The consequence of auditory-acoustic contrast on perception and recognition of English /s/ and /ʃ/

Roger Yu-Hsiang Lo[1], Charlotte Vaughn[2], Michael McAuliffe[3], and Molly Babel[1]

[1]University of British Columbia (Canada), [2]University of Maryland (USA), [3]McGill University (Canada)

Introduction: A basic mantra in our research community is that speech is highly variable. Individuals vary in their acoustic realizations based on a host of learned and physiological/anatomical factors. Individuals also vary within themselves as a function of social, pragmatic, and linguistic context. And, listeners are sensitive to this phonetic variation at multiple levels, attending to within-category variation [1,2]. In a seminal study, Newman, Clouse, and Burnham [3] quantified the degree of /s/-/ʃ/ contrast in American English, using skewness and centroid in a small sample (n = 20), and used selected productions in a series of fricative categorization experiments. They found that listeners' responses to talkers with more variable fricative productions were slower, though listeners' ability to categorize the varied fricatives was robust.

Experiment: In this study, we capitalize on a large dataset of North American English-speaking voices producing /s/ and /ʃ/ words to revisit the consequences of category distinctiveness on various aspects of perception. We use six voices and a diverse set of listeners from our university's speech community. The six voices were selected from a set of 121 individuals, whose /s/ and /ʃ/ data are described in Lo et al. [4]. In brief, productions of word-initial /s/ and /ʃ/ were characterized based on a time-series of peak ERB_N numbers, which are a psychoacoustic measure of peak frequency [5]. These time-series were decomposed with functional principal components analyses (FPCA; [6]), and individuals' degree of auditory-acoustic contrast was then summarized using the Mahalanobis distance [7], which is a generalization of Cohen's d [8] to higher dimensions, based on FPCA loadings. Such a measure considers degree of contrast and variability in tandem. The tokens from the three talkers with the largest contrast and the three with the smallest contrast were selected as stimuli for the three perception studies described below. The fricative distributions from the selected talkers are shown in Figure 1.

Procedure: Listeners were asked to categorize (1) the isolated fricative (C-only; $n_{listeners}$ = 62; n_{tokens} = 15,998), (2) the fricative-vowel sequence (CV; $n_{listeners}$ = 55; n_{tokens} = 14,071), or (3) to complete a speeded-shadowing task where listeners were auditorily presented with the full words and asked to identify the words by repeating them as quickly and accurately as possible ($n_{listeners}$ = 128; n_{tokens} = 29,068). All three tasks were conducted online using Gorilla [9]. Categorization choice and response time were registered for the fricative categorization task, and participant productions were recorded for the speeded-shadowing task. The analyses on the speeded-shadowing task are based on automatic detection of vocal onset from a Praat script, with tokens where no onset or multiple onsets were detected being excluded. The hypothesis is that listeners will be able to identify the fricatives more accurately and faster when produced by the talkers with the larger production distance, and that these same voices will elicit shorter onset latencies in speeded shadowing.

Results: Data were fitted with Bayesian mixed effects models using `brms` [10] in R [11]. As shown in Figure 2, broadly, the fricatives from talkers with greater contrast were identified more accurately and more quickly, and this was observed with greater effect size for the C-only condition and /s/ productions. These results indicate that listeners leverage information from the formant transitions in distinguishing these fricatives (e.g., [12]), and suggest that a reduced contrast in /s/ and /ʃ/ is the result of /s/ having more high amplitude lower frequency components. The speeded-shadowing results indicate the participants are faster at identifying the words produced by the *least* distinctive voices. This is the opposite of the expected pattern and, minimally, underscores the robustness of whole-word recognition. Coupling C-only, CV, and word-level responses paints a more accurate picture of how talker differences in auditory-acoustic contrast affect categorization and intelligibility.

