THE RELATIONSHIP BETWEEN PERCEPTUAL SIMILARITY JUDGMENTS AND VOT CONVERGENCE IN A SHADOWING TASK

Jessamyn Schertz[1,2], Melissa Paquette-Smith[3], Elizabeth K. Johnson[1,2]

University of Toronto Mississauga, University of Toronto, University of California Los Angeles
jessamyn.schertz@utoronto.ca, paquettesmith@psych.ucla.edu, elizabeth.johnson@utoronto.ca

ABSTRACT

This work examines the relationship between phonetic and perceptual metrics for convergence in a shadowing task, focusing on whether listeners are sensitive to acoustic convergence in voice onset time (VOT) when judging similarity of word productions across talkers. Adults and 6-year-olds shadowed model talkers with artifically extended VOT; VOTs of shadowed vs. baseline productions were compared to assess phonetic convergence. Listeners completed a discrimation task on baseline and shadowed tokens produced by shadowers who showed systematic convergence in VOT. The magnitude of VOT convergence correlated with listeners' choice of the shadowed production as more similar to the model than the baseline production. Listeners were less accurate in a follow-up study, where VOT of the shadowed productions was equalized to the same value as baseline (i.e. "removing" the convergence on the dimension of VOT), suggesting that VOT plays an independent role in similarity judgments.

Keywords: shadowing; VOT; perceptual vs. acoustic correlates of phonetic convergence.

1. INTRODUCTION

A large body of work has demonstrated phonetic convergence during shadowing tasks: when repeating words directly after a recorded "model" talker, the productions of the "shadowers" become more similar to the model's productions (see review in [10]). However, it is not well understood how different acoustic dimensions (e.g. VOT, f0) contribute to listeners' perceptions of convergence. In this work, we examine the relationship between convergence in VOT and listeners' similarity judgments, and we test whether VOT plays an independent role in perception by comparing performance across two experiments with VOT differences present vs. absent.

Past work has used both "acoustic" and "perceptual" measures to quantify convergence in shadowing. Comparisons of baseline vs. shadowed productions suggest that convergence occurs on multiple acoustic dimensions, including vowel formants [2], duration [8, 10], f0 [1], and VOT [3, 12, 11]. However, these effects are often complex and inconsistent, with substantial individual differences in effects [10]. Furthermore, focusing on a small subset of acoustic dimensions can lead to an incomplete picture of convergence, particularly given that the specific nature of adaptation can differ substantially across shadowers and model talkers [10].

Pardo et al. (2017) [10] emphasized the importance of considering listeners' perceptual judgments of similarity between the shadower and model talker: this more holistic measure integrates multiple phonetic dimensions not necessarily captured in acoustic analyses. At the same time, considering listeners' judgments alone risks leaving important gaps in our understanding of convergence. Using judgments as a "gold standard" to quantify similarity in speakers' productions is only valid insofar as listeners' judgments faithfully reflect properties of the signal. If not, then using similarity judgments could fail to capture systematic patterns of convergence, either because listeners may selectively attend to some dimensions over others, or because the speakers' modifications, although systematic, are too small to be perceptible by listeners.

A comprehensive understanding of convergence requires taking both acoustic and perceptual measures into account, and looking into the relationship between them. To this end, several studies have used both metrics and examined to what extent listeners' judgments are correlated with measurable phonetic differences. The most comprehensive of these tested to what extent changes in three phonetic properties of vowels (duration, formants, and f0) predicted listeners' judgments of similarity of productions of 92 talkers shadowing 12 different talkers [10]. When each of the acoustic measures was considered independently, there was not consistent convergence. However, the extent of phonetic convergence was predictive of listeners responses (see also [9, 13]). This was taken as evidence that listeners are sensitive to similarity along these acoustic dimensions, and that therefore their judgments can be used as a holistic metric to assess convergence.

Correlations between convergence on a given acoustic dimension and listeners' judgments are consistent with the idea that listeners use that dimension to inform their judgments. However, when using natural productions, it is likely that other dimensions are co-varying with the target dimension. In this situation, it is impossible to know which dimensions are actually driving listeners' responses. For example, in a hypothetical experiment, tokens with greater convergence in f0 might be more likely to elicit greater similarity judgments by listeners. However, these same tokens might also show convergence along other dimensions, e.g. vowel formants, such that the perceptual results could be driven by either f0 or formants. In order to test whether listeners are sensitive to a given dimension, it must be manipulated independently, with all other acoustic information held constant.

