VARIATION IN PERCEPTION AND PRODUCTION OF /s/-/ʃ/ IN ENGLISH

Roger Yu-Hsiang Lo¹, Suyuan Liu¹, Charlotte Vaughn²,³, Michael McAuliffe⁴, and Molly Babel¹

¹Linguistics, University of British Columbia, Canada, ²Language Science Center, University of Maryland, United States, ³Linguistics, University of Oregon, United States, ⁴Linguistics, McGill University, Canada
roger.lo@ubc.ca, suyuan.liu@ubc.ca, {cvaughn@umd.edu, cvaughn@uoregon.edu}, michael.emcauliffe@gmail.com, molly.babel@ubc.ca

ABSTRACT

In this exploratory study, we examine the relationship between perception and production in /s/ and /ʃ/ in North American English. To this end, peak ERB N trajectories fitted with functional PCA characterize the fricative space on an individual level for 121 participants. Mahalanobis distances quantify the degree of category overlap and distance. Participants also completed a categorization task on /s/-/ʃ/ continua with four minimal pairs, the data of which were fit to a mixed-effects logistic regression. Individual logistic slopes are compared with the by-individual Mahalanobis distance, suggesting a subtle link between production and perception. An additional analysis suggests that the strength of the perception-production link is moderated by participants’ multi-/monolingual status.

Keywords: sibilant fricatives, production, perception, phonetic variation

1. INTRODUCTION

This project began as a theoretical inquiry about perceptual learning with /s/ and /ʃ/. In discovering that listener samples from two universities had surprisingly different categorization thresholds for the /s/-/ʃ/ contrast, we decided to look at production as a source of the perception differences. That led to the current data, which are exploratory in nature.

Community production patterns often mirror community perception patterns. Spanish-speaking individuals, for example, have a /b/-/p/ perceptual category boundary that reflects the voice onset time (VOT) distribution in Spanish productions, exhibiting a lower VOT crossover point than English-speaking individuals, whose own threshold reflects English productions [1, 2]. On an individual level, however, the relationship between production and perception is more variable. While there is supporting evidence for some type of connection between perception and production, it is not a straightforward reflex [3, 4, 5, 6]. We structure our exploration of perceptual category boundary variation and acoustic-phonetic production variation in the context of this body of literature.

We focus on /s/ and /ʃ/ in North American English, first characterizing the degree of acoustic-auditory overlap on a by-talker basis using functional Principal Components Analysis (FPCA) on fricative trajectories. We use a large number of voices (n = 121) producing a large number of real words (n = 40). We consider the relationship between one’s own realization of the /s/-/ʃ/ contrast in production and one’s categorization function in recognition of /s/-/ʃ/ minimal pairs. Our focus on word recognition, as opposed to lower-level perceptual acuity [7], connects to individuals’ production-perception link at the word level [8, 9]. Because our investigation is exploratory in nature, we take the opportunity to consider how bi-/multilingualism may affect the link between perception and production, theorizing about the impacts of multilingualism on the perception and production relationship in Section 3. The nature of our data set ultimately allows the larger project to also serve as an opportunity to replicate the seminal work of [10].

2. EXPERIMENTS

2.1. Participants

Participants recruited at the University of British Columbia (UBC) and the University of Oregon (UO) were given partial course credit or cash for their time. Data from 121 participants are included in the current analysis. Participants had heterogeneous language backgrounds that are are representative of the speech communities under study, and all acquired English before the age of five.

2.2. Production: Materials and Methods

First, participants read 252 single words, including 100 /s/-words, 47 /ʃ/-words, and 105 filler words not containing sibilant sounds. Words containing
sibilants were either bi- or tri-syllabic. Of the 100 /s/-words, 50 words had /stu/ clusters, and 50 were /sV/ words. Of the 47 /ʃ/-words (all /NV/), 20 had /ʃ/-word-initially and 27 had /ʃ/-word-medially. Participants had the option to repeat words if they wished, and only the final repetition was saved. To match the perception data, only items with fricatives in initial position and with following vowels (e.g., not the /stu/ clusters) are analyzed in this paper (n = 40).

