9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Non-segmental conditioning of sibilant variation in American English

Jacob B. Phillips[1], *Daniel Chen*[2], *Alan C. L. Yu*[1]

[1]Phonology Laboratory, University of Chicago, United States
[2]Institute for Advanced Study, Toulouse School of Economics, France
jbphillips@uchicago.edu, daniel.li.chen@gmail.com, aclyu@uchicago.edu

Abstract

Variation, both between and within speakers, is ubiquitous in language. Examining and understanding this variation is crucial not only to questions of sociolinguistics and sound change but also to the study of prosody and phonology more broadly. The present study contributes to the literature on inter- and intra-speaker variation in speech production, examining the phonetic realization of prevocalic /s/ in American English using recordings from a longitudinal phonetic corpus of oral arguments before the Supreme Court of the United States. Specifically, this study examines the role of non-segmental factors conditioning the observed variation, focusing on prosodic prominence, segment duration, phonological contrast and lexical frequency. Significant effects for segment duration, speakers' average /sh/ centroid frequency and word position were observed, with higher centroid frequency values observed in instances of /s/ with a longer relative duration, in word-initial positions, or for speakers with a higher mean /sh/ centroid frequency. There was also significant individual variation in the effects of duration, prosodic prominence and phonological contrast. These results provide further evidence for place of articulation contrast strengthening in prominent positions with novel evidence for place of articulation contrast strengthening in sibilants.
Index Terms: corpus phonetics, inter-speaker variation, intra-speaker variation, prominence, lexical frequency

1. Introduction

The acoustic realization of human speech is marked by substantial individual variation. This study focuses on the nature of sibilant variation using a longitudinal corpus of spontaneous speech. Sibilant realization may vary due to gender, class, or dialect [1, 2]. Even speakers of the same gender and dialect have nonetheless been shown to differ, for example, in terms of frication centroid frequencies and skewness [3]. Sibilant realization also robustly varies by the phonological context, with adjacent segments exerting a high degree of coarticulatory influence on /s/ production with, for example, subsequent vowel rounding [4] or following consonants, especially /r/[5, 6], leading to lower centroid frequency values. This study focuses on the role of non-segmental factors including prosody, lexical frequency and phonological contrast, in the conditioning of the observed sibilant variation.

Lexical frequency has been shown to affect segmental realization and has been suggested to contribute to the the actuation and propagation of sound change, but it is unclear how it affects the realization of sibilants. It has been proposed that sound changes propagate across high frequency words and trickle down to lower frequencies [7]. Lin et al. [8] also demonstrate that gestural reductions occur at higher rates in high frequency words. Here, we seek to determine the role of lexical frequency in sibilant realization.

Additionally, prosodic factors have been demonstrated to influence segmental variation, with segments strengthened at prosodic boundaries and in stressed or accented positions. Most robustly, segments have been demonstrated to lengthen adjacent to a prosodic boundary [9, 10]. Furthermore, segments have been suggested to enhance contrast cues in prominent positions [11]. For example, /i/ in English has been demonstrated to be higher adjacent to boundaries, and fronter in accented positions [11]. Additionally, English stops have been demonstrated to enhance voicing contrasts [12], but little evidence has been robustly demonstrated for place contrasts. Cho and Keating [13] demonstrate changes in the spectral energy in stop bursts, with a higher centroid frequency observed in stressed syllables and individual variation in the effect of utterance position on the burst centroid frequency. Specific examinations of sibilants have demonstrated less prominence effects than other consonants [14], with greater evidence for voicing contrast enhancement than place enhancements [15]. In this study, we continue the investigation of prosodic strengthening of sibilants, asking whether place contrast is enhanced at word boundaries.

A final factor examined in this study is the role that the phonological contrast between two sounds play in shaping phonetic variation. Previous studies have suggested that phonetic variation may be constrained by the phonological system itself. Vowel-to-vowel coarticulation, for example, has been shown to vary according to the size of the vowel inventory of the language; the larger the number of contrastive vowels, the less vowel-to-vowel coarticulation is observed [16] (though see [17]). VOT realization has also been shown to co-vary across place of articulation across speaker [18]. Here, we ask whether the realization of /s/ is constrained by the realization of /ʃ/ and, if so, whether this contrast effect would modulate the prosodic and frequency effects discussed above.

