SPONTANEOUS IMITATION OF ENGLISH SIBILANTS BY NATIVE AND NON-NATIVE SPEAKERS

Ivy Hauser, Emily Graham and Xinwen Zhang

University of Texas at Arlington
ivy.hauser@uta.edu

ABSTRACT

This study investigates spontaneous phonetic imitation of the English alveolar sibilant /s/ by monolingual native speakers, bilingual native speakers, and non-native speakers with different language backgrounds in a delayed shadowing task. Participants first produced /s/-initial words, then were exposed to model speech with enhanced spectral mean (SM) on /s/, and finally produced /s/-initial words again post-exposure. All participants increased SM after exposure, converging towards the model talker. There were no significant differences in degree of imitation across the different language backgrounds. These results also provide an example of the starting distance bias associated with the difference-in-distance measure of imitation, as apparent significant differences across language backgrounds are accounted for by differences in starting distance to the model.

Keywords: phonetic imitation, speech acoustics, accommodation, bilingualism, methodology

1. INTRODUCTION

There is a large body of work demonstrating that talkers unconsciously alter their production towards speech to which they are exposed, even in lab settings without explicit instruction to imitate [1, 2]. Spontaneous imitation in native speech is constrained by social attitudes [3] as well as linguistic factors like phonological category boundaries and lexical frequency [2]. These factors also affect imitation in non-native speech, where degree of imitation is affected by belief about the linguistic proficiency of the model talker [4] and phonological representations in the second language (L2) [5, 6].

Linguistic and social distance between interlocutor pairs has been shown to be a factor in degree of convergence, but the direction of the effect is unclear. Some studies found more convergence for same dialect pairs than different dialect/language pairs [7], while other studies found more convergence for different dialect pairs [8, 9]. In addition, degree of convergence for L2 speakers specifically may depend on whether the model talker is a native speaker of the target language. In [10], Polish learners of English were found to converge more towards English vowel duration when the model talker was a native speaker of English. These divergent findings could be due to differences in the type of task; [7] used a conversational task while [8, 9, 10] used shadowing. Ultimately, it is unclear whether imitation is facilitated or inhibited when the model talker / interlocutor is of a different language background than the participant.

The present study builds on this by examining how speakers of different language backgrounds imitate a native English speaker in laboratory task. The study is designed to examine convergence across different language backgrounds, not the influence of native phonology in non-native shadowing, so the stimuli use the alveolar sibilant /s/ which is contrastively present in all participant languages. Participants include native English monolingual speakers, English-dominant native English-Spanish bilingual speakers, and non-native speakers with L1 Spanish, Vietnamese, or Urdu backgrounds. All participants were exposed to enhanced spectral mean (SM) on /s/, and imitation was expected for all groups as none were speakers of languages where enhancing SM on /s/ would threaten phonological contrast (which has been shown to inhibit spontaneous imitation [2]).

If different language backgrounds inhibit phonetic imitation (as in [7]), the native English speakers are predicted to converge to the model talker more than the speakers of other L1 backgrounds. If different language backgrounds facilitate phonetic imitation (as in [8, 9]), the non-native speakers are predicted to converge to the model talker more than the native speakers. It is also possible that the model talker's status as a native speaker of English could facilitate imitation for the non-native speakers independently of the degree of linguistic distance between the model and participants (as in [10]).

This study also provides a test case of the

difference-in-distance (DID) measure often used to quantify phonetic imitation, which has been shown to exhibit a 'starting distance bias' [11, 12]). When using DID, speakers who have greater baseline distance from the model will be found to have converged more than those who have smaller baseline distance from the model. This is the case even in datasets that lack convergence [11]. This is a particular problem in datasets like the one presented here, where differences in baseline starting distance co-vary with other variables of interest.

An alternative to DID is linear combination, which quantifies imitation by modeling post-exposure values as the dependent variable in a linear regression with baseline values and model talker values as predictors [11, 12]. This is different from most DID approaches, which typically model DID as the dependent variable. In a linear combination model, significant effects of both the baseline and model values indicate convergence. The analysis here compares DID and linear combination. Apparent DID differences across language background are accounted for by differences in starting distance to the model and do not appear to reflect systematic differences in degree of convergence based on L1/L2 status.

2. METHODS

The procedure is delayed shadowing with a baseline word-naming task, an exposure (listening) block, and a post-exposure word-naming task with no explicit instruction to imitate. The stimuli were designed to follow [2], who examined spontaneous imitation of enhanced English stops.

