenty three undergraduate students from Michigan State University (Mean Age = 20.1; 83 female, 2 non-binary, 5 unreported gender) received either course credit or a small monetary reimbursement (\$10) for participating in the study.¹⁴ All speakers identified as American English speakers and did not report any hearing problems.
605

4.1.2. Materials

The LDT for Experiment 3 uses a subset of the stimuli used in Experiment 2. The list of training words can be found in Appendix A. The overall number of unique words used in the LDT for Experiment 3 is reduced to
610 eight. Of the eight words, two are training items and the remaining six are fillers. Two of the filler items are real English words and the other four are phonotactically licit English nonce words. For the training items, the [f]-word is more frequent (Mean=31.61/million; MeanLog=3.21) than the [s]-word is
615 (Mean=8.55/million; MeanLog=2.64). Both training items were disyllabic and had the crucial segment in the onset position (word-initial). Using disyllabic words gives more information to the listener to identify it as a real English word

¹⁴The original version had 70 participants (67 for analysis), but we added 50 participants to increase power. Out of these added participants, 3 participants did not make the threshold LDT criterion.

and therefore gives a better chance that the ambiguous token will be recognized as a normal token of [f].

620 The filler items are all either mono- or di-syllabic. Two forms of the list were created. The $FASA2_{test}$ LDT list was sampled from the $FASA$ LDT list from Experiment 1 and therefore had the word containing [f] replaced with [? $_{fs}$]. The $FASA2_{control}$ LDT list was identical to the $FASA2_{test}$ list, but it had the normal pronunciation for all words, including the word containing [f]. Therefore, the
625 only difference between the two lists was in the single [f]-containing training token. The materials for the phonetic categorization tests are identical to the ones used in Experiment 2. The same most ambiguous segment [? $_{fs}$] as in Experiment 2 was used in Experiment 3 to create the single modified real word in the LDT list of the test group.

630 4.1.3. Procedure

The general procedure was the same as previous experiments. The few differences are outlined here. Participants were randomly assigned into one of two groups: $FASA2_{test}$ or $FASA2_{control}$. For the phonetic categorization tasks, both groups categorized the fa~sa continuum. For the LDT, participants heard
635 the list that corresponded to their assigned group. Since the LDT lists for this experiment only contain 8 total words, each word was now presented 17 times. The number of repetitions was set to 17 because that is the number of unique training words containing [f] or [s] in Experiments 1 and 2. Participants therefore heard 136 tokens in random running order during the LDT. While
640 the overall number of words participants heard was reduced by 14 as compared to Experiment 2, they did hear the same number of training tokens between Experiment 2 and Experiment 3. PsychoPy was once again used and all other procedural methods were the same as Experiments 1 and 2.

4.2. Results

645 The same criteria as in the previous experiments was used to exclude any participants from analysis. Nine participants failed to identify target segments accurately and were therefore removed. Both groups had lower overall accuracy rates than the previous experiments (86% for both $FASA2_{control}$ and $FASA2_{test}$), but this was brought down primarily by the nonce words. For the
650 training words containing the target fricatives, both groups had greater than 96% accuracy. The same analysis window was used as in Experiment 2 since all of the materials containing [? $_{fs}$] were identical. The categorization functions are in Figure 5.