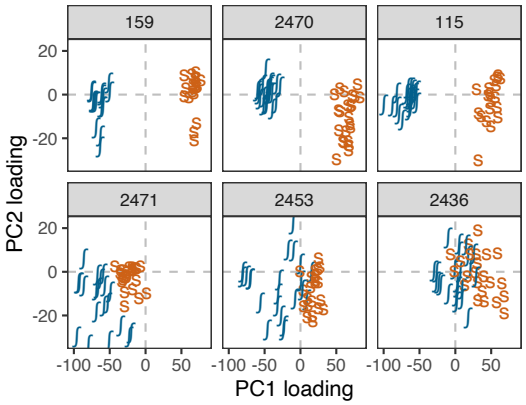

Fig. 1. Distributions of /s/ and /ʃ/ for talkers with the largest (159, 2470, 115) and smallest (2471, 2453, 2436) contrast, in terms of the loadings of the first and second principal components (PCs). Each s/ʃ label represents a s/ʃ-initial word production from a talker.

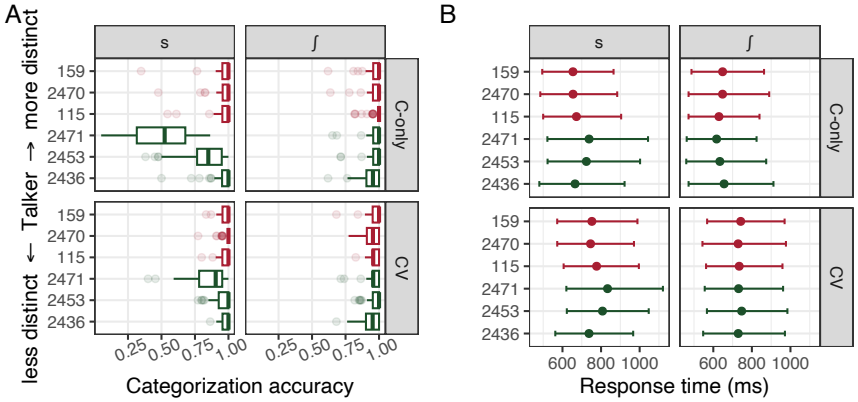

Fig. 2. Listener performance as a function of fricatives (/s/ and /ʃ/), stimulus types (C-only and CV), and talkers with different degrees of /s/-/ʃ/ contrast (red for largest contrast, and green for smallest): (A) categorization accuracy and (B) response time (the dots represent the means, and the whiskers span one standard deviation above and below the mean).

References

[1] McMurray, B., R. N. Aslin, M. K. Tanenhaus, M. J. Spivey & D. Subik (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance*, 34, 1609-1631.

[2] Munson, B., J. Edwards, S. K. Schellinger, M. E. Beckman & M. K. Meyer (2010). Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of *Vox Humana*. *Clinical Linguistics & Phonetics*, 24, 245-260.

[3] Newman, R. S., S. A. Clouse & J. L. Burnham (2001). The perceptual consequences of within-talker variability in fricative production. *Journal of the Acoustical Society of America*, 109, 1181-1196.

[4] Lo, R. Y.-H., S. Liu, C. Vaughn, M. McAuliffe & M. Babel (2023). Variation in perception and production of /s/-/ʃ/ in English. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 624-628). Guarant International.

[5] Reidy, P. F. (2016). Spectral dynamics of sibilant fricatives are contrastive and language specific. *Journal of the Acoustical Society of America*, 140, 2518-2529.

[6] Gubian, M., F. Torreira & L. Boves (2015). Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics*, 49, 16-40.

[7] Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2, 49-55.

[8] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

[9] Anwyl-Irvine, A. L., J. Massonnié, A. Flitton, N. Kirkham & J. K. Evershed (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 52, 388-407.

[10] Bürkner, P. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80, 1-28.

[11] R Core Team (2019). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing (https://www.R-project.org/).

[12] Wagner, A., M. Ernestus & A. Cutler (2006). Formant transitions in fricative identification: The role of native fricative inventory. *Journal of the Acoustical Society of America*, 120, 2267-2277.