Production tokens were discarded if they contained incorrect pronunciations or stress, or any disfluencies. Productions were force aligned using the Montreal Forced Aligner [11]. Following any disfluencies. Productions were force aligned contained incorrect pronunciations or stress, or

2.3. Perception: Materials and Methods

Perception stimuli came from a previous study [12]. An 11-step continua between monosyllabic /s/-/ʃ/ minimal pairs (sack-shack, sigh-shy, sin-shin, sock-shock) was created using Tandem-STRAIGHT [13]. A white college-aged English-L1 male from UBC produced the endpoints. Pretests were conducted at both testing locations (UBC and UO) to find the center portion of each continuum. The categorization for listeners at location UBC was shifted a half to a full step more towards /s/ compared to listeners at location UO. A 7-step continuum centered around each location’s midpoint was presented to listeners at that location. Each step of each continuum was repeated 7 times, giving a total of 196 trials (7 steps × 7 repetitions × 4 continua) per participant. For each trial, the listener categorized the word as either the /s/-word (e.g., sin) or the /ʃ/-word (e.g., shin) of the minimal pair.

2.4. Analysis and Results

The goal of the analysis is to derive metrics to quantify the degree of contrast between /s/ and /ʃ/ in both production and perception, and to examine the correlation across modalities. Mahalanobis distance [14] was used for production and the logistic beta-coefficient for perception to quantify the degree of contrast. It should be noted that, despite analyses on production and perception data being described separately, the degree of contrast across the two modalities was assessed in a single Bayesian model to account for uncertainty surrounding estimates. All analyses were performed in R [15], with Bayesian models fitted using CmdStanR [16].

2.4.1. Production

Using the script developed in [17], a trajectory of peak ERBN numbers, henceforth peak ERBN, which represent frequencies with the greatest amount of excitation on the multitaper psycho-acoustic spectrum calculated over a sequence of 20-ms windows, was extracted over the duration of each fricative with a 10-ms stride. Tokens with fricative duration shorter than 60 ms and those with within-token standard deviation in peak ERBN beyond the 0.95 quantile were removed. After removal, 5,199 tokens entered the analysis.

We analyzed peak ERBN trajectories with FPCA, following the steps laid out in [18]. FPCA offers a way to succinctly quantify global variation in peak ERBN. In essence, FPCA treats each peak ERBN contour as a composition of a grand mean curve and a small number of principal component (PC) curves (each of which encodes a different deformation of the mean curve) weighted by corresponding PC loadings, which can be studied using conventional statistical tests. Here we focus on the first three PC components (i.e., PC1, PC2, and PC3), which jointly explain 96.5% of the variance and whose effects on the grand mean peak ERBN curve are shown in Figure 1. PC1 chiefly controls the extent of dipping/rising at both ends of the curve, and PC3 alters the direction of curvature. PC1 alone accounts for 89.3% of the variance and serves to bipartite the two fricatives, though to different degrees across participants, as shown in Figure 2A. To more holistically capture this degree of contrast, the Mahalanobis distance, which can be understood as a generalization of Cohen’s d [19] to higher dimensions, was estimated for each individual based on the loadings of their first three PCs. The distribution of individual (posterior mean) Mahalanobis distances is shown in the top histogram of Figure 2C, and individuals with the largest as well as smallest Mahalanobis distance are also featured in Figure 2A. These results indicate that individuals do come with different degrees of contrast, mainly in terms of the height proximity of peak ERBN trajectories of the two fricatives.

2.4.2. Perception

We excluded perceptions trials where no response was registered. The responses of the remaining trials were modeled with a mixture model: one component was a mixed-effects logistic regression that had step as the fixed effect as well as by-participant and by-word random intercepts and slopes, and the other
2. Speech Acoustics

Figure 1: PCs of peak ERBₜ trajectories. The dotted line represents the grand mean curve \( \mu(t) \) while the solid/dashed curves are obtained by adding (orange) or subtracting (blue) from \( \mu(t) \) the curve \( \sigma(\text{PC}_i \text{ loadings}) \cdot \text{PC}_i \), where \( \sigma \) denotes the standard deviation, and \( i = 1, 2, 3 \).

was sheer randomness due to the participant making accidental button-push errors. That is, each response had a chance, \( \gamma \), of being generated by a random process, and a probability of \( 1 - \gamma \) that it came from the logistic function of the step variable.