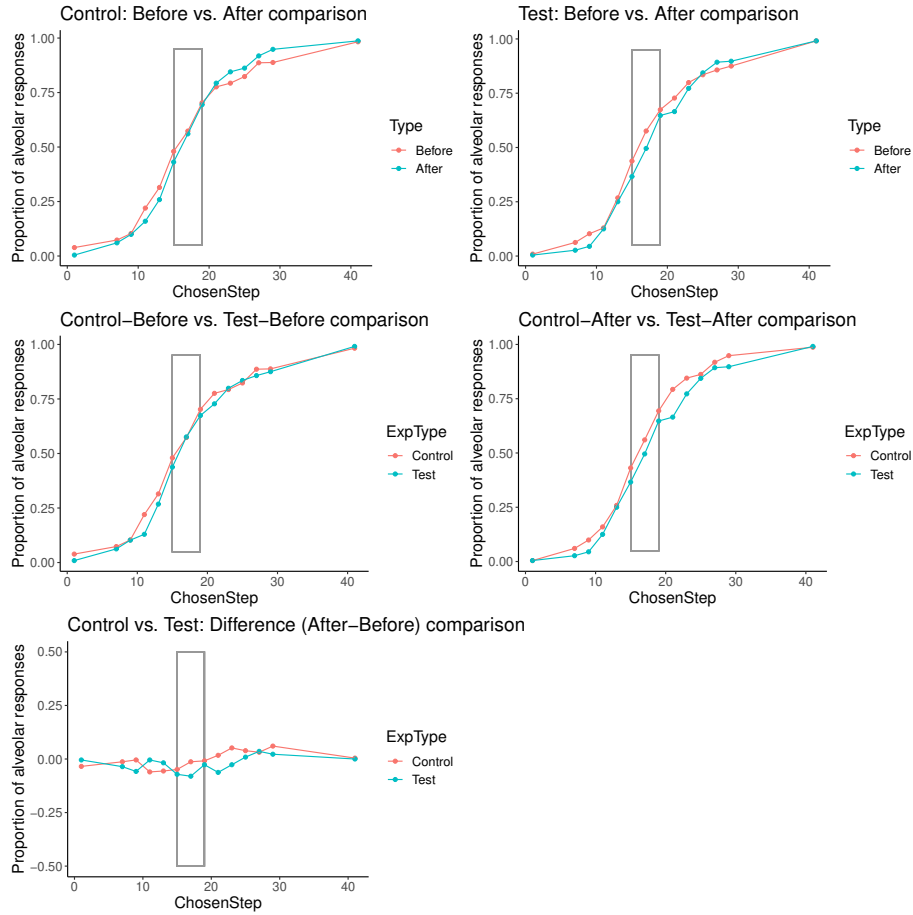


Figure 5: **Categorization results for Experiment 3.** First row, left column is $FASA\mathcal{L}_{control}$ before vs after comparison; First row, right column is $FASA\mathcal{L}_{test}$ before vs after comparison; Second row, left column is the before responses of $FASA\mathcal{L}_{control}$ vs the before responses $FASA\mathcal{L}_{test}$; Second row, right column is the after responses of $FASA\mathcal{L}_{control}$ vs the after responses $FASA\mathcal{L}_{test}$; Bottom row is the difference between after and before responses between the two groups. Boxed area in each graph indicates the analysis window.

655 The statistical analysis that follows is identical to that in Experiments 1 and 2. We first present the pairwise comparisons, and then proceed to find the best model that accounts for the overall alveolar responses in the test and control groups. Unlike with Experiments 1 and 2, there were no statistically clear differences for any of the pairwise comparisons (Table 7).

	Comparison	Coefficient	SE	z	p ($> z $)
Control Group	After vs. Before	-0.16	0.20	-0.8	0.41
Test Group	After vs. Before	-0.36	0.23	-1.6	0.12
“Before” Responses	Test vs. Control	-0.26	0.38	-0.68	0.49
“After” Responses	Test vs. Control	-0.42	0.38	-1.11	0.27

Table 7: The mixed effects logistic results for the different pairwise comparisons in Exp. 3

660 Following the strategy laid out in Experiment 1 and 2, we sought to test if the test group saw a bigger shift in alveolar responses than the control group. Based on the AIC values, the best logistic regression model identified was the model with just a fixed effect of EXPERIMENTTYPE; crucially, the model with the interaction term was not the best model (Table 8). Suggesting that there was no clear evidence of a retuning effect in this experiment.

Model	AIC	Chi.Sq.	Chi.Sq. df	p-value
Only Intercept	2966			
Int. + CONTYPE	2967	0.985	1	0.32
Int. + EXPTYPE	2961			
Int. + CONTYPE + EXPTYPE	2962	0.986	1	0.32
Int. + CONTYPE*EXPTYPE	2963	1.53	1	0.22

Table 8: Comparison of different logistic regression models of all the data in Exp. 3. The best model is boldfaced. Each p-value represents the comparison of the model on that line with that on the preceding line. Note, Chi-squared tests are possible only for nested models, hence not all cells have Chi-squared values and associated p-values.

665 Despite not being the best model, we present the model with the interaction term below (Table 8). As can be observed, there is no statistically clear interaction effect. We take this as minimally suggesting that the evidence for retuning was lacking in the test group.¹⁵ Note, this is not to say that there is no such effect. It is just that the effect was not clearly observable in our experiment.

¹⁵Note, as with Experiment 2, we also ran a fully bayesian logistic mixed effects model incorporating the full fixed effect structure. The two-sided 95% credible interval did include zero, suggesting that there is no clear evidence that *the change in alveolar responses* in the test group was larger than in the control group.

Effect	Coefficient	SE	z-value	p ($> z $)
Intercept	0.51	0.44	1.17	0.24
Test	-0.21	0.34	-0.61	0.55
After	-0.14	0.13	-1.06	0.29
Test:After	-0.23	0.19	-1.25	0.21

Table 9: The model with maximal fixed effects structure for the categorization results in Experiment 3.

