Acoustic convergence to an unfamiliar language and an unfamiliar accent in shadowed speech

Amy E. Hutchinson[1], Alexis Zhou[2], Yuhyeon Seo[2], and Olga Dmitrieva[2]

[1]Boston University, [2]Purdue University
ahut@bu.edu, atews@purdue.edu, seo86@purdue.edu, odmitrie@purdue.edu

ABSTRACT

The present study investigated the propensity of native speakers of English (two groups of 15 participants each) to spontaneously imitate a model who spoke a different language (Russian) or an accented version of their own language (Russian-accented English). The study consisted of four phases: baseline, exposure, shadowing, and post-test. The change in the voice onset time (VOT) of word-initial voiceless stops from the baseline to shadowing to post-test was assessed across the two groups. The results showed that participants in the Russian condition significantly shortened the VOT of their voiceless stops during the shadowing phase, indicating convergence towards the native speaker model. The values, however, returned to baseline levels in the post-test, indicating that imitation of the phonetic properties of Russian speech did not extend to participants' native English. Participants in the accented English condition did not converge towards the shorter VOT of the model during shadowing or post-test phases.

Keywords: Acoustic convergence, Russian, English, accented speech, shadowing

1. INTRODUCTION

Previous research has demonstrated the ability of talkers to spontaneously adjust the acoustic properties of their speech to sound more like their interlocutor or speech model. This phenomenon is known as phonetic convergence or phonetic accommodation and has been shown to affect a wide variety of acoustic properties including segmental properties, such as voice onset time (VOT) and vowel quality [1]–[5], as well as suprasegmental properties, such as intonation contour and voice quality [4], [6]–[8]. A majority of previous work examined acoustic convergence between two native speakers of the same language, while less is known about convergence to accented speech and/or unfamiliar languages. The present study aims to address this gap by investigating acoustic convergence during shadowing of an unfamiliar accent and an unfamiliar language.

Acoustic convergence is often approached within the framework of the Communication Accommodation Theory (CAT) [9], which postulates that talkers accommodate other talkers in verbal communication in order to reduce the social distance between themselves and their interlocuter. The present study expands the assumptions of the CAT beyond a native language and/or monolingual setting.

Additionally, by examining acoustic convergence to another language or to accented speech, we can begin to assess the potential contribution of long-term convergence to phonetic changes in the speaker's native language (phonetic drift) [10], [11]. For example, could accommodation to a new language or accented speech, such as that of a language instructor, contribute to phonetic drift in a learner's native language (L1)?

While L1 drift in individual second language (L2) learners is believed to be reversal and temporary, it has also been proposed that prolonged contact with accented versions of a language could, over time, lead to permanent sound change in the language itself [12]. This process could also be construed as L1 phonetic drift, but on a larger, societal scale. Assuming that inter-speaker accommodation would be the driving force behind such sound changes, the present study can help us evaluate the viability of this proposal.

In order to compare acoustic convergence to an unfamiliar language and an unfamiliar accent in shadowed speech, the present study examined word-initial voiceless stop shadowing by native speakers of American English who believed they were shadowing either Russian speech or Russian-accented English. Stop voicing was assessed via VOT.

VOT is the primary correlate of voicing in both English and Russian, but different VOT values are used to express phonologically equivalent categories in the two languages. English realizes voiceless stops in word-initial position as voiceless aspirated (long lag VOT >30ms), while Russian realizes voiceless stops as voiceless unaspirated (short lag VOT <30ms) [13], [14]. Therefore, if English speaking participants were to converge towards Russian or Russian-accented speech, we would expect to observe shorter VOT values for their voiceless stops.

2. METHODS

2.1. Participants

A total of 30 native speakers of American English have participated in the current study thus far (22 female and 8 male; mean age 26.63 years)[1]. Participants were self-selected volunteers recruited via flyers on the campus of a large Midwestern university. Twenty-six of the participants indicated low proficiency in at least one language other than English, with Spanish being the most prevalent (16 participants) distantly followed by French (6 participants). No participants indicated that they had knowledge of Russian.