2.1. Stimuli

The test words for the baseline word-naming task comprised 40 high frequency [s]-initial words, 40 low frequency [s]-initial words, and 30 sonorant-initial filler words. The listening block contained the 80 [s]-initial words from the baseline production block and 40 sonorant-initial fillers that were different from those in the baseline production. The test words for post-exposure production comprised the baseline list plus 20 novel low frequency [s]-initial words and 20 novel low frequency [z]-initial words (the [z] words are not analyzed here).

All test words had initial stress and no onset clusters. Word frequency data was obtained from CELEX [13]. Thresholds for low and high frequency were below 300 and above 1000 per 17.9 million respectively. Test words were additionally counterbalanced for phonological neighborhood density (data obtained from the English Lexicon Project [14], number of syllables (1-3), and rounding of the following vowel.

The 80 target words and 40 filler words for the listening block were recorded by a phonetically trained female native American English speaker. Recording took place in a sound-attenuated booth at a university lab with a Shure SM35 headworn microphone and Audacity software [15]. Recordings were sampled at a rate of 44.1 kHz with a bit depth of 16.

The model speech with enhanced SM was created by shifting the spectrum of the sibilant noise up using the "shift frequencies" function in Praat [16]. The initial sibilants were segmented using TextGrids, so only the sibilant noise was shifted while the rest of the word was not. The spectra were shifted up 15% of the raw SM value. The manipulated stimuli were evaluated by two native American English speakers and one L2 English speaker for naturalness (to ensure the stimuli did not sound so unnatural that it interfered with participants' perception).

2.2. Participants

Twelve female English speakers between ages of 18-25 were recruited at a large American university. 4 were monolingual native speakers with no early exposure to other languages and no reported proficiency in any other languages, 2 were native bilingual Spanish-English speakers, 2 were non-native L1 Urdu speakers, 2 were non-native L1 Spanish speakers, and 2 were non-native L1 Vietnamese speakers. All non-native speakers reported high English proficiency and high L1 proficiency. Proficiency was determined with self-report on a questionnaire. High English proficiency is additionally verified by the fact that all participants are students at a US university where the language of instruction is English.

2.3. Procedure

The procedure had four blocks: warm-up reading, baseline word-naming, listening (exposure), and post-exposure word-naming. Most participants took 30-40 minutes to complete the study. The recording was done using the same parameters as noted in the above section for the stimulus recording. The stimuli were presented in a random order inside the sound-attenuated booth on an external monitor. Before beginning, participants were told the researchers were interested in how they speak naturally, as if talking to a friend.

The warm-up block for silent reading follows previous work [2, 1] and familiarizes participants with the words used in the baseline word-naming task to reduce potential hyperarticulation of low-frequency words. In the baseline block, the participants were presented with the same words as in the warm-up but they were instructed "to name the word they see by saying it aloud." In the listening exposure block, the participants were asked to listen to the manipulated words in the listening list using headphones. The instructions read, "In this section you will hear words but won't see any on the screen. You do not need to speak, just listen to what you hear." The last block, the post-exposure production task, used the same instructions as the baseline word-naming task, but included the additional novel words.

The recordings were force aligned [17] and all sibilant boundaries were hand edited by the researchers. Time-averaged SM values for each sibilant were extracted using a Praat script [18].

3. RESULTS

This section presents the imitation results by comparing two methods of quantifying imitation: difference-in-distance (DID; e.g. [3]) and linear combination [12, 11]. Figure 1 displays SM across the baseline and post-exposure blocks. On average, participants in each language group increased SM after exposure to the stimuli with enhanced SM. DID was calculated for each token as $|(SM_{baseline} - SM_{model})| - |(SM_{post} - SM_{model})|$ such that positive values indicate convergence and negative values indicate divergence. DID values across each language group are graphed in Figure 2. While all groups show an average positive DID indicating convergence, there is a visual difference in apparent degree of convergence. In particular, the L1 Vietnamese speakers' DID values are higher relative to the other groups.