670 If indeed there is a retuning effect in the presence of type variation (Experiment 2), and no retuning effect in the absence of type variation during exposure (Experiment 3), then it is possible to make another useful statistical prediction, namely that the crucial interaction effect in Experiment 2 should be larger in magnitude than the interaction effect in Experiment 3 (Gelman & Stern, 2006; 675 Nieuwenhuis et al., 2011).¹⁶ To test this prediction, we fitted a mixed effects logistic regression model to all the data from Experiment 2 and 3, with seven independent variables: EXPERIMENTNUMBER (2 vs. 3; 2 was the baseline), EXPERIMENTTYPE (Control vs. Test; Control was the baseline) as between-subjects factors, CONDITIONTYPE (After vs. Before; Before was the baseline) 680 as a within-subject factor, and all four possible interactions of the three variables. The model had random intercepts of participant and ChosenStep, and by-participant random slopes of EXPERIMENTNUMBER and EXPERIMENTTYPE. The crucial expectation based on the statistical prediction is that there should be a 3-way interaction between all three main variables; however, Table 10 shows 685 that there was no clear 3-way interaction.¹⁷

¹⁶We would like to thank Holger Mitterer for bringing this to our attention.

¹⁷A reviewer suggests that the results of Experiment 2 form the basis for a perfect prior for a Bayesian analysis. We disagree with this claim. Experiment 2 is a single experiment, and there is no reason to automatically trust that the effect size we found is the correct one; in fact, this is true for any single experiment. One would need a well conducted meta-analysis of similar experiments (that has access to all experiments similar to Experiment 2, not just the significant ones) to arrive at a good estimate. This is one reason why it is helpful to look at each experiment separately too, as we did.

Effect	Coefficient	SE	z-value	p ($> z $)
Intercept	0.15	0.74	0.21	0.83
Test	0.53	0.30	1.75	0.08
After	-0.18	0.07	-2.52	0.011
Exp3	0.24	0.31	0.786	0.43
Test:After	-0.26	0.10	-2.54	0.01
After:Exp3	0.16	0.11	1.52	0.13
Test:Exp3	-0.74	0.42	-1.79	0.07
Test:After:Exp3	0.06	0.15	0.43	0.67

Table 10: The model with maximal fixed effects structure comparing the results of Experiment 2 vs. Experiment 3.

While some have argued that the above statistical test of interaction (in Table 10) is the right comparison (Gelman & Stern, 2006; Nieuwenhuis et al., 2011), it is important to recognize that this is indeed a *separate* prediction from the prediction of a null effect in Experiment 3, and in fact suffers from its own weakness. As Brysbaert (2019) has pointed out recently, tests of interactions (even more so for 3-way interactions than for 2-way interactions) in psychological experiments likely need a large number of participants (sometimes, in the hundreds) to have sufficient power. Therefore, such comparisons which test for interactions between experiments are likely to be quite under-powered given typical samples sizes (an issue which thereby creates interpretational difficulties of its own).¹⁸ While we have tried to add more participants in our experiments (resulting in considerably more participants than in previous re-tuning experiments), it is still quite likely that such experiments in general are under-powered to test 3-way interactions (particularly, those involving between-subjects factors). Consequently, such comparisons make it more likely to get unclear or “non-significant results” than the traditional comparisons. The fact that the crucial 3-way interaction, though non-significant, was none-the-less in the expected direction very weakly suggests this possibility for our results, and we note it here as a word of caution for future research on the topic.

4.3. Discussion

In Experiment 2, it was observed that listeners could generalize to a new context in the testing condition when the training condition contained one unique vowel environment presented in multiple unique training words. Experiment 3 kept the same testing condition and the same single vowel environment for the training condition, but repeated the same training word multiple times. In

¹⁸Note, studies that are under-powered, with a specific effect size in mind, can also result in statistically significant effects, but not in a consistent manner. Therefore, the interpretation of non-significant effects across experiments becomes more difficult. On the other hand, a consistent non-significant result in a set of well-powered studies, given a specific minimum effect size, is indicative of an absence of an effect.

other words, it traded type experience for token experience to better match the conditions used in visually-guided retuning experiments. The results suggest that there is no clear evidence of listeners generalizing the lexical retuning effect, despite hearing the same number of raw tokens in the training condition
715 for both experiments. Further comparison between the results of Experiment 2 and Experiment 3 showed no clear statistical difference in the magnitude of the *change in alveolar responses*. The absence of a clear difference in the comparison between Experiments 2 and 3 make it difficult to directly claim that there is no generalization in the absence of type variation. However, the lack of a clear
720 evidence of retuning in Experiment 3 and the direction of the difference in the interaction effect between Experiments 2 and 3 are potentially suggestive that the non-significant difference could be due to insufficient power in our experiments, and prompt future higher powered experiments. It is equally important to point out that what this experiment has shown, in conjunction with the previous experiments, is that the presence of type experience in the training data
725 allows us, as researchers, to more clearly see the generalization effect in lexical retuning experiments.