1.1. The current work

While several studies have found convergence in VOT, no previous work has examined whether listeners are sensitive to this dimension when assessing similarity. Given that the average extent of covergence on this dimension is very small (around 10 ms or less [3, 11, 12]), it is questionable whether this is a dimension listeners would attend to in assessing similarity. We investigate the relationship between VOT convergence and perceptual judgments first as in previous work, via a correlation analysis between listeners' similarity judgments on an XAB discrimination task and the extent of phonetic convergence. We then test to what extent VOT exerts an *independent* effect on listeners' judgments by playing listeners the same stimuli after removing the VOT differences between baseline and shadowed productions. If listeners use VOT in assessing similarity, we expect to see 1) a positive correlation between the extent of VOT convergence and listeners' accuracy, and 2) a decrease in accuracy when VOT differences are artificially removed. An ancillary question, based on the fact that children's VOTs for voiceless stops are more variable than adults' [4], is whether listeners rely less on VOT when assessing similarity of child vs. adult productions.

2. METHODS

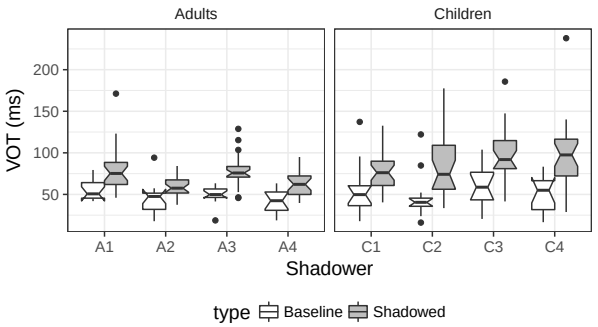
2.1. Imitation task

The stimuli for the discrimination tasks are a subset of data from a larger study [7] in which 6-year-olds and adults shadowed one of two model talkers. Participants completed a baseline phase, where they were presented with pictures of each of the target words and asked to name them, followed by a shadowing phase, where they saw a picture, heard a model talker's production of the relevant word, then repeated the word. VOTs of the model talkers had been systematically extended in stop-initial words. The exact VOT value varied by word, but the values across the two model talkers for each word were set to be identical, with an average VOT of 120ms (50ms above the average value of the natural productions). This manipulation was done in order to elicit convergence along the VOT dimension [5, 6].

Acoustic measures: VOT was measured from burst release until onset of voicing in the following vowel. For each shadowed token, we calculated a difference score, as in previous work, to quantify convergence [8, 10]. This measurement, which we call ΔVOT (1), indicates how much closer in VOT the shadowed production is to the model talker as compared to the baseline production. Larger ΔVOT is interpreted as more convergence in VOT.

(1) $$\Delta VOT = (VOT_{Shad} - VOT_{Mod}) - (VOT_{Base} - VOT_{Mod})$$

Figure 1: Distributions of baseline and shadowed VOT values by the eight shadowers used for stimuli.



2.2. Discrimination task

2.2.1. Experiment 1

Stimuli: Baseline and shadowed productions of four children and four adults who showed the greatest convergence in VOT were used as stimuli. Since the purpose of this study was to determine whether listeners were sensitive to convergence along VOT, we wanted to ensure that the shadowers did indeed show convergence[1]. Stimuli consisted of 1 baseline and 2 shadowed tokens of /p/-initial words: where this full set was not available due to mispronunciation, the whole set was omitted. Fig. 1 shows the distribution

of VOTs of baseline and shadowed tokens for each of the shadowers used in the stimuli.

Participants and procedure: 33 native English listeners completed an XAB discrimination task. In each trial, listeners heard a model talker's production of a word (X), followed by a baseline and one of the shadowed versions (A and B, order of baseline and shadowed tokens were randomized) of the same word. Listeners were asked to indicate which of the two final words sounded most like the first word (i.e. the model). There were four blocks, grouped by model talker and shadower age (e.g. in one block, the model was always Talker A and baseline/shadowed productions came from the two adults who had shadowed that talker). The order of blocks, and the order of trials within each block, were randomized for each participant. There were 200 total trials (8 talkers * 12-13 words * 2 repetitions), and the task took about 15 minutes.

2.2.2. Experiment 2

Experiment 2 was identical to Experiment 1, but the VOT of each shadowed token was manipulated to be equivalent to the baseline token of the same word spoken by the same shadower, using the PSOLA algorithm in Praat. In other words, for a given XAB trial, the VOT values of A and B (the shadowed and baseline tokens) were identical. 27 native English listeners participated.

2.3. Statistical analysis

We used two logistic mixed-effects models to test the factors influencing listeners' similarity judgments. Our response variable for both models is listeners' choice of the shadowed (vs. baseline) token as more similar to the model, which we refer to as "accuracy."