To estimate the strength of this effect, DID was analyzed as the dependent variable using a basic DID model as in [11]: DID $\sim$ Language Group + Lexical Frequency + (1|speaker) + (1|word). The fixed effect of frequency was included because infrequent words have been shown to exhibit more convergence [1, 2], but there was no significant effect of frequency on DID in the model. There was a significant effect for the L1 Vietnamese group ($\beta = 674.26, p < 0.001^{***}$) indicating significantly higher DID and more convergence relative to the intercept L1 monolingual English group. However, DID is known to exhibit bias according to starting

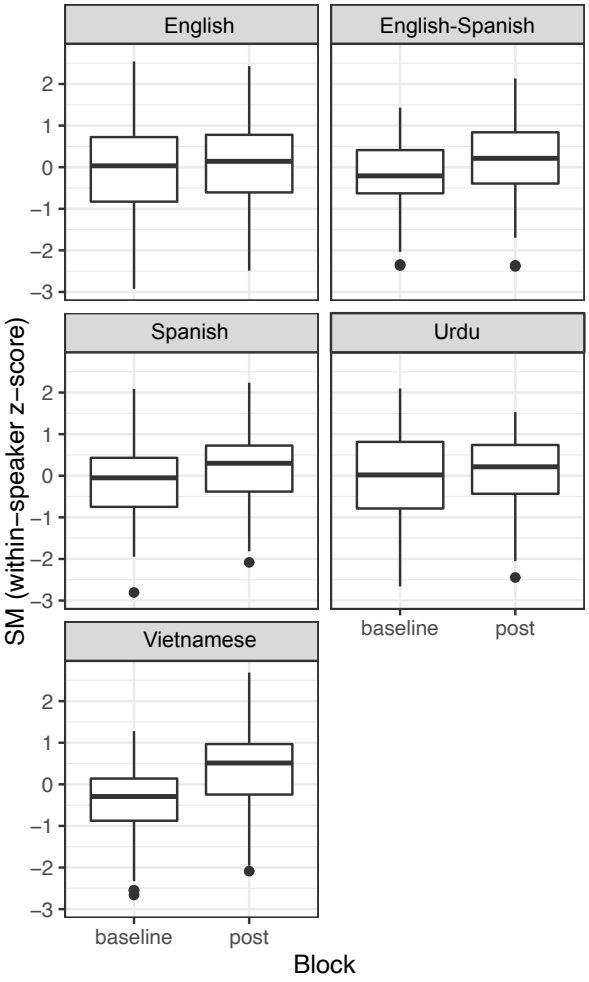

Figure 1: SM across blocks for each language group. Panels are labeled with L1's of participants.

distance and this result does not obtain in a linear combination model which takes baseline SM into account.

Differences in average baseline SM across the language groups are shown in Table 1. The L1 Vietnamese group had the largest DID values, but they also exhibited the lowest average baseline SM. This means that they had more space to converge to the model relative to the other groups. Because of the small number of speakers in each group (2-4), it is not clear whether the differences in baseline SM are due to language background, individual differences, or other factors. However, it is possible to determine whether these baseline SM differences affected the DID results.

Another method of quantifying imitation that takes baseline distance into account is linear combination. A linear combination model following [12, 11] was constructed with post-exposure

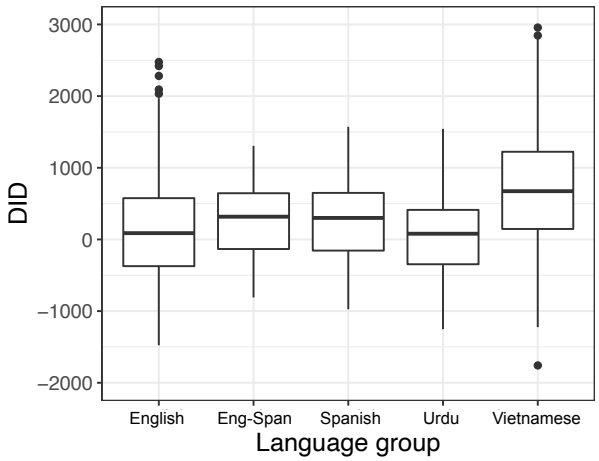

Figure 2: DID for each language group. Positive values indicate convergence.

Language group	Average baseline SM (Hz)
L1 English	7717
L1 English/Spanish	9127
L1 Spanish	7834
L1 Urdu	7791
L1 Vietnamese	6658
Model (enhanced)	10649

Table 1: Baseline SM for each language group and the model speech with enhanced SM.

SM as the dependent variable: Post SM $\sim$ Baseline SM + Model SM + Language Group + Model SM:Language Group + Model SM:Lexical Frequency + (1|speaker) + (1|word). In this model, baseline SM is significant ($\beta = 0.33, p < 0.001^{***}$) indicating that speakers are overall consistent between baseline and post-exposure blocks. Model SM is also significant ($\beta = 0.30, p < 0.001^{***}$) indicating overall convergence to the model talker in the post-exposure block. The interaction between Model SM and Frequency remains insignificant, indicating no significant difference in convergence based on lexical frequency. Unlike in the DID model, none of the Language Group terms or their interactions reached significance. Crucially, the interaction between Model SM and the L1 Vietnamese group is not significant ($\beta = 0.03, p = 0.81$). This means that the degree of convergence for the L1 Vietnamese group does not differ significantly from the intercept L1 English group.