5. Discussion and Conclusion

5.1. Summary of Results

This paper investigated the generalizable nature of perceptual learning in speech using the lexical retuning paradigm developed by Norris et al. (2003). Three new experiments were run to confirm that lexical retuning allowed for generalization to new phonetic contexts, particularly in the presence of type variation.
730

The results from Experiment 1 showed that generalization of the perceptual learning effect in lexical retuning experiments was possible when using a multitude of different training environments and testing on an environment that listeners only directly heard briefly throughout the training phase. This was observed by comparing the *change in alveolar responses* from Before to After
735 between the test group and the control group. Experiment 1 was therefore important not only to corroborate past results and give more supporting evidence for generalization within lexical retuning, but also to show that methodologically there is likely a need to expand the experimental/analytical comparisons used within these perceptual learning paradigms.
740

Experiment 2 observed that lexical retuning could still lead to generalization, even in more strictly controlled training and testing conditions. Overall, the results from the experiment challenge what was shown with visually-guided retuning experiments, mainly that the phonetic environment needs to be identical in training and testing conditions in order to see an effect (Reinisch et al.,
745 2014).
750

Experiment 3 was used to test whether or not type experience could explain why lexical retuning experiments allowed for generalizability. The training and testing environments for Experiment 3 were identical to the ones used in Experiment 2, but the overall number of training items were reduced. Results

755 from Experiment 3 show that listeners displayed no clear evidence of generaliz-
ing their perceptual training when presented with the same ambiguous stimuli
multiple times rather than multiple, unique stimuli. However, while the direc-
tion of the difference in the magnitude of the *change in alveolar responses* in
Experiments 2 and 3 is suggestive of a possible role for type variation, they are
760 ultimately inconclusive and thus provide no clear evidence for the claim that
generalization is only possible in the presence of type variation.

5.2. General Discussion

5.2.1. Methodological

The experiments in this paper contained design choices that deviate slightly
765 from the standard lexical retuning paradigm. These design choices were made
in order to measure the *change from “Before” to “After” between the Test and
Control groups* which most clearly displayed the lexical retuning effect in Exper-
iments 1 and 2. In fact, the standard comparisons masked the retuning effect.
It is therefore possible that previous experiments have missed certain insights
770 and retuning effects by not making this comparison. In what follows, we discuss
some of the benefits and potential weaknesses of the methodological innovations
introduced in this paper.

One significant difference between traditional lexical retuning experiments
and the experiments presented in this paper is the control group. In their
775 original experiment, Norris et al. (2003) include nonce words containing an
ambiguous segment and real words containing [f] and [s] in their control group’s
lexical decision task (LDT) wordlist. They argue that because the ambiguous
segment is in nonce words, there should not be a lexical retuning effect as there
is no lexical information to tap into. Since this was never explicitly confirmed,
780 there is a possibility that the presence of the ambiguous segment still had an
influence on the control group despite appearing in nonce words. To err on the
side of caution, the control group training sets in our experiments contained no
instances of an ambiguous segment, but rather just the clear [f] and [s] tokens
and the same filler words as the test group. This will hopefully control for
785 any subtle effect hearing the ambiguous segment may have had on those in the
control group. One could fairly argue that the manipulation presented by Norris
et al. (2003) is also relevant to weed out potential confounds, and we hope to
return to it in future work.

A second difference between the experiments in this paper and the typical
790 lexical retuning experiment is the lack of a second test group. The experiments
in this paper contained only 1 test group in which members hear words that
normally contain [f], but with all instances of [f] having been replaced by [?_fs].
However, most recent experiments omit the control group altogether and simply
use two test groups: one that is trained with words containing segment *A* re-
795 placed with [?_{AB}] and the other that is trained with words containing segment
B replaced with [?_{AB}](Eisner & McQueen, 2006; Kraljic & Samuel, 2005, 2006;
Kraljic et al., 2008; Cutler et al., 2008; Jesse & McQueen, 2011; Mitterer et al.,
2011; Reinisch & Holt, 2014; Mitterer et al., 2016; Reinisch & Mitterer, 2016).

It is worth noting that Norris et al. (2003) included both the test groups and
800 a control baseline, and the retuning effect looked somewhat symmetrical in the
plots, when compared to the control group. However, while the comparison
that involved the two test groups was statistically significant, neither test group
when compared against the baseline was statistically significant. Under these
statistical results, it is not clear if there is a genuine symmetrical shift for both
805 sets of test conditions or not. Therefore, our choice to compare a single test
group against the control case was driven by our belief that the [f] retuning case
needs to be studied separately from the [s] retuning case in order to understand
if both segments (or places of articulation) are susceptible to retuning. This
is particularly important given the observation of asymmetric patterns of per-
810 ception related to place of articulation and other phonological features (Cornell
et al., 2011; Scharinger et al., 2012; Cornell et al., 2013; Hestvik & Durvasula,
2016; Schluter et al., 2017; Hestvik et al., 2020). Note, our comparison makes it
potentially harder, not easier, for one to find a retuning effect (since the effect
size may be smaller). Furthermore, the experiments and the set of comparisons
815 included in the current research need quite a large set of participants as it is,
and adding more conditions would have made the number of participants even
larger. For these reasons, we have not pursued the possibility of a contrast-
ing test group (where the ambiguous segment was located in [s]-words); we do
however however hope to pursue this option in the future.

