The perceptual generalization of normalized cue distributions across speakers

Wei Lai (wei.lai@vanderbilt.edu)
Department of Psychology and Human Development, Vanderbilt University
Aini Li (liaini@sas.upenn.edu)
Department of Linguistics, University of Pennsylvania

Abstract

Listeners adapt to specific speakers' speech cue distributions and generalize the adaptation to the perception of a different speaker. It remains unclear whether listeners track and generalize the distributional statistics of raw, un-normalized cues or normalized cue distributions relative to the speaker's acoustic space. To address this question, we adopted a perceptual generalization paradigm to investigate whether manipulating context properties of a training speaker (Female A)'s speech would lead to different categorization results of critical phonemes in a test speaker (Female B)'s speech. Experiment 1 showed that learning Female A's speech containing the same set of sibilants but shifted vowel formants would lead to different categorization of Female B's sibilants: listeners exposed to raised vowel formants were more likely to identify an /s/ and those exposed to lowered vowel formants were more likely to identify /ʃ/ in Female B's speech, compared with participants exposed to unaltered vowel contexts. Experiment 2 showed that learning of Female A's speech containing the same set of stops but manipulated word frame duration would lead to different categorization results of Female B's stops: listeners temporally exposed to expanded word frames were more likely to identify a /t/ and those exposed to compressed contexts were less likely to identify a /t/ in Female B's speech, compared to participants exposed to unaltered temporal cues. These results suggest that listeners keep track of normalized cue distributions relative to the speaker's acoustic space and generalize those distributions to guide their speech perception behaviors.

Keywords: perceptual learning; speaker normalization; vocal tract normalization; speech rate normalization; generalization

Introduction

Human speech is highly variable. Listeners are known to adjust their perceptual expectations rapidly to be better aligned with the production of specific speakers via a process of "perceptual learning" (Norris, McQueen, & Cutler, 2003). Listeners also generalize the distributional properties of one speaker's linguistic units to their perception of a different speaker's speech (Kraljic & Samuel, 2006, 2007; Reinisch & Holt, 2014; Xie et al., 2018; Tamminga, Wilder, Lai, & Wade, 2020; Lai, 2021). Perceptual learning can generalize across speakers of the same gender (Reinisch & Holt, 2014; Tamminga et al., 2020; Lai, 2021) and of different genders (Kraljic & Samuel, 2006, 2007; Reinisch & Holt, 2014; Van der Zande, Jesse, & Cutler, 2014; Lai, 2021); it has also been found for different types of phonemes, including fricatives (Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2007; Reinisch & Holt, 2014; Tamminga et al., 2020) and stops (Kraljic & Samuel, 2006, 2007; Van der Zande et al., 2014).

The encoding of phonemic contrasts not only involves raw acoustic statistics of the target phonemes, but also their relationship with cue distributions of their surrounding acoustic signals, or contexts (e.g. word frames). The mechanism by which listeners uncover the intended meaning of variable speech by scaling phonetic cues according to contextual baselines is called *normalization* (Diehl, Souther, & Convis, 1980; Johnson, 1990; Johnson, Strand, & D'Imperio, 1999; Dilley & Pitt, 2010, etc.). This study investigates the incorporation of two kinds of normalization processes in perceptual learning.

The first kind is vocal tract length normalization (Johnson, 2018) and its impact on the perception of /s-ʃ/. Speakers' vocal tract lengths correlate with the spectral energy distributions of their sibilants and vowels (Kent, 1993): longer vocal tract leads to lower frequencies of vowel formants and lower sibilant spectral energy. Listeners can infer the information of a speaker's vocal tract size from his/her vowel formants (Reby & McComb, 2003; Smith & Patterson, 2005; Lammert & Narayanan, 2015), and use this information to locate the distributional properties of the sibilant frequency of this speaker (Johnson, 2018). For example, Strand and Johnson showed that when female voices were manipulated into male voices through lowering vowel formants and F_0, listeners were more likely to perceive /s/ rather than /ʃ/ for sounds on the same sibilant continuum. This is partly because male speakers tend to have a longer vocal tract than female speakers which leads to lower spectral energy frequencies.