2.2. Stimuli

The present study included three separate sets of stimuli: a baseline set, an exposure/shadowing set, and a post-test set. Each set contained eight CVC target words and 23 CVC fillers. Vowels used in the targets were limited to /ɑ/, /i/, and /ʌ/ and their closest Russian counterparts: /o/, /i/, and /a/. Across stimuli sets, the onset consonants in target words were voiceless stops: /p, t, k/.

The baseline list (Set A) consisted of English target words *pall, pod, posh, pup, tin, top, kill,* and *cod*. The post-test list (Set C) contained English target words *pot, pond, pox, pug, tip, tot, kiss,* and *cob*. The exposure/shadowing stimuli (Set B) was a set of eight English-Russian near-homophones that were recorded as Russian words by a female native speaker of Russian (Table 1). The exposure/shadowing list also contained 23 fillers that were different depending on condition. Fillers in the Russian condition were real Russian words, while fillers in Russian-accented English condition were perceived as English words produced with Russian accent.

	English	Russian	IPA
1	pot	пот (sweat)	[pot]
2	poll	пол (gender)	[pol]
3	pop	поп (priest)	[pop]
4	pun	Пан (Pan)	[pan]
5	tick	тик (teak)	[tik]
6	tock	ток (electric current)	[tok]
7	call	кол (stake)	[kol]
8	kit	кит (whale)	[kit]

Table 1: Set B: English-Russian near-homophones. IPA is representative of the model talker's pronunciation of each word (presented as either Russian or Russian-accented English to participants).

2.3. Procedures

The experiment consisted of four phases: baseline, exposure, shadowing, and post-test. During the baseline phase, all participants completed a word list reading task (baseline Set A) during which 8 English targets and 23 fillers were recorded twice. Items were presented one by one in a randomized order on a screen and participants were instructed to read each word in their normal speaking voice. Items were presented in two blocks and there was a 500 ms ISI between words.

Following the baseline recording, participants were randomly assigned to one of two conditions for the exposure and shadowing portions of the experiment. While the target items were acoustically identical in each condition (English-Russian near-homophones; exposure/shadowing Set B), prior to the start of the exposure phase, participants were explicitly informed that they would hear Russian-accented English words or Russian words. During the exposure phase, participants listened to 8 target items and 23 fillers produced by the model, presented in random order, while images representative of the word meanings were simultaneously displayed. Following the playback of each item, participants responded to a simple multiple-choice question about the sounds they heard (e.g., *what was the first sound of the word?*) by pressing a key on a keyboard. The exposure phase was designed to prepare participants for the shadowing phase, by familiarizing them with the model talker and the target words that they would be producing. Pictures were used to make sure that participants in the accented English condition understood the words that they were to shadow despite the accented pronunciation. For participants in the Russian condition pictures were used to demonstrate that the sequences of sounds they heard in Russian were real lexical items (i.e., had meaning). Questions were used to ensure that the participants maintained focus and to direct their attention to the sound structure of the items that they heard.

In the shadowing phase, participants were aurally presented the same set of words as in the exposure phase (Set B), in a different randomized order and were given 2,000 ms to repeat each word into a microphone as accurately as they could. Each item was presented twice over two blocks.

Following the shadowing portion of the study, participants completed a post-test word-list reading task that was identical in structure to the baseline phase but contained different lexical items (Set C). The post-test phase was conducted to test whether potential acoustic convergence in the shadowing phase persisted over time and generalized to novel English words (that is, words not presented during

exposure/shadowing phase). In addition to generalization, post-test words differed from the baseline words, as previous research has noted that talkers are less likely to converge in items that they already pronounced [15].