To test whether accuracy was related to the magnitude of VOT convergence, we entered the data from Experiment 1 into a model predicting accuracy from ΔVOT[2]. To test for an independent effect of VOT, we built a second model, using data from both experiments, predicting accuracy from Experiment (1 vs. 2)[3]. Both models included a fixed effect of Shadower Age (Adult vs. Child), as well as its interaction with the other fixed effects, to test the hypothesis that listeners place more reliance on VOT when listening to adults than to children. ΔVOT was centered and converted to z-scores for analysis, while categorical factors were simple-coded (-.5, .5), such that all coefficients represent the (change in) log-odds of a "shadowed" response across all levels of the other condition(s). The full random effects structure justified by the design was included.

3. RESULTS

In Experiment 1, listeners chose the shadowed token on average 69% of the time. Fig. 2 shows the proportion of shadowed responses for each trial, as a function of the extent of convergence (ΔVOT) on that trial. Table 1 shows the results of the model predicting listeners' responses from ΔVOT and Shadower Age. The significant positive coefficient for the intercept indicates that listeners' accuracy was above chance, while the significant effect of ΔVOT indicates that the likelihood of an accurate response increases as ΔVOT increases (as modeled by the best-fit regression line in Fig. 2). Neither shadower age nor its interaction with ΔVOT were significant, so there is no evidence that listeners showed different accuracies, or more sensitivity to convergence along the VOT dimension, for child vs. adult shadowers.

Figure 2: Accuracy as a function of ΔVOT in Experiment 1. Each dot represents the proportion of accurate responses (i.e. choice of the shadowed token) for one trial. Larger values of ΔVOT represent more convergence toward the model talker. The regression line shows the best-fit logistic curve.

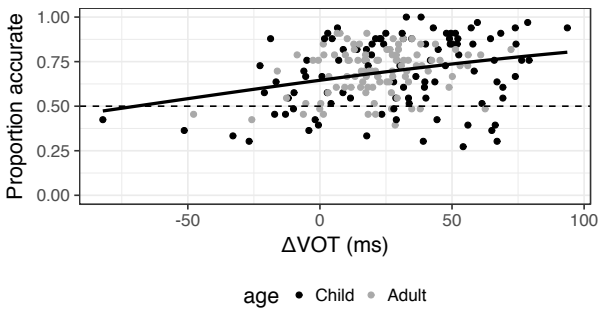


Table 1: Results of a mixed-effects logistic regression model predicting choice of shadowed response from ΔVOT and shadower age.

	β	SE	z	p
Intercept	0.79	0.16	4.99	$< .001$*
ΔVOT	0.12	0.59	2.04	0.041*
Shadower age	0.07	0.26	0.298	0.766
ΔVOT * Age	-0.05	0.12	-0.45	0.657

This relationship suggests that listeners may be sensitive to VOT. However, these results could also be because those tokens with larger values of ΔVOT also exhibited convergence on other dimensions, which the listeners could use to inform their judg-

ments. To determine whether VOT plays an independent role, we compared the results of Experiment 1 to Experiment 2, where VOT differences were removed. Fig. 3 shows the proportions of accurate response, broken down by Experiment and Shadower Age. Overall, accuracies are slightly lower for Experiment 2 in Experiment 1, although the difference is very small (with 69% accuracy overall in Experiment 1, compared to 65% in Experiment 2).Table 2 shows the results of the model predicting listeners' accuracy from Experiment and Shadower Age. We again see a significant intercept, indicating that accuracy overall was above chance. In response to our primary research question, we see a negative coefficient for Experiment, indicating significantly lower accuracy for Experiment 2. However, even with this slight decrease, overall accuracy for Experiment 2 is still well above chance. As before, Age was not significant, nor was the interaction between Age and Experiment.

Figure 3: Distribution of listeners' mean accuracy rates across Experiment 1 and Experiment 2, broken down by performance on trials with child vs. adult shadowers. Error bars represent 95% confidence intervals of by-listener means.

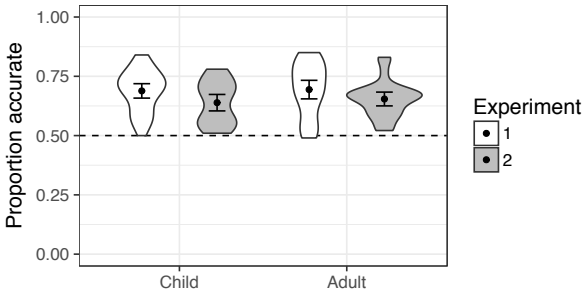


Table 2: Results of a mixed-effects logistic regression model predicting choice of shadowed response as a function of Experiment (2 vs. 1) and Shadower Age.