The second kind is speech rate normalization, which refers to the phenomenon where listeners use the temporal characteristics of the contextual information to segment speech units and identify speech targets (Port, 1979; Summerfield, 1975; Dilley & Pitt, 2010). This mechanism plays a critical role in the identification of phonemic contrasts signaled by temporal cues such as voice-onset-time (VOT). On top of the tendency that longer VOT leads to more frequent identifications of voiceless stops rather than voiced ones, it is also reported that the perception of stop voicing also varies with the overall speech rate of an adjacent phrase (Port, 1979; Summerfield, 1975).

Normalization mechanisms have also been documented to affect the perceptual integration of various other speech cues than spectral and temporal, such as F_0 (Wong & Diehl, 2003).

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Although the process is widely recognized, its relationship with the mechanism of perceptual learning and generalization has not been well elaborated. In specific, it still remains unresolved whether listeners keep track of talker-specific distributional statistics of the raw, un-normalized cues or their relative distributions in a particular speaker's acoustic space. Perceptual learning models that take raw acoustic values of a phonemic contrast as the model input successfully predict the perceptual boundary shift after adaptation (Kleinschmidt & Jaeger, 2015). Recently, there have also been attempts to incorporate normalization mechanisms into perceptual learning models, but relevant studies have mainly focused on the learning of intonation (Xie, Buxó-Lugo, & Kurumada, 2021; Kurumada & Roettger, 2021). To our knowledge, no study has been conducted to evaluate the interference of speech normalization in the perceptual generalization of phonemic contrasts at the segment level. This forms the main goal of the current study.

The present study

The present study probes the nature of listeners' mental representations of speech distributions that they keep track of during perceptual adaptation and later generalize to guide their subsequent speech perception. We adopt the experimental paradigm of the generalization of perceptual learning across speakers, where the perceptual learning with one speaker continues to affect the perception of a different speaker's speech. Cross-speaker perceptual generalization provides a suitable testing ground to evaluate whether perceptual learning involves tracking normalized cue distributions or absolute acoustic values, because the presence of multiple speakers involves acoustic variability of the speech context and therefore presents the necessity to clarify whether the generalized cue distributions are speaker-dependent.

Two experiments are conducted to investigate the generalization of two distinct kinds of cues. Experiment 1 investigates the generalization of spectral cues to a sibilant contrast, asking how listeners generalize their learning of sibilant frequencies across speakers with different sizes of vocal tracts. Experiment 2 investigates the generalization of temporal cues to a plosive contrast, asking how listeners generalize their learning of voice-onset-time across speakers with different speech rates. Both experiments adopt a between-subject design. Participants in different experimental conditions hear training stimuli that have identical acoustic signals of the critical target phonemes, but the acoustic distributions of the context information, i.e. word frames where the target sounds occur, are different.

We evaluate two competing hypotheses. A *raw cue distribution* hypothesis claims that listeners keep track of the absolute values of acoustic cues to a phonemic contrast and use these distributions to uncover phoneme identities in subsequent speech perception behaviors. On the contrary, a *normalized cue distribution* hypothesis claims that listeners keep track of the relevant cues' relative distributions in the speaker's acoustic space and generalize those relative distributions to further guide their speech perception activities. The two hypotheses we evaluate make different predictions about the categorization results obtained from different experimental conditions. The *raw cue distribution* hypothesis predicts that participants in different conditions should have similar results since they were tracking the acoustic statistics of the target phoneme that remains identical across conditions; whereas the *normalized cue distribution* hypothesis predicts that participants in different conditions would categorize the test stimuli in ways that are consistent with the manipulation of their contextual information. Specific predictions for each experiment will be further provided in the following sections.

Experiment 1: vocal tract normalization

Experiment design

Experiment 1 investigates whether manipulating spectral properties of speech contexts of a training speaker would shift the identification of the same set of sibilants from a test speaker. Figure 1 sketches the design of Experiment 1. The experiment has three conditions, each containing two phases: a training phase and a test phase. In the training phase, participants in all conditions were exposed to Female A's spoken words. Then, in the test phase, they completed a phoneme categorization task with Female B's /s-ʃ/ continuum. Participants were randomly assigned to one of the three conditions – a *raised vowel formant* condition, a *normal vowel formant* condition, and a *lowered vowel formant* condition. The three conditions shared identical stimuli of Female B's speech in the test phase, and identical sibilants in Female A's speech in the training phase. However, the vowel formants of Female A's lexical contexts would be raised, lowered, or normal, depending on the specific condition.