820 The final primary departure from previous work using lexical retuning is
the addition of a pre-LDT phonetic categorization task in addition to the post-
LDT phonetic categorization task. Typically, lexical retuning experiments have
only looked at between-subjects comparisons of post-LDT phonetic categoriza-
tion tasks with no reference to the baseline performance of the different groups
825 (Norris et al., 2003; Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2006;
McQueen et al., 2006b; Cutler et al., 2008; Sjerps & McQueen, 2010; Jesse &
McQueen, 2011; Scharenborg et al., 2011; Mitterer et al., 2011, 2013b; Reimisch
& Holt, 2014; Mitterer et al., 2016).¹⁹ While random assignment is thought
to wash out any variation in groups given a large enough sample, there is still
830 a risk that each group had different pre-retuning baseline performances on the
identification task.²⁰ Including both a “Before” and “After” categorization fur-
ther removes the risk of any sampling idiosyncrasies. The addition of a pre-LDT
phonetic categorization task also allows for a second between-group comparison
which is the difference in the change from “Before” to “After” between both
835 groups; we argue that this is necessary, as it controls for differences in base-
lines between the test and control groups and allows for differences related to
experimental manipulation to be better noticed.

¹⁹Eisner & McQueen (2006) is an exception. In their study, they found no clear difference
in the pre-test values for either group.

²⁰In fact, we see this in our own Experiment 2, where the baseline differences persisted
despite us adding many more participants during the review process. Another way to solve
the problem we are raising is to run multiple exact replications of specific studies in order to
tease apart random variation in group differences from systematic variation.

As we have mentioned above, we think knowing pre-training categorization functions is important in interpreting the retuning effect. However, this change
840 to the experimental design did not come without certain risks. By adding the pre-LDT phonetic categorization task, one runs the risk of alerting participants of the crucial segments importance and therefore altering the way they interact with the LDT. The “After” phonetic categorization task in our experiment is therefore subtly different from the phonetic categorization task that comes after
845 an LDT in other lexical retuning experiments due to how much meta-linguistic information participants have while performing the task. We do hope to return to this issue in future work, where we could address this concern by replacing the two categorization tasks with priming tasks, as in McQueen et al. (2006a), to ensure that the pre-training and post-training components do not allow the
850 participant to focus directly on the crucial segments.

Related to the same experiment design issue, it has been observed that the size of the lexical retuning effect is reduced when participants hear good exemplars of the crucial segments beforehand (Kraljic et al., 2008). Central to this finding was that both the good and ambiguous exemplars were presented
855 to individual participants during the lexical decision task, making them part of the training data. In the current experiment, the good exemplars appear in the non-lexical phonetic categorization task alongside many other ambiguous exemplars. The random and variable nature of a phonetic categorization task is different than a lexical decision task and, while we cannot be completely certain,
860 we believe that this mitigates the risk of the few good exemplars significantly affecting the retuning effect.

5.2.2. Theoretical

Reinisch et al. (2014) provide the strongest argument against generalization across phonetic contexts using the visually-guided retuning paradigm, and they interpret their results as suggesting that retuning generally targets context-
865 dependent sub-lexical units. However, our results (particularly from Experiment 2) point to the fact that lexical retuning does generalize to new phonetic contexts. Consequently, the results presented here suggest that there is an abstract context-independent representation that is used in speech perception and during
870 lexical retuning. One might be tempted to think that the result is not too surprising as fricatives have much more consistent properties across contexts than nasal or stop consonants. However, it is known that neighboring vowels can have a substantial effect on the spectral properties of fricatives (Yeni-Komshian & Soli, 1981; Soli, 1981; Whalen, 1983). Given the substantial effect of vocalic
875 context on the acoustics of fricatives, the reason for the generalization of retuning across contexts in the case of fricatives is not because of the immutability of fricative acoustics across contexts, but because the perceptual recovery of the abstract categories is easier and more stable across contexts in the case of fricatives (Whalen, 1983), and is better than the perceptual recovery of abstract
880 categories related to other segments (Hura et al., 1992). To re-iterate, the across-context retuning generalization effect, even in the case of fricatives, requires a fair degree of abstraction from the acoustic token and from the phonetic

context, and suggests that retuning targets abstract context-general categories and not context-specific (sub-lexical) categories.