Using data culled from a longitudinal corpus of spontaneous speech, the speech of 419 individuals were examined from an eight year period. The present study examines inter- and intra-speaker variation in sibilant production, focusing on the role non-segmental factors including sibilant duration, word position, stress, lexical frequency, and /s/-/ʃ/ contrast in conditioning the observed variation.

2. Materials & Methods

2.1. The SCorpus

The present study examines the SCorpus [19], a longitudinal corpus of the oral arguments before the Supreme Court of the United States (SCOTUS), focusing on the 2006–2013 terms. The recordings and their associated transcripts were drawn from the Oyez Project (`http://www.oyez.org/`), a multimedia archive of the cases before the SCOTUS hosted by the Chicago-Kent College of Law. Over the course of each term, lasting roughly 200 days, the Court hears arguments from approxi-

10.21437/SpeechProsody.2018-201

mately 80 cases, although the number varies from term to term. 40 cases were selected from each term, yielding approximately 40 hours of audio per year and 320 hours in total. The 320 cases were heard on 225 unique days over an eight an a half year period. During these arguments, attorneys for both sides of the case, the petitioner and respondent, are given 30 minutes to argue their case before the bench of 9 justices. The justices are active participants, interrupting with questions, comments, and occasionally jokes.The attorneys vary on a case-by-case basis, but the average justice serves on the bench for 16.5 years. The present work studies the speech of both the justices and the attorneys arguing before the bench, examining 11 justices and 408 attorneys. The SCorpus provides a unique opportunity for examining variation, as it is comprised of spontaneous, interactive speech recorded in a highly controlled environment over a period of years. Furthermore, as the tone veers formal and speakers are disproportionately older, educated men, sociolinguistic variation is less expected and any such variation observed sheds light on the extent of intra- and inter-speaker variation in language. The present study is restricted to recordings after 2005 because the Supreme Court switched in October 2005 from a reel-to-reel taping system to digital recording with audio compressed as MP3.

2.2. Segmentation

All transcripts were manually edited for accuracy against the audio recordings and subsequently forced-aligned using Penn Forced Aligner [20], whose modeled were originally trained on the SCorpus. The Penn Force Aligner determines phone-level boundaries using the HTK toolkit [21] and the CMU American English Pronouncing Dictionary [22] to determine phonemic representations of words, including stress values.

2.3. Extraction

In the present study, centroid frequency will be used as the primary acoustic cue to characterize and distinguish the fricatives [23]. Using the boundaries determined by force alignment, centroid frequency values were extracted for all intervals labeled as /s/ or /ʃ/ using a Praat [24] script originally created by DiCanio [25] and modified by the researchers. Given the size of the corpus, we rely on the automatic alignment and measurements without manual correction. All measurements were time-averaged to minimize potential errors in the spectral measurements [26]. Time-averaged centroid frequency measurements were calculated on the middle 80% of the sibilant (to exclude transitions) using six 15 ms windows with preemphasis at 80 Hz and an examined frequency range from 500 to 12000 Hz.

3. Analysis

All instances of prevocalic /s/ were analyzed when in word-initial or word-medial position in either primary stressed or unstressed syllables. Sibilants were excluded if they were contained in incomplete words, if there was any simultaneous or overlapping speech from another talker, or if lexical frequency values could not be determined, for example proper names or legal jargon. Additionally sibilants were excluded if the speaker was not successfully identified in the transcript. These criteria yielded a total of 230,169 analyzed instances of /s/.

While centroid frequency measurements were extracted for /ʃ/, these values served as explanatory variables rather than dependent variables in our analysis. The average centroid frequency of /ʃ/ was calculated for each speaker using the total 97,123 observed instances of /ʃ/ regardless of prosodic position or lexical word.

Centroid frequency values of the target instances of /s/ were modeled using linear mixed-effects models in `R` [27], using the `lmer()` function from the `lme4` package [28]. The model included AVG-SH, DURATION, STRESS (unstressed vs. primary stress), POSITION (initial vs. medial), log-transformed WORDFREQUENCY (values from SUBTLEX$_{US}$ [29]). To reduce multicollinearity between predictors, continuous variables were scaled and centered at 0 (AVG-SH, DURATION, WORDFREQUENCY) and categorical variables were contrast coded, with sum-coding for STRESS (UNSTRESSED = base) and POSITION (ONSET = base). Two- and three-way interactions were included between the explanatory variables.