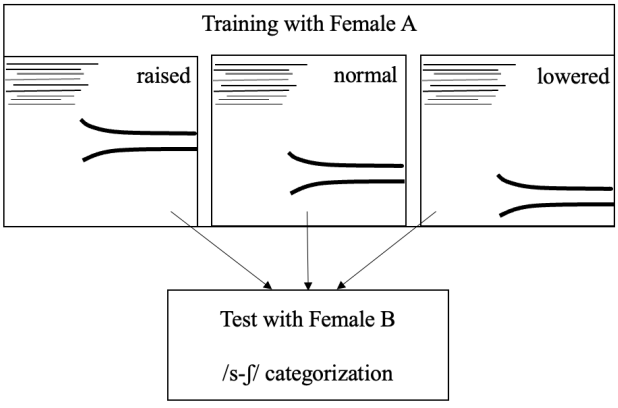

Figure 1: The design of Experiment 1

The planned analysis was to compare across conditions and evaluate whether the differences in contextual vowel formants

in the training phase would affect listeners' categorization of /s-ʃ/ with a different speaker in the test phase.

Predictions

The two competing hypotheses we set up in the previous section make different predictions.

A *raw-distribution* hypothesis predicts that participants in different experimental conditions would have similar categorization results, since they were exposed to identical acoustic signals of /s-ʃ/ of Female A in the training phase.

By contrast, a *normalized-distribution* hypothesis predicts that participants in the three conditions would exhibit categorization patterns that are different from each other, in ways consistent with the vowel formant manipulation on the word frames in each condition. Participants in the *lowered vowel formants* condition were exposed to lower-vowel-formant frames, which entailed relative higher frequencies of the spectral energy for sibilants in Female A's acoustic space. Therefore, these participants should expect higher spectral energy frequency in the identification of /s/ and report fewer perception responses of /s/ along the continuum, compared with participants in the *normal vowel formants* condition. Similarly, participants in the *raised vowel formants* condition heard higher vowel formants and therefore relative lower spectral energy frequencies of the sibilants in the speaker's acoustic space. Therefore, they should expect lower spectral energy frequencies for the perception of /s/ and report more /s/ along Female B's sibilant continuum, compare with participants in the *normal vowel formant* condition. Taken together, the predicted ranks across conditions in terms of /s/ report rate is *raised* > *normal* > *lowered*.

Method

Participants 45 adult participants (20 men and 25 women) were recruited from Prolific, a subject pool for online experiment (Palan & Schitter, 2018), to complete Exp 1 online. Five of them were under the age of 20, eight of them were above 40, and the rest of them fell into the range of 20-40 years old. They were self-reported to be native English monolinguals and have no hearing disorders.

Materials Two female native American English speakers recorded stimuli for the experiments presented here. They will be referred to throughout as Female A and B. All the stimuli were recorded in a sound-proofed recording booth, with a Yeti microphone at a sampling rate of 44.1 kHz.

Construction of training stimuli The training stimuli were drawn from Female A's speech. They consisted of 51 words. 17 of them contained *s* word-medially, 17 contained *sh* word-medially, and the remaining 17 did not contain *s* or *sh* anywhere in the word. The *s*-containing words and *sh*-containing words were matched in lexical frequency as determined using the SUBTLEX corpus (Brysbaert & New, 2009) FREQcount measure. All the words were normalized to 70 dB, and they were used as the training stimuli in one of the experimental conditions – the *normal* vowel formant condition.