Recordings were made in a sound-attenuated booth using an ART Tube MP Project Series preamplifier and a Shure KSM32 Embossed Single-Diaphragm microphone. Audio was captured using Audacity version 2.3.2 at a 44.1 sampling rate. The experimental interface was created using Psycho-Py [16]. All instructions (recorded by the native Russian talker) and exposure/shadowing stimuli were presented to participants via Sennheiser HD 380 Pro headphones.

2.4. Analysis

Word-initial VOT in the target words in the baseline, shadowing, and post-test recordings were measured using Dr. VOT – a machine-learning system based on a recurrent neural network [17]. The pre-trained program automatically measured the time between the release of a stop and the onset of periodic vibration of the following vowel. Each measurement was manually checked for errors prior to statistical analyses and any instances of pre-voicing/negative VOT were excluded from the analysis (8 total tokens; .06% of the total data set). A total of 1,395 tokens (Accented English: 691 tokens; Russian: 704 tokens) were submitted to a mixed-effects linear regression model with voiceless VOT duration as a dependent variable, using R (V. 4.2.1) [18] and the 'lme4' package [19]. The fixed effects of the model included Condition (English: reference, Russian), Phase (Baseline: reference, Shadowing, Post), and their interaction. Additionally, as many participants had knowledge of a foreign language that realized voicing in the same way as Russian (i.e., 'voicing languages' like Spanish, French, etc.), the model included a categorical fixed effect of Voicing Lang Knowledge (No: reference, Yes) and its interaction with Condition. The model also included by-subject random intercepts and slopes for subject by Phase, as well as by-word random intercepts.

3. RESULTS

3.1 Model talker acoustics

The model talker's voiceless VOT duration was first measured in order to confirm that it was within native-like expectations for Russian voiceless stops. On average, the model talker produced their voiceless stops with an average VOT duration of 3.42 ms, which is in line with previous literature on Russian stop voicing duration [14].

3.2 Voiceless stop shadowing by participants

Table 2 summarizes the fixed effects of the mixed-effects linear regression model. In the model, the intercept equals the estimated mean of English VOT produced by the accented English group (β_0 = 89.1, SE = 7.51). The statistical results showed a significant interaction effect of Condition with Phase. Specifically, participants significantly reduced their VOT durations by 20.4 ms on average when shadowing speech they believed to be Russian, compared to the baseline phase (β = -20.4, SE =8.69, t = -2.35, p = .03). A following post-hoc analysis using a pairwise coefficient Tukey adjustment additionally revealed that VOT in the shadowing phase also significantly differed from VOT in the post-test phase (β = 24.20, SE =7.53, t = 3.22, p = .03). By contrast, VOT values were not significantly different between the baseline and post-test phases (p = .39) in the Russian condition. The pairwise contrasts are also visually confirmed by the right panel of Figure 1 that demonstrates the raw VOT values.

Effects	Estimate	*t*	*p*
Intercept	**89.057**	**11.86**	**< .001**
ConditionRussian	4.576	.36	.72
PhaseShadowing	-3.766	-.52	.61
PhasePost	6.68	1.04	.30
VoicingLangKnow	**-20.567**	**-3.31**	**.003**
ConditionRussian:PhaseShadowing	**-20.398**	**-2.34**	**.02**
ConditionRussian:PhasePost	-3.043	-.86	.40
ConditionRussian: VoicingLangKnow	-1.649	-.13	.90

Table 2: Fixed-effects of the mixed-effects linear regression model.

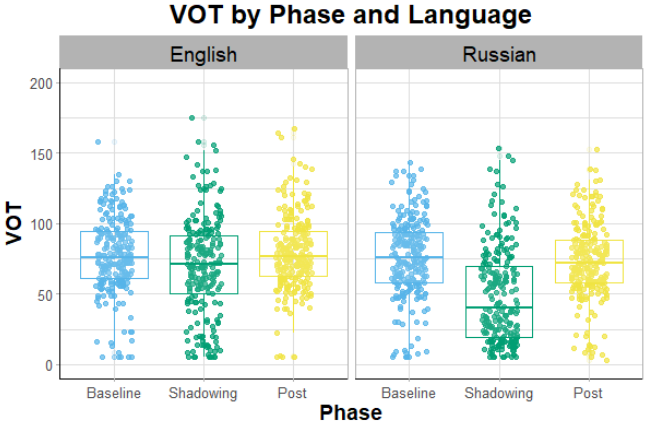

Figure 1: VOT values by Phase and Condition (left: English, right: Russian).