The model included random intercepts for SPEAKER and WORD to account for speaker-specific and word-specific variation in the acoustic measurements. As the aim of this research is to model non-segmental influences on /s/ production, the model included the coarticulatory triggers PRECEDINGPHONE and FOLLOWINGPHONE as random intercepts as well. By-subject random slopes for AVG-SH, DURATION, STRESS, and POSITION were included in the model to account for speaker-specific variation in the effects of each of the explanatory variables.

4. Results

The mixed effects model for the centroid frequency of /s/ included significant main effects of DURATION, AVG-SH and POSITION, summarized in Table 1. The main effect of DURATION ($\beta = 136.18, t = 28.21, p < 0.001$) suggests that the centroid frequency of /s/ increases as segmental duration increases. The main effect of AVG-SH ($\beta = 244.99, t = 10.97, p < 0.001$) suggests that the higher an individual's mean centroid frequency of /ʃ/, the higher the predicted centroid frequency of /s/. And the main effect of POSITION ($\beta = -19.16, t = -2/98, p < 0.01$) suggests that the centroid frequency of word-initial instances of /s/ are higher than word-medial instances of /s/.

Table 1: *Summary of main effects*

Predictor	β	SE	t	p
(int)	6077.94	32.71	185.80	$<$ 0.001
AVG-SH	244.99	22.32	10.97	$<$ 0.001
DURATION	136.18	4.83	28.21	$<$ 0.001
POSITION	-19.16	6.41	-2.98	0.002
STRESS	-17.70	20.13	-0.88	0.379
WORDFREQ	-9.71	5.05	-1.92	0.054

Significant two- and three-way interactions are summarized in Table 2. Of particular note, the interaction of DURATION and POSITION ($\beta = 13.87, t = 5.50, p < 0.001$) suggests that the difference between word-initial and medial centroid frequency values diminishes as the duration of the interval increases. This interaction is mediated by the effect of AVG-SH as suggested by the significant three-way interaction of DURATION, POSITION and AVG-SH ($\beta = -8.91, t = -4.84, p < 0.001$). In particular, the speaker's mean centroid frequency of /ʃ/ counteracts the diminishing effect that DURATION has on POSITION. As illustrated in Figure 1, the higher the mean centroid frequency of /ʃ/, the smaller the DURATION effect in reducing the difference between the prosodic position.

There is also a significant interaction between DURATION

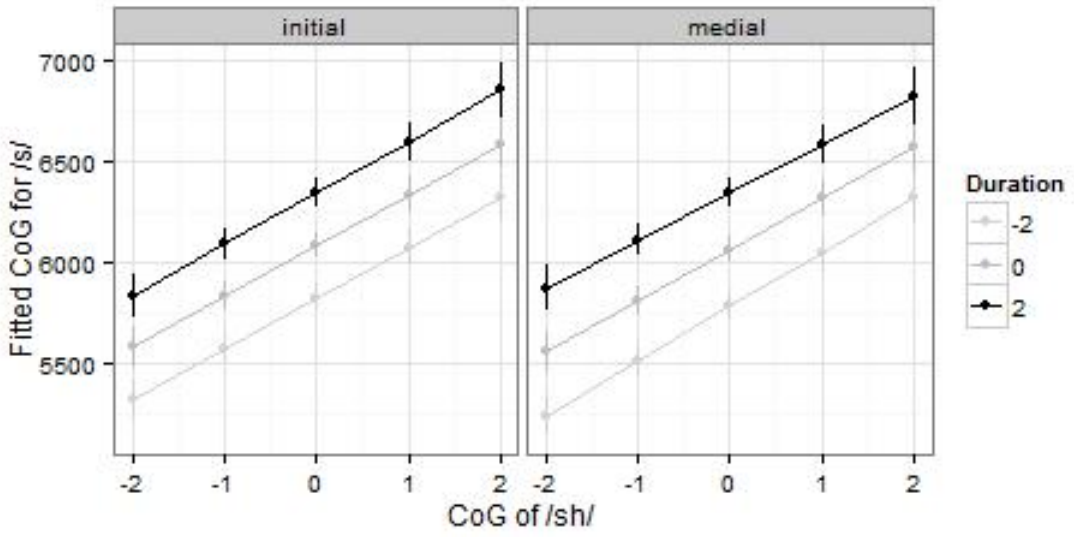

Figure 1: *Fitted centroid frequency values by the interaction of* DURATION *(color: mean duration in dark gray with two standard deviation above (black) and below (light gray) the mean),* AVG-SH *(x-axis, centered at the mean with two standard deviations above and below the mean), and* POSITION *(left panel: word-initial, right panel: word-medial).*