Then we manipulated the vowel formants of the 51 stimuli in Praat to construct training stimuli for the two other experimental conditions. Stimuli in the *raised* vowel formant condition were created by multiplying all vowel formants by a factor of 1.2, and stimuli in the *lowered* vowel formant condition were created by multiplying vowel formants by a factor of 0.8. Sibilants were cut off from the critical items before manipulation and spliced back afterward to ensure that they were not manipulated. The formants of the filler stimuli were also scaled in ways consistent with the critical stimuli in their condition, and manipulation was performed throughout the whole word. These stimuli were also normalized to 70 dB.

Construction of test stimuli The test stimuli were 51 spoken words produced by Female B. They consisted of 35 critical items in word frames of /s/- /ʃ/ minimal pairs and 16 filler words that had no /s/ or /ʃ/. To generate the critical items, we first made a five-step s-sh continuum with Female B's typical /s/ and /ʃ/ sounds by blending the two sounds at five steps of proportion ratios, varying from 0.3 /s/ 0.7 /ʃ/ to 0.7 /s/ 0.3 /ʃ/ with an increase of 0.1 /s/ and a decrease of 0.1 /ʃ/ at each interval. Then the five instances were each spliced onto seven word frames that form minimal pairs of /s/ and /ʃ/. The word frames were made from the spoken words of *sign*, *same*, *seat*, *self*, *shake*, *shell*, and *shy*. We removed the original sibilant onsets of these words and concatenated each of them to the five steps of sibilants. These stimuli were also normalized to 70 dB.

Procedure All experiments were programmed and implemented through the PCIbex online experimental platform (Zehr & Schwarz, 2018). Participants in all three conditions completed a training block on Female A's spoken words, followed by a test block on Female B's spoken words. In each block, listeners needed to complete 51 trials, where they listened to a spoken word once and needed to choose from two written options which word they heard. For the training trials, the two options consisted of the correct word and a foil that was orthographically similar but not contrastive on the critical sound (e.g., *voca<u>t</u>ion* vs. *vaca<u>t</u>ion*). For test trials, however, the two options were always contrastive on the critical sound (e.g., *<u>s</u>ame* vs. *<u>sh</u>ame*). The trial order was randomized within blocks for each participant, and the order between the two options was randomized for each trial.

Results

Analyses were conducted using the R Statistical environment (R Core Team, 2014); linear models were run using the lme4 library (Bates, Mächler, Bolker, & Walker, 2015), and plots were created using ggplot (Wickham, 2016).

Figure 2 shows the means and standard errors of the /s/ identification rates at each fricative step in three experimental conditions. Generally, participants in all three conditions

showed more /s/ responses as the proportion of /s/ mixed in the test stimulus increased. Moreover, although listeners were exposed to identical sibilants in the training stage, their categorization boundary of /s-ʃ/ shifted with different manipulations of vowel formants of the training stimuli. Listeners who listened to stimuli with vowel formants shifted downwards in the spectrum reported the perception of /s/ less frequently, and listeners who listened to stimuli with vowel formants shifted upwards in the spectrum reported more frequent perception of /s/, compared with those who were exposed with unmanipulated vowels.

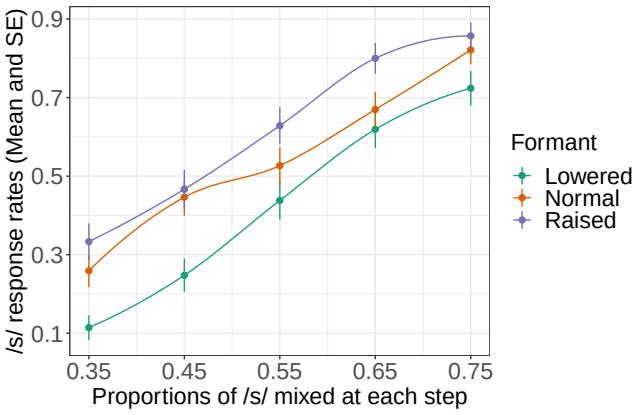

Figure 2: The means and standard errors of /s/ response rates at each step in three experimental conditions