885 Furthermore, the results from Experiment 3 are reminiscent of findings on L2 acquisition that suggest that high-variability training data lead to better learning of phonological categories (Logan et al., 1991; Lively et al., 1993). These findings have since been supplemented by results from other linguistic domains (Gomez, 2002; Gerken & Boltt, 2008; Endress & Hauser, 2011; Denby et al., 890 2018). That being said, the lack of an interaction effect between Experiments 2 and 3 in our paper prevents us from inferring that perceptual learning *needs* type variation within the training stimuli. Minimally, our results suggest that the retuning effect is more easily observed when there is type variation in the training stimuli. Further inquiry into the relationship between type variation 895 and perceptual learning may prove fruitful; however, we note again this is likely to involve a much larger number of participants than is currently the norm in such experiments.

Finally, our results also encourage future research within the visually-guided retuning paradigm using type variation in the training stimuli to see if there 900 is generalization to new contexts. If such a generalization were absent, then it would suggest that visually-guided retuning does not tap into the same perceptual learning mechanism as lexically guided retuning.

Acknowledgements

There are many people who have helped make this article possible. First, 905 we would like to thank James McQueen, Holger Mitterer, and one anonymous reviewer who provided valuable feedback and criticism that helped us greatly improve the article. Second, we would like to thank Russ Werner and CeLTA for help with running the experiments. Third, we would like to thank Dr. Yen-Hwei Lin, Dr. Silvina Bongiovanni, and members of the Michigan State Phonology- 910 Phonetics group for many helpful discussions. Finally, we would like to thank audiences at LabPhon 16, MidPhon 23, and LSA 93 for helpful comments, critique and discussion.

References

References

- 915 Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. doi:10.1109/TAC.1974.1100705.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of* 920 *Memory and Language*, *68*, 255–278.

- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. doi:10.18637/jss.v067.i01.
- 925 Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: a mcgurk aftereffect. *Psychological Science*, *14*, 592–597.
- Boersma, P., & Weenink, D. (2016). *Praat: doing phonetics by computer [Computer program]*. Version 6.0.19, retrieved 13 June 2016 from <http://www.praat.org/>.
- 930 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.
- Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? a tutorial of power analysis with reference tables. 935 *Journal of Cognition*, *2*. doi:<http://doi.org/10.5334/joc.72>.
- Brysbaert, M., & New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, *41*, 977–990.
- 940 Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference - A Practical Information-Theoretic Approach*. New York: Springer-Verlag.
- Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*, 395–411. doi:10.32614/RJ-2018-017.
- 945 Chládková, K., Podlipský, V. J., & Chionidou, A. (2017). Perceptual adaptation of vowels generalizes across the phonology and does not require local context. *Journal of Experimental Psychology: Human Perception and Performance*, *43*, 414.
- Cornell, S., Lahiri, A., & Eulitz, C. (2011). What you encode is not necessarily 950 what you store: evidence for sparse feature representations from mismatch negativity. *Brain Research*, *1394*, 79–89. doi:<https://doi.org/10.1016/j.brainres.2011.04.001>.
- Cornell, S., Lahiri, A., & Eulitz, C. (2013). Inequality across consonantal contrasts in speech perception: evidence from mismatch negativity. 955 *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 757–772. doi:<https://doi.org/10.1037/a0030862>.
- Cutler, A., McQueen, J. M., Butterfield, S., & Norris, D. (2008). Prelexically-driven perceptual retuning of phoneme boundaries, .

- 960 Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America*, *27*, 769–773.
- Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1962). Formant transitions and loci as acoustic correlates of place of articulation in american fricatives. *Studia linguistica*, *16*, 104–122.
- 965 Denby, T., Schecter, J., Arn, S., Dimov, S., & Goldrick, M. (2018). Contextual variability and exemplar strength in phonotactic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 280.
- Durvasula, K., & Kahng, J. (2015). Illusory vowels in perceptual epenthesis: The role of phonological alternations. *Phonology*, *32*, 385–416.
- 970 Durvasula, K., & Nelson, S. (2018). Lexical retuning targets features. In *Proceedings of the Annual Meetings on Phonology*. volume 5.
- Dushoff, J., Kain, M. P., & Bolker, B. M. (2019). I can see clearly now: Reinterpreting statistical significance. *Methods in Ecology and Evolution*, *10*, 756–759. URL: <https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13159>. doi:10.1111/2041-210X.13159. arXiv:<https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13159>.
- Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, *4*, 99 – 109. URL: <http://www.sciencedirect.com/science/article/pii/0010028573900066>. doi:[https://doi.org/10.1016/0010-0285\(73\)90006-6](https://doi.org/10.1016/0010-0285(73)90006-6). 980
- Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Attention, Perception, & Psychophysics*, *67*, 224–238.
- Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, *119*, 1950–1953.
- 985 Endress, A. D., & Hauser, M. D. (2011). The influence of type and token frequency on the acquisition of affixation patterns: Implications for language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 77.
- Ganong, W. F. (1980). Phonetic categorization in auditory word perception. 990 *Journal of experimental psychology: Human perception and performance*, *6*, 110.
- Gaskell, M. G., & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human perception and performance*, *22*, 144.
- 995 Gaskell, M. G., & Marslen-Wilson, W. D. (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 380.