4. DISCUSSION AND CONCLUSION

This study investigated spontaneous imitation of English /s/ across participants with different language backgrounds. After exposure to a native model talker with enhanced SM on /s/, participants in all groups increased their own SM to converge towards the model talker. When analyzing the results using the common DID measure, it appeared that the L1 Vietnamese group converged more than the other language groups. However, when the same results were analyzed using linear combination, there were no significant differences in degree of convergence across groups. This is because the L1 Vietnamese speakers had a lower average baseline SM.

Methodologically, these results provide another test case for the 'starting distance bias' associated with DID in phonetic imitation research. Using DID led to an epiphenomenal significant effect because differences in baseline acoustic values were correlated with other variables of interest (in this case, language background). It is not the case that the L1 Vietnamese speakers imitated the model talker more. Those speakers simply had baseline values that were more distinct from the model, which inflated DID values. These results therefore provide support for methods like linear combination, which take baseline values into account when quantifying degree of convergence.

Empirically, enhancement of the English alveolar sibilant /s/ was imitated across all language backgrounds examined. These results suggest that imitation is not necessarily inhibited or facilitated by a mismatch in L1 status between the participant and model talker when the relevant phonological category structure is similar across the two languages.

5. REFERENCES

[1] S. D. Goldinger, "Echoes of echoes? An episodic theory of lexical access." *Psychological Review*, vol. 105, no. 2, p. 251, 1998.

[2] K. Nielsen, "Specificity and abstractness of VOT imitation," *Journal of Phonetics*, vol. 39, no. 2, pp. 132–142, 2011.

[3] M. Babel, "Evidence for phonetic and social selectivity in spontaneous phonetic imitation," *Journal of Phonetics*, vol. 40, no. 1, pp. 177–189, 2012.

[4] F. Jiang and S. Kennison, "The Impact of L2 English Learners' Belief about an Interlocutor's English Proficiency on L2 Phonetic Accommodation," *Journal of Psycholinguistic Research*, vol. 51, no. 1, pp. 217–234, 2022.

[5] J. E. Flege and W. Eefting, "Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation," *The Journal of the Acoustical Society of America*, vol. 83, no. 2, pp. 729–740, 1988.
[6] M. Llompart and E. Reinisch, "Imitation in a second language relies on phonological categories but does not reflect the productive usage of difficult sound contrasts," *Language and Speech*, vol. 62, no. 3, pp. 594–622, 2019.
[7] M. Kim, W. S. Horton, and A. R. Bradlow, "Phonetic convergence in spontaneous conversations as a function of interlocutor language distance," *Laboratory Phonology*, vol. 2, no. 1, pp. 125–156, 2011.
[8] M. Babel, "Dialect convergence and divergence in New Zealand English," *Language in Society*, vol. 39, pp. 437–456, 2010.
[9] A. Walker and K. Campbell-Kibler, "Repeat what after whom? exploring variable selectivity in a cross-dialectal shadowing task," *Frontiers in Psychology*, vol. 6, p. 546, 2015.
[10] M. Zając and A. Rojczyk, "Imitation of English vowel duration upon exposure to native and non-native speech," *Poznan Studies in Contemporary Linguistics*, vol. 50, no. 4, pp. 495–514, 2014.
[11] B. MacLeod, "Problems in the difference-in-distance measure of phonetic imitation," *Journal of Phonetics*, vol. 87, p. 101058, 2021.
[12] U. C. Priva and C. Sanker, "Limitations of difference-in-difference for measuring convergence," *Laboratory Phonology*, vol. 10, no. 1, 2019.
[13] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "The celex lexical database, release 2," 1996.
[14] D. A. Balota, M. J. Yap, K. A. Hutchison, M. J. Cortese, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson, and R. Treiman, "The English lexicon project," *Behavior research methods*, vol. 39, no. 3, pp. 445–459, 2007.
[15] Audacity Team, *Audacity(R): Free Audio Editor and Recorder*, http://audacity.sourceforge.net/, 1999-2021.
[16] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5:9/10, pp. 341–345, 2001.
[17] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner [computer program] version 0.9.0, retrieved from http://montrealcorpustools.github.io/montreal-forced-aligner/." 2017.
[18] C. DiCanio, "Time averaging for fricatives 2.0," Praat script published online, 2021.