A mixed-effects logistic regression model was fitted to predict the response of each trial (*/ʃ/*=0, */s/*=1). The model included Condition (ternary, treatment coded, baseline: Lowered), Step (numeric, scaled, and centered), and Phoneme (the original consonant produced with the word frames of the test stimuli, sum-coded) as fixed effects, Step $\times$ Condition as an interaction term, and Step by Participant and Step by Word as random slopes. The model showed a significant effect of Step ($\beta = 1.82, SE = 0.24, p < 0.001$) and Phoneme ($\beta = 0.75, SE = 0.24, p = 0.001$). Condition is significant for the Raised Formant condition ($\beta = 1.54, SE = 0.68, p < 0.02$) but not the Normal Formant condition ($\beta = 0.95, SE = 0.66, p = 0.15$). The model revealed no significant interactions. We then evaluated the difference between the Normal Formant condition and the Raised Formant condition by re-coding the latter as the baseline level and re-ran the model. Again, we found no significant effect of the Normal Formant condition ($\beta = -0.59, SE = 0.66, p = 0.38$).

Discussion

The result of Experiment 1 lends some support to the *normalized-distribution* hypothesis of the perceptual generalization of spectral cues. We found that shifting vowel formants of the word frames of the training speaker leads to a shift of sibilant perception boundary for the test speaker. This means that altering vowel formant frequencies has yielded a different relative distribution of spectral energy frequencies of sibilants in the learning process. The hypothesis is also supported by the finding that the /s/ response rates rank as predicted across the three conditions, although a statistical difference was only detected between the two most extreme experimental conditions (the Lowered and the Raised conditions). This can be partially attributed to the lack of power since there were only 15 participants in each condition. More data therefore is needed to evaluate whether our current finding is robust.

Experiment 2: Speech rate normalization

Overview

Experiment 2 investigates whether manipulating temporal properties of word frames of a training speaker would shift the identification of the same set of stops from a test speaker. Figure 3 illustrates the design of Experiment 2. Similar to Experiment 1, the experiment has three conditions, each containing two phases. In the training phase, participants in all conditions were exposed to Female A's spoken words that contained /t/ and /d/; then they completed a phoneme categorization task with female B's /t-d/ continuum in the test phase. Participants were randomly assigned to one of the three conditions, namely, a *shortened* frame condition, a *normal* frame condition, and a *lengthened* frame condition. Participants across the three conditions were tested on the same set of stimuli produced by Female B. Crucially, they were exposed to identical critical stops produced by female A. However, the word frames in which stops were situated were temporally compressed, expanded, or unaltered, depending on the specific condition participants were in.

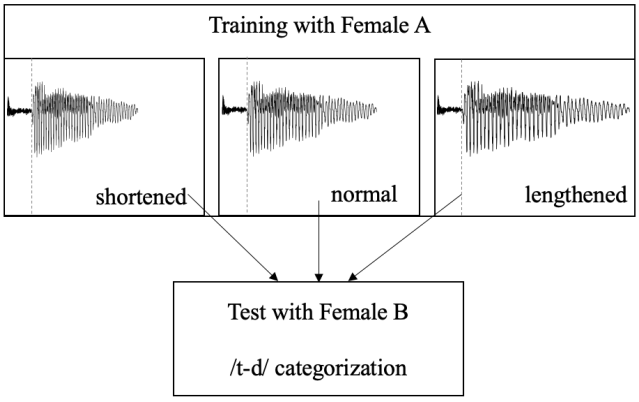

Figure 3: The design of Experiment 2

The planned analysis was to compare across conditions to evaluate whether manipulating the temporal properties of speech contexts of Female A in the training phase would affect listeners' categorization results of /t-d/ with Female B in the test phase.

Predictions

The two competing hypotheses we set up make different predictions about the results. A *raw-distribution* hypothesis

predicts that participants in different experimental conditions would have similar categorization results, since they were exposed to identical acoustic signals of /t-d/ of Female A in the training phase. In contrast, a *normalized-distribution* hypothesis predicts that participants in the three conditions would exhibit categorization patterns that differ from each other in the following way: participants in the *shortened frame* condition were exposed to shorter word frames, and therefore larger proportions of VOT of /t-d/ in the frames, which would make them expect longer VOTs for /t-d/. As a result, these participants were predicted to show a lower probability of reporting a /t/ along the same VOT continuum than participants in the *normal frame* condition. Similarly, participants in the *lengthened frame* condition heard longer word frames and therefore smaller proportions of VOT of /t-d/ relative to the frames, which would make them expect smaller VOT values for /t-d/. As a result, they would be more willing to report /t/ at a smaller VOT step along the same VOT continuum, in comparison with participants in the *normal frame* condition.