The only independent variable—step—was treated as a continuous variable and \( z \)-transformed with respect to the original manipulation steps (e.g., Step4 was consistently mapped to 0, regardless of listeners). The default level for the response was /s/-word (i.e., /s/-word responses were coded with 0, and /ʃ/-word responses were coded with 1), so a positive \( \beta_{\text{step}} \) means that larger step values elicit more /ʃ/-word responses. All individuals demonstrate a clear sigmoid function, suggesting a robust perceptual contrast for /s/ and /ʃ/. However, as shown in the right histogram of Figure 2C and highlighted in Figure 2B, individuals also vary with respect to the slope of their sigmoid function. Overall, these results speak to the existence of cross-listener variation in the “crispness” of the boundary between the two perceptually robust fricatives [20].

2.4.3. Production–Perception Link

Finally, we regressed individual logistic slopes in perception on individual Mahalanobis distances in production to examine the correlation across the modalities. An initial analysis of all participants revealed a positive, but negligible, relationship (mean corr. = .053, 95% credible interval (CrI) = [−.049, .186], \( p(\text{corr.} > 0) = .85 \), as depicted in Figure 2C. An exploratory analysis separating monolingual and bi-/multilingual individuals found a consistent positive relationship for the English monolingual group (mean corr. = .118, 95% CrI = [−.008, .254], \( p(\text{corr.} > 0) = .96 \) and a more variable pattern for the bi/multilingual individuals (mean corr. = −.010, 95% CrI = [−.135, .112], \( p(\text{corr.} < 0) = .57 \). The resulting correlations, separated by language background, are visualized in the scatter plot of Figure 3.

3. DISCUSSION AND CONCLUSION

Our findings echo prior work demonstrating that individuals vary in the distinctiveness of /s/ and /ʃ/ in North American English [10]. Talkers exhibited considerable variation in the degree of acoustic-auditory contrast in production, ranging from highly separable /s/ and /ʃ/ categories and those with considerable overlap. Similarly, in perception, while all individuals perceived a contrast, there was variation in the slope of the curve, indicating variation in the perceptual distinctiveness of the two categories. We used these data to explore the within-individual production-perception link. That there is a relationship between perception and production is at the heart of many models of sound-based knowledge (e.g., exemplar-based models [21] and motor-theory [22]), in addition to being a core mechanism for contemporary theorizing about sound change [3]. Considering the dataset as a whole, we did not find an overall strong correlation between production and perception.

We divided our participants into two groups, separating individuals who are monolingual and bi-/multilinguals. The initial observed lack of a strong relationship holds for the bilingual group and a suggestive link emerges for the monolinguals. A weak or nonexistent correlation between bilinguals’ perception and production was also attested in [25]. Specifically, that study found a strong correlation between English monolinguals’ production and perception, but the relationship did not hold for their Spanish-English bilingual group. The bilingual group produced the contrasts in production, indicating phonetic sensitivity in that modality; [23] reason that the conflicting phonological status across English and Spanish contorts the perception space. Similarly, while [24] finds that Spanish-Catalan bilinguals maintain the Catalan mid-vowel contrasts in production, their listeners find the distinction more challenging in perception and may not robustly encode the contrasts at the lexical level; see also [25].
Why might the perception-production link be absent for bi-/multilingual speakers? It is not due to an across-the-board lack of a relationship in bilinguals [9, 8]. Crosslinguistic influence in bilingual linguistic systems may affect both production and perception [26]. How observable that influence is varies as a function of language dominance and language mode (e.g., [2]). As alluded to above, however, while early bilinguals may maintain distinctions in production and phonetic level discrimination, bilinguals may experience competition in categorization and recognition that masks the cross-modal link or impedes the encoding of phonetic detail in the lexicon. Additional theorizing about the nature of the perception-production link for bilinguals is necessary, and it may shed light on the variable nature of the perception-production connection in monolingual listeners as well (e.g., [27]).

To summarize, this project observed individual variation in both production and perception of North American English /s/ and /ʃ/, showing a positive correlation between the two modalities for monolingual, but not bi-/multi-lingual speakers. Our future work will explore the consequences of this production variation in perception and recognition for listeners, as a conceptual replication of [10].

4. ACKNOWLEDGMENTS

Thank you to Stephanie Chung, Yu Cen Gao, Sam Elliott, Celina Maldonado, Omar Ortiz, and Blair
Prater for their assistance with this work. Funding for this work has come from a SSHRC Grant to Molly Babel.

5. REFERENCES


\[^1\] In [17], 17 equal-distance peak ER B N were extracted for each trajectory. We also performed analyses on the 17-point data and obtained very similar results.