	β	*SE*	z	p
Intercept	0.79	0.15	5.15	$< .001$*
Experiment	-0.23	0.11	-2.07	0.038*
Shadower age	0.06	0.25	0.24	0.81
Experiment * Age	0.05	0.15	0.33	0.73

4. DISCUSSION

Our results suggest that VOT plays a role in listeners' similarity judgments, supported by the fact that the extent of VOT convergence in a shadowing task is predictive of listeners' choice of the shadowed token as more similar to the model talker than a baseline production (Experiment 1). The fact that listeners' performance decreased when the VOT of the shadowed production was equalized to the baseline value (Experiment 2) further shows that VOT plays an independent role. No differences were found in sensitivity to VOT when listening to children vs. adults.

Our findings are consistent with the idea that perceptual judgments of similarity incorporate multiple acoustic dimensions [10]. However, it is important to note that despite the fact that VOT seems to play a role in similarity judgments, this role appears to be very small, given that there was only a very small decrease in accuracy when the VOT differences were removed. Furthermore, even though all of the shadowers used in our study showed similar extents of phonetic convergence, listeners' sensitivity to VOT appeared to vary considerably by shadower (i.e. there was not a decrease in accuracy for all shadowers when VOT was removed). Therefore, the relationship between the perceptual and phonetic measures is not straightforward.

We have shown that listeners are sensitive to VOT, but the nature of this sensitivity remains to be explored. For example, it is possible that the effect of VOT may be driven by those trials with relatively large convergence effects. We are not able to explore this question systematically with the current dataset, but future work could test the threshold for which listeners are sensitive VOT, and in which circumstances (e.g. for which model talkers, and as a function of the presence and/or strength of other phonetic cues) they use it more or less.

These results demonstrate that VOT is part of the constellation of acoustic dimensions that informs listeners' similarity judgments. However, it appears that there is not a straightforward relationship between perception and VOT convergence, even in a case where the acoustics have been controlled by choosing shadowers who showed similar extents of VOT convergence. Therefore, listeners' judgments cannot be used as a replacement for acoustic measures in assessing convergence, and it cannot be assumed that listeners' perception faithfully reflects properties of the signal. Using manipulations such as the one used here can help to isolate specific dimensions and inform our understanding of the relationship between perceptual and acoustic measures of similarity.

5. REFERENCES

[1] Babel, M., Bulatov, D. 2012. The role of fundamental frequency in phonetic accommodation. *Language and Speech* 55(2), 231–248.

[2] Babel, M., McAuliffe, M., Haber, G. 2013. Can mergers-in-progress be unmerged in speech accommodation? *Frontiers in Psychology* 4.

[3] Fowler, C. A., Brown, J. M., Sabadini, L., Weihing, J. 2003. Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language* 49(3), 396–413.

[4] Lowenstein, J. H., Nittrouer, S. 2008. Patterns of acquisition of native voice onset time in English-learning children. *The Journal of the Acoustical Society of America* 124(2), 1180–1191.

[5] Nielsen, K. 2011. Specificity and abstractness of VOT imitation. *Journal of Phonetics* 39(2), 132–142.

[6] Nielsen, K. 2014. Phonetic imitation by young children and its developmental changes. *Journal of Speech, Language, and Hearing Research* 57(6), 2065–2075.

[7] Paquette-Smith, M. 2018. *The Effect of Accent Exposure on Social Cognition and Language Acquisition in Early Childhood*. PhD thesis The University of Toronto.

[8] Pardo, J. S. 2013. Measuring phonetic convergence in speech production. *Frontiers in Psychology* 4.

[9] Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., Lewandowski, E. 2013. Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language* 69(3), 183–195.

[10] Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J. 2017. Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics* 79(2), 637–659.

[11] Sanchez, K., Miller, R. M., Rosenblum, L. D. 2010. Visual influences on alignment to voice onset time. *Journal of Speech, Language, and Hearing Research* 53(2), 262–272.

[12] Shockley, K., Sabadini, L., Fowler, C. A. 2004. Imitation in shadowing words. *Perception & Psychophysics* 66(3), 422–429.

[13] Walker, A., Campbell-Kibler, K. 2015. Repeat what after whom? exploring variable selectivity in a cross-dialectal shadowing task. *Frontiers in Psychology* 6.

[1] Half of the participants (A1, A2, C1, C2) shadowed a Canadian English model talker, while the other half shadowed an Australian English talker. We did not have different predictions for listeners' use of VOT for the different accents, nor did we see any different patterning of results, so we do not include this factor in the analysis below

[2] `glmer(response~ΔVOT*Age+(ΔVOT+Age||listener)+(ΔVOT+Age||shadower)+(1|word))`

[3] `glmer(response~Exp*Age+(Age||listener)+(Exp+Age||shadower)+(Exp||word))`