Method

Participants 45 participants (23 men and 22 women) were recruited from Prolific (Palan & Schitter, 2018) to complete the experiment. Five of them were under the age of 20, seven of them were above 40, and the rest of them fell into the range of 20-40 years old. They self-reported to be native English monolinguals and have no hearing disorders.

Materials The stimuli used in Experiment 2 were produced by the same two female speakers whose speech was used in Experiment 1. The recording procedure was also identical to Experiment 1.

Construction of training stimuli The training stimuli were manipulated from 51 spoken words of Female A. They consisted of 17 words that contained /t/ word-medially, 17 words that contained /d/ word-medially, and 17 words that did not contain /t/ or /d/ anywhere in the word. The 51 words were further manipulated to create training stimuli for three context duration conditions: a *lengthened* condition, a *shortened* condition, and a *normal* condition. In these conditions, the target stops in the critical words remained unaltered, while the remaining word frame was either expanded by a factor of 1.7 (*lengthened*), compressed by a factor of 0.7 (*shortened*), or scaled by 1.0 (*normal*). Duration interpolation was implemented using the PSOLA algorithm in Praat to avoid potential signal mismatch or missing material at target boundaries associated with splicing. The duration of filler words was manipulated consistently with the critical items in each condition, except that the manipulation was implemented throughout the whole word due to the absence of target phonemes. All the stimuli were normalized to 70 dB.

Construction of test stimuli The test stimuli were all manipulated from Female B's speech. They were 35 critical items in lexical frames of /t-d/ minimal pairs and 16 filler words that had no /t/ or /d/. The critical items were manipulated from seven /t/-initial spoken words that form a different word with /t/ substituted by /d/ (i.e., *tear*, *tie*, *tip*, *toes*, *touch*, *town*, *time*). For each word, we manipulated their onset into a five-step /t-d/ continuum by temporally scaling the VOT proportion of the /t/ onset to be 0.2, 0.4, 0.6, 0.8, and 1.0 of its original length. Again, temporal compression was conducted using the PSOLA algorithm in Praat and its graphical user interface for duration interpolation. All stimuli are normalized to 70 dB.

Procedure The experiment was implemented through the PCIbex platform (Zehr & Schwarz, 2018). Participants in all conditions completed a training block on Female A's spoken words and then a test block on Female B's spoken words (51 trials in each block). Same as Experiment 1, listeners heard each spoken word once and needed to choose from two written options which word they heard. The two options in a training trial never contrasted on the critical phoneme, whereas the two options in a test trial always contrasted on the critical phoneme. The trial order was randomized within blocks for each participant.

Results

Figure 4 shows the means and standard errors of the /t/ identification rates at each VOT step in the three experimental conditions.

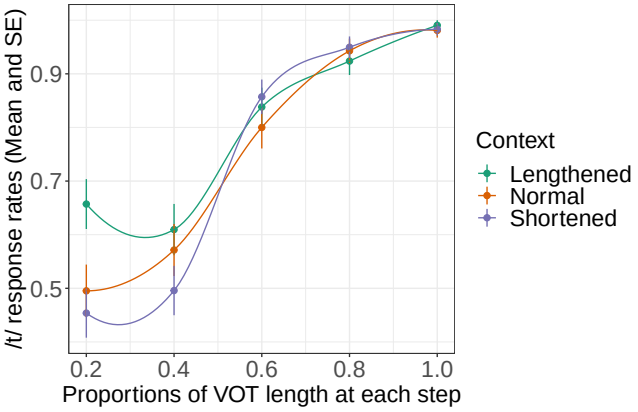

Figure 4: The means and standard errors of /t/ response rates at each step in three experimental conditions

In general, participants showed a strong tendency to identify /t/ throughout the continuum such that the /t/ identification rate has already reached 50% at the first step. Then, as VOT became longer at larger steps, participants increased their /t/ responses. Between-condition differences in /t/ responses were most clearly shown at the first two VOT stops: Participants exposed to training stimuli with shorter word frames perceived the fewest /t/ instances in the test phase; similarly, participants exposed to training stimuli with longer word frames perceived the most /t/ responses; participants exposed with training stimuli with unaltered

word frames duration reported a /t/ response rate that lied in between. This difference between conditions disappeared at the third to fifth VOT step.