Table 2: *Summary of significant interactions*

Interaction	β	SE	t	p
DUR:POS	13.87	2.52	5.50	< 0.001
DUR:STRESS	17.70	2.62	6.77	< 0.001
DUR:WF	16.01	2.19	7.30	< 0.001
DUR*POS*SH	-8.91	1.84	-4.84	< 0.001
DUR*WF*SH	5.84	1.44	4.07	< 0.001
DUR*POS*STRESS	6.77	1.99	3.40	< 0.001
DUR*POS*WF	-9.44	2.44	-3.87	< 0.001
POS*WF*SH	5.25	2.21	2.38	0.017

and STRESS ($\beta = 17.70, t = 6.77, p < 0.001$), suggesting that the duration effect is enhanced by lexical stress, with greater higher centroid frequency for /s/ in stressed syllables. However, this interaction is modulated by POSITION ($\beta = 6.77, t = 3.40, p < 0.001$). That is, for medial /s/ in stressed syllables, as the segmental duration increases, centroid frequency is higher, reducing the prosodic position difference between initial and medial sibilants.

Additionally, there is a significant interaction between DURATION and WORDFREQUENCY ($\beta = 16.01, t = 7.30, p < 0.001$), suggesting that the duration effect is amplified by lexical frequency, with higher centroid frequency in more frequent words. This interaction is further amplified by AVG-SH ($\beta = 5.84, t = 4.07, p < 0.001$), suggesting a higher centroid frequency for /s/ in high frequency words for speakers with higher average /ʃ/ centroid frequencies. Conversely, the interaction between DURATION and WORDFREQUENCY is diminished by POSITION ($\beta = -9.44, t = -3.87, p < 0.001$), similar to the effect of the speaker's mean centroid frequency for /ʃ/ on this interaction. As illustrated in Figure 2, the higher the word frequency, the smaller the duration effect in reducing the difference between word-initial and medial instances of /s/.

5. Discussion

The research aims of this project are to determine the role of non-segmental factors in the production of /s/ in English. The strongest effect, the raising effect of duration, however, in some ways indirectly captures certain segmental information. This is largely due to the robust observation that segments with longer durations exhibit less coarticulation [9]. The presence of vowel rounding [4] or a subsequent /r/ articulation [5] can lead to lower centroid frequencies, and, as the duration of the sibilant increases, these potential dampening effects diminish. The robustness of the effect, however, suggests a role of duration in environments that cannot be exclusively be accounted for by coarticulatory explanations.

The parallel raising effect of the speaker's average centroid frequency of /ʃ/ serves to capture two important and interrelated factors. The first is physiological: a speaker with a high centroid frequency for /ʃ/ is expected to have a higher centroid frequency for /s/ due in large part to the length of that speaker's vocal tract. The second is phonological: English contrasts /s/ and /ʃ/ largely using centroid frequency. Thus if a speaker has a high frequency /ʃ/ (regardless of their vocal tract length), then there is a phonological motivation for a high frequency /s/ in order to maintain and maximize the contrast between the sibilants. Taken together, the speaker's average centroid frequency of /ʃ/ allows the SPEAKER random intercepts to capture individual variation that is not related to these two factors.

The effect of word position suggests a form of prosodic strengthening manifested by a small but reliable raising in the centroid frequency of sibilants at prosodic boundaries, here word-initially. This effect is similar to the prosodic lengthening that also indirectly leads to a higher centroid frequency word-initially. As /s/ and /ʃ/ are contrasted using primarily centroid frequency [23], raising the centroid frequency would serve to enhance the place of articulation contrast between the sibilants in American English. These results provide novel evidence for prosodic strengthening of place contrasts for sibilants, contra