- 1000 Gelman, A., & Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician*, *60*, 328–331. URL: <https://doi.org/10.1198/000313006X152649>. doi:10.1198/000313006X152649. arXiv:<https://doi.org/10.1198/000313006X152649>.
- 1005 Gerken, L., & Boltt, A. (2008). Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms and constraints. *Language Learning and Development*, *4*, 228–248.
- Gomez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431–436.
- Gow, D. W. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics*, *65*, 575–590.
- 1010 Hestvik, A., & Durvasula, K. (2016). Neurobiological evidence for voicing underspecification in english. *Brain and Language*, *152*, 28 – 43. URL: <http://www.sciencedirect.com/science/article/pii/S0093934X15300274>. doi:<https://doi.org/10.1016/j.bandl.2015.10.007>.
- 1015 Hestvik, A., Shinohara, Y., Durvasula, K., Verdonschot, R. G., & Sakai, H. (2020). Abstractness of human speech sound representations. *Brain Research*, *1732*, 146664. URL: <http://www.sciencedirect.com/science/article/pii/S0006899320300202>. doi:<https://doi.org/10.1016/j.brainres.2020.146664>.
- 1020 Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, *97*, 3099–3111.
- 1025 Hura, S. L., Lindblom, B., & Diehl, R. L. (1992). On the role of perception in shaping phonological assimilation rules. *Language and Speech*, *35*, 59–72. URL: <https://doi.org/10.1177/002383099203500206>. doi:10.1177/002383099203500206. arXiv:<https://doi.org/10.1177/002383099203500206>. PMID: 1287392.
- Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review*, *18*, 943–950. doi:10.3758/s13423-011-0129-2.
- 1030 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.
- Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive psychology*, *51*, 141–178.
- 1035 Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic bulletin & review*, *13*, 262–268.

- Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological science*, *19*, 332–338.
- 1040 Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, *54*, 358.
- Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training japanese listeners to identify english /r/ and /l/. ii: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the acoustical society of America*, *94*, 1242–1255.
- 1045 Llompart, M., & Reinisch, E. (2018). Acoustic cues, not phonological features, drive vowel perception: Evidence from height, position and tenseness contrasts in german vowels. *Journal of Phonetics*, *67*, 34–48.
- 1050 Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training japanese listeners to identify english /r/ and /l/: A first report. *The Journal of the Acoustical Society of America*, *89*, 874–886.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.
- 1055 McMurray, B., & Jongman, A. (2016). What comes after /f/? prediction in speech derives from data-explanatory processes. *Psychological science*, *27*, 43–52.
- McQueen, J. M., Cutler, A., & Norris, D. (2006a). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*, 1113–1126.
- 1060 McQueen, J. M., Norris, D., & Cutler, A. (2006b). The dynamic nature of speech perception. *Language and speech*, *49*, 101–112.
- Mitterer, H., Chen, Y., & Zhou, X. (2011). Phonological abstraction in processing lexical-tone variation: Evidence from a learning paradigm. *Cognitive Science*, *35*, 184–197.
- 1065 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.
- Mitterer, H., Kim, S., & Cho, T. (2013a). Compensation for complete assimilation in speech perception: The case of korean labial-to-velar assimilation.
- 1070 *Journal of Memory and Language*, *69*, 59–83.
- 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*, 1133–1147.

- 1075 Mitterer, H., Reinisch, E., & McQueen, J. M. (2018). Allophones, not phonemes in spoken-word recognition. *Journal of Memory and Language*, *98*, 77–92. URL: <http://www.sciencedirect.com/science/article/pii/S0749596X17300748>. doi:<https://doi.org/10.1016/j.jml.2017.09.005>.
- 1080 Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013b). Phonological abstraction without phonemes in speech perception. *Cognition*, *129*, 356–361. URL: <http://www.sciencedirect.com/science/article/pii/S0010027713001443>. doi:<https://doi.org/10.1016/j.cognition.2013.07.011>.
- Moreton, E. (2002). Structural constraints in the perception of english stop-sonorant clusters. *Cognition*, *84*, 55–71.
- 1085 Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature neuroscience*, *14*, 1105–1107.
- Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*, 299–325.
- 1090 Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *30*, 1113–1126.
- Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior research methods*, (pp. 1–9).
- 1095 Poellmann, K., McQueen, J. M., & Mitterer, H. (2011). The time course of perceptual learning. In *The 17th International Congress of Phonetic Sciences (ICPhS XVII)* (pp. 1618–1621). Department of Chinese, Translation and Linguistics, City University of Hong Kong.
- 1100 Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of experimental psychology*, *77*, 353.
- Quinn, P. C., & Bhatt, R. S. (2010). Learning perceptual organization in infancy: The effect of simultaneous versus sequential variability experience. *Perception*, *39*, 795–806.
- 1105 R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: <http://www.R-project.org> ISBN 3-900051-07-0.
- 1110 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*, 539.