A mixed-effects logistic regression model was fitted to predict the response of each trial (/d/=0, /t/=1) with Condition (treatment coded, baseline: Lengthened), Step (numeric, scaled, and centered) in a two-way interaction and the random slopes of Step by Participant and Step by Word. The model showed a significant effect of Step ($\beta = 1.18, SE = 0.27, p < 0.001$) but no effect of Condition for Normal ($\beta = -0.21, SE = 0.29, p = 0.48$) and Shortened ($\beta = -0.16, SE = 0.28, p = 0.56$) conditions. However, the interaction between Step and Condition (Shortened) turned out to be significant ($\beta = 0.54, SE = 0.21, p = 0.01$), which indicated that the smaller proportions of /t/ responses on the first two steps in the Shortened condition in Figure 4 drove the categorization slope to be sharper than that of the Lengthened condition.

Discussion

The result of Experiment 2 provides preliminary evidence for the *normalized-distribution* hypothesis of the perceptual generalization of temporal cues, by showing the predicted ranking of /t/ response rates across conditions according to this hypothesis. However, the supporting pattern was only found at the first two steps of the VOT continuum. This might be because the whole /t-d/ continuum in the test phase was heavily /t/-biased, such that stop instances at the third to fifth VOT steps were relatively unambiguous. Unlike other stop voicing identification studies where VOTs were manipulated non-linearly across continuum steps (Toscano & Lansing, 2019), we kept the increment of VOT amount fixed throughout the continuum, which might have contributed to this bias.

This bias could also be attributed to the /t/-biasing cues revealed by the critical phonemes and the word frames in the test trials of Experiment 2. For the critical phonemes, the aspiration noises remained somewhat perceivable after manipulation, adding an extra favoring cue to the perception of /t/. For the word frame, all the word frames used for the test phase were /t/-initial, whereas in Experiment 1, the seven word frames used for the test phase were four /s/-initial (*sign*, *same*, *seat*, *self*) and three /ʃ/-initial (*shake*, *shell*, *shy*). Both factors might have contributed to the predominant /t/ responses in Experiment 2.

General Discussion

Through two experiments, we show that changing the acoustic cues in the contextual materials would affect how listeners learn and generalize the cue distributions of a critical phonemic contrast. Similar results have been observed for investigations on the integration of both spectral cues and temporal cues. Experiment 1 showed that raising the spectral energy frequencies of contextual information would cause listeners to learn and generalize relatively lowered spectral energy frequencies for sibilants, and lowering the contextual frequencies would lead to the learning and generalization of higher sibilant frequencies. Experiment 2 showed that lengthening contextual materials would cause listeners to pick up and generalize relatively shortened VOT distributions for stops, and shortening contextual materials would lead to the learning and generalization of relatively larger proportions of VOTs in the temporal dimension. Taken together, these results lend some support for an account where listeners track and generalize speaker-normalized distributions on the fly.

One piece of supporting evidence we found for a normalized-distribution account is that identification rates across conditions were ranked as predicted robustly in both experiments. However, we did not detect a significant difference between any two of the three conditions in each experiment. Significant differences were only found between conditions with opposite manipulations. This raised a follow-up question: whether the small effect size was a by-product of under-powered subject samples or a part of the nature of this mechanism. It also remains unknown to what degree the effect size would be modulated by the relationship between the acoustic distributions of the target phonemes and their lexical contexts in speech production. Does the magnitude of normalization that listeners implement correspond to the ratios between the target acoustics and the contextual acoustics that listeners pick up in their real-world language experience?