In contrast to the Russian condition, the results of the mixed-effects model indicated that participants'

VOT did not change significantly in either the shadowing or post-test phases compared to the baseline phase in the accented English condition (shadowing; p = .6, post; p = .3) despite a visually apparent tendency towards shorter VOTs in the shadowing phase (Figure 1, left panel). Furthermore, the pairwise post-hoc analysis confirmed that VOT values in the shadowing phase did not differ significantly from those in the post-test phase (p = .8). The results suggest that participants' shadowed speech did not demonstrate convergence towards the short lag VOTs of Russian-accented voiceless stops. In contrast, the results suggest that exposure to Russian speech had a significant effect on shadowed voiceless stops of participants in the Russian condition. The VOT of their voiceless stops in the shadowing phase was significantly shorter than their own baseline recordings, suggesting convergence towards, or imitation of the model speaker. Nevertheless, the effect did not generalize to the English words recorded in the post-test phase.

Finally, results of the mixed-effects model revealed a significant effect of Voicing Lang Knowledge (β = -20.6, SE =6.22, t = -3.31, p = .003), suggesting that participants who had existing knowledge of a language that realized voicing in the same manner as Russian had significantly shorter voiceless VOT across all conditions and phases: 25.1 ms shorter on average than participants without this knowledge.

4. DISCUSSION

The present study found convergence in voiceless VOT towards an unfamiliar language (Russian condition) but not towards an unfamiliar accent (accented English condition). Regarding the CAT, it appears that talkers are willing/able to imitate other languages, but not accented versions of their own, native language. However, the CAT approach may not be completely suitable to the setting of the present experiment. Imitating speech in an unfamiliar language is akin to speaking another language and likely reflects different processes than those involved in adjusting one's speech to accommodate an acoustically different version of one's L1.

The difference between the conditions could be further exacerbated by some of our methodological choices. In the accented English condition, participants may have detected the non-native (shorter) VOT in words like "pot" but were aware of the target phoneme due to the accompanying illustrations. As a result, they may have produced the words with native, English-like pronunciation. Comparatively, in the Russian condition, participants had no knowledge of the phonemic content of the words they shadowed and had to rely on the model's pronunciation in reproducing each item.

Moreover, as participants in the Russian condition were not familiar with the Russian language and were not provided with any orthographic representation of the words they shadowed, it is possible that, at least in some case, Russian voiceless stops were shadowed 'faithfully', that is, with short lag VOT, because they were mis-analyzed by participants as voiced (English short lag VOT is characteristic of voiced stops). This could also explain the lack of carry-over of this imitation effect into participants' English in the post-test phase, since there is no reason to expect that their pronunciation of Russian 'voiced' stops should affect their pronunciation of English voiceless stops. Future analyses of other acoustic correlates of voicing, as well as acoustic properties of other segments, such as vowels, would help paint a more complete picture of acoustic imitation of an unfamiliar language.

At face value, these findings do not lend support to the theory that exposure to accented speech, or to another language, can lead to L1 drift or to sound change in the ambient/dominant language via long-term phonetic convergence. Although participants imitated VOTs of Russian voiceless stops in the shadowing phase, there was no carry-over of this imitation into their English speech in the post-test. Furthermore, in the accented English condition, there was no convergence and no carry-over. That said, evidence of L1 drift was not completely absent from the present study, as results indicated that participants with existing knowledge of voicing languages, regardless of proficiency, produced significantly shorter voiceless VOT than their peers.

While present results do not suggest the tendency for talkers to converge with accented speech, since only one type of accent was examined in the study and the language, we should be cautious about generalizing these findings. It is possible that talkers may be more willing to accommodate certain speakers and accents than others. This may especially be true for Russian, considering the global political climate at the time of research. Therefore, future research should examine other L2 accents.

Moreover, it needs to be acknowledged that in this study participants underwent relatively short exposure in rather artificial conditions, and future work should consider longer exposure in more naturalistic settings in order to corroborate these results.

5. REFERENCES

[1] M. Babel, "Evidence for phonetic and social selectivity in spontaneous phonetic imitation," *J. Phon.*, vol. 40, no. 1, pp. 177–189, 2012, doi: 10.1016/j.wocn.2011.09.001.

[2] D. Kim and M. Clayards, "Individual differences in the link between perception and production and the mechanisms of phonetic imitation," *Lang. Cogn. Neurosci.*, vol. 34, no. 6, pp. 769–786, 2019.

[3] K. Nielsen, "Specificity and abstractness of VOT imitation," *J. Phon.*, vol. 39, no. 2, pp. 132–142, 2011, doi: 10.1016/j.wocn.2010.12.007.

[4] J. S. Pardo, "Measuring phonetic convergence in speech production," *Front. Psychol.*, vol. 4, 2013, doi: 10.3389/fpsyg.2013.00559.

[5] G. Zellou, R. Scarborough, and K. Nielsen, "Phonetic imitation of coarticulatory vowel nasalization," *J. Acoust. Soc. Am.*, vol. 140, no. 5, pp. 3560–3575, 2016, doi: 10.1121/1.4966232.

[6] M. Babel and D. Bulatov, "The role of fundamental frequency in phonetic accommodation," *Lang. Speech*, vol. 55, no. 2, pp. 231–248, 2012.

[7] R. Levitan and J. B. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," presented at the INTERSPEECH, 2011.

[8] J. S. Pardo, I. C. Jay, and R. M. Krauss, "Conversational role influences speech imitation," *Atten. Percept. Psychophys.*, vol. 72, no. 8, pp. 2254–2264, 2010, doi: 10.3758/BF03196699.

[9] H. Giles, *Communication accommodation theory: Negotiating personal relationships and social identities across contexts*. Cambridge University Press, 2016.

[10] C. B. Chang, "A novelty effect in phonetic drift of the native language," *J. Phon.*, vol. 41, no. 6, pp. 520–533, 2013, doi: 10.1016/j.wocn.2013.09.006.

[11] C. B. Chang, "Phonetic drift," in *The Oxford handbook of language attrition*, Oxford University Press, 2019.

[12] C. G. Clopper, "Sound change in the individual: Effects of exposure on cross-dialect speech processing," *Lab. Phonol.*, vol. 5, no. 1, pp. 69–90, 2014, doi: 10.1515/lp-2014-0004.

[13] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964, doi: 10.1080/00437956.1964.11659830.

[14] C. Ringen and V. Kulikov, "Voicing in Russian stops: Cross-linguistic implications," *J. Slav. Linguist.*, vol. 20, no. 2, pp. 269–286, 2012.

[15] S. Dufour and N. Nguyen, "How much imitation is there in a shadowing task?," *Front. Psychol.*, vol. 4, p. 346, 2013.

[16] J. Peirce *et al.*, "PsychoPy2: Experiments in behavior made easy," *Behav. Res. Methods*, vol. 51, no. 1, pp. 195–203, 2019.

[17] Y. Shrem, M. Goldrick, and J. Keshet, "Dr. VOT: Measuring positive and negative voice onset time in the wild," *ArXiv Prepr. ArXiv191013255*, 2019.

[18] R Core Team. (2022). *R: A language and environment for statistical computing*. URL https://www.R-project.org/.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *ArXiv Prepr. ArXiv14065823*, 2014.

[1] This study did not collect information about participants' race, which we recognize as a limitation considering the effect it has on sociolinguistic variation in the United States.