- Reinisch, E., & Mitterer, H. (2016). Exposure modality, input variability and the categories of perceptual recalibration. *Journal of Phonetics*, *55*, 96 – 108. URL: <http://www.sciencedirect.com/science/article/pii/S0095447015001084>. doi:<https://doi.org/10.1016/j.wocn.2015.12.004>.
- 1115 Reinisch, E., Wozny, D. R., Mitterer, H., & Holt, L. L. (2014). Phonetic category recalibration: What are the categories? *Journal of phonetics*, *45*, 91–105.
- Samuel, A. G. (1986). Red herring detectors and speech perception: In defense of selective adaptation. *Cognitive psychology*, *18*, 452–499.
- Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*, 1207–1218.
- 1120 Samuel, A. G., & Lieblich, J. (2014). Visual speech acts differently than lexical context in supporting speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 1479.
- Scharenborg, O., Mitterer, H., & McQueen, J. M. (2011). Perceptual learning of liquids. In *Interspeech 2011: 12th Annual Conference of the International Speech Communication Association* (pp. 149–152).
- 1125 Scharinger, M., Bendixen, A., Trujillo-Barreto, & N.J., J., Obleser (2012). A sparse neural code for some speech sounds but not for others. *PLoS ONE*, *7*. doi:<https://doi.org/10.1371/journal.pone.0040953>.
- 1130 Schluter, K. T., Politzer-Ahles, S., Al Kaabi, M., & Almeida, D. (2017). Laryngeal features are phonetically abstract: Mismatch negativity evidence from arabic, english, and russian. *Frontiers in Psychology*, *8*, 746. URL: <https://www.frontiersin.org/article/10.3389/fpsyg.2017.00746>. doi:10.3389/fpsyg.2017.00746.
- 1135 Schuhmann, K. S. (2015). *Perceptual learning in second language learners*. Ph.D. thesis The Graduate School, Stony Brook University: Stony Brook, NY.
- Sjerps, M. J., & McQueen, J. M. (2010). The bounds on flexibility in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 195.
- 1140 Soli, S. D. (1981). Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America*, *70*, 976–984.
- Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r/ by japanese adults learning english. *Perception & psychophysics*, *36*, 131–145.
- 1145 Ullas, S., Formisano, E., Eisner, F., & Cutler, A. (2020). Audiovisual and lexical cues do not additively enhance perceptual adaptation. *Psychonomic Bulletin & Review*, (pp. 1–9).

- 1150 Van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 1483.
- Vroomen, J., van Linden, S., Keetels, M., De Gelder, B., & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Communication*, *44*, 55–61.
- 1155 Whalen, D. (1983). Vowel information in postvocalic fricative noises. *Language and Speech*, *26*, 91–100. URL: <https://doi.org/10.1177/002383098302600106>. doi:10.1177/002383098302600106. arXiv:<https://doi.org/10.1177/002383098302600106>. PMID: 6621206.
- 1160 Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. URL: <https://CRAN.R-project.org/package=tidyverse> r package version 1.2.1.
- Yeni-Komshian, G. H., & Soli, S. D. (1981). Recognition of vowels from information in fricatives: Perceptual evidence of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America*, *70*, 966–975.

1165 **Appendix A. Critical training words for lexical decision tasks**

Experiment 1		Experiment 2/3	
bluff	truss	cliff	kiss
chef	chess	reef	geese
cliff	bliss	thief	piece
deaf	less	whiff	bliss
poof	deuce	film	silk
whiff	kiss	filth	sick
beef	geese	fish	sim
cough	boss	fill	sing
fudge	such	fifth	sip
felt	sect	fib	seek
food	soup	fiend	seem
fab	sash	feeble	seagull
fade	say	female	seated
fig	sip	fetal	seeing
fish	silt	fever	seeker
fool	soon	fiji	seeping
full	sulk	feline	seething

Table A.11: **Training words for Lexical Decision Tasks.** The left side of the table contains the 34 training words used in Experiment 1. The right side of the table contains the 34 training words used in Experiment 2. The bolded subset of words on the Experiment 2 side are the training words used in Experiment 3.