Although we did find evidence for the interference of speech normalization in perceptual learning, we cannot rule out the possibility that perceptual generalization mechanisms based on raw distributions may occur when tasks are different. More research is needed to identify the scope and extent where each of these different mechanisms plays a role. Finally, our results reveal an impetus for future speech perception models to incorporate normalization mechanisms (Miller, 1989; Patterson & Irino, 2013; Johnson & Sjerps, 2021, among others) into perceptual learning models.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Brysbaert, M., & New, B. (2009, Nov). 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. *Behavioral Research Methods*, *41*(4), 977-90. doi: 10.3758/BRM.41.4.977

Diehl, R. L., Souther, A. F., & Convis, C. L. (1980). Conditions on rate normalization in speech perception. *Perception & Psychophysics*, *27*(5), 435–443.

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, *21*(11), 1664–1670.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception &*

psychophysics, *67*(2), 224–238.

Johnson, K. (1990). Contrast and normalization in vowel perception. *Journal of Phonetics*, *18*(2), 229–254.

Johnson, K. (2018). Vocal tract length normalization. *UC Berkeley PhonLab Annual Report*, *14*(1).

Johnson, K., & Sjerps, M. J. (2021). Speaker normalization in speech perception. *The handbook of speech perception*, 145–176.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*(4), 359–384.

Kent, R. D. (1993). Vocal tract acoustics. *Journal of Voice*, *7*(2), 97–117.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, *122*(2), 148.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive psychology*, *51*(2), 141–178.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic bulletin & review*, *13*(2), 262–268.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15.

Kurumada, C., & Roettger, T. B. (2021). Thinking probabilistically in the study of intonational speech prosody. *PsyArXiv. May*, *31*.

Lai, W. (2021). *The online adjustment of speaker-specific phonetic beliefs in multi-speaker speech perception*. Unpublished doctoral dissertation, University of Pennsylvania.

Lammert, A. C., & Narayanan, S. S. (2015). On short-time estimation of vocal tract length from formant frequencies. *PloS one*, *10*(7), e0132193.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *The Journal of the Acoustical society of America*, *85*(5), 2114–2134.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology*, *47*(2), 204–238.

Palan, S., & Schitter, C. (2018). Prolific. ac — a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Patterson, R. D., & Irino, T. (2013). The role of size normalization in vowel recognition and speaker identification. In *Proceedings of meetings on acoustics ica2013* (Vol. 19, p. 060038).

Port, R. F. (1979). The influence of tempo on stop closure duration as a cue for voicing and place. *Journal of Phonetics*, *7*(1), 45–56.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Reby, D., & McComb, K. (2003). Vocal communication and reproduction in deer. *Advances in the Study of Behavior*, *33*, 231–264.

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

Smith, D. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America*, *118*(5), 3177–3186.

Strand, E. A., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In *Konvens* (pp. 14–26).

Summerfield, Q. (1975). How a full account of segmental perception depends on prosody and vice versa. In *Structure and process in speech perception* (pp. 51–68). Springer.

Tamminga, M., Wilder, R., Lai, W., & Wade, L. (2020). Perceptual learning, talker specificity, and sound change. *Papers in Historical Phonology*, *5*, 90–122.

Toscano, J. C., & Lansing, C. R. (2019). Age-related changes in temporal and spectral cue weights in speech. *Language and speech*, *62*(1), 61–79.

Van der Zande, P., Jesse, A., & Cutler, A. (2014). Cross-speaker generalisation in two phoneme-level perceptual adaptation processes. *Journal of Phonetics*, *43*, 38–46.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. Retrieved from `http://ggplot2.org`

Wong, P. C., & Diehl, R. L. (2003). Perceptual normalization for inter-and intratalker variation in cantonese level tones. *Journal of Speech, Language, and Hearing Research*, *46*(2), 413–421.

Xie, X., Buxó-Lugo, A., & Kurumada, C. (2021). Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition*, *211*, 104619.

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *The Journal of the Acoustical Society of America*, *143*(4), 2013–2031.

Zehr, J., & Schwarz, F. (2018). Penncontroller for internet based experiments (Ibex). *URL https://doi. org/10.17605/OSF. IO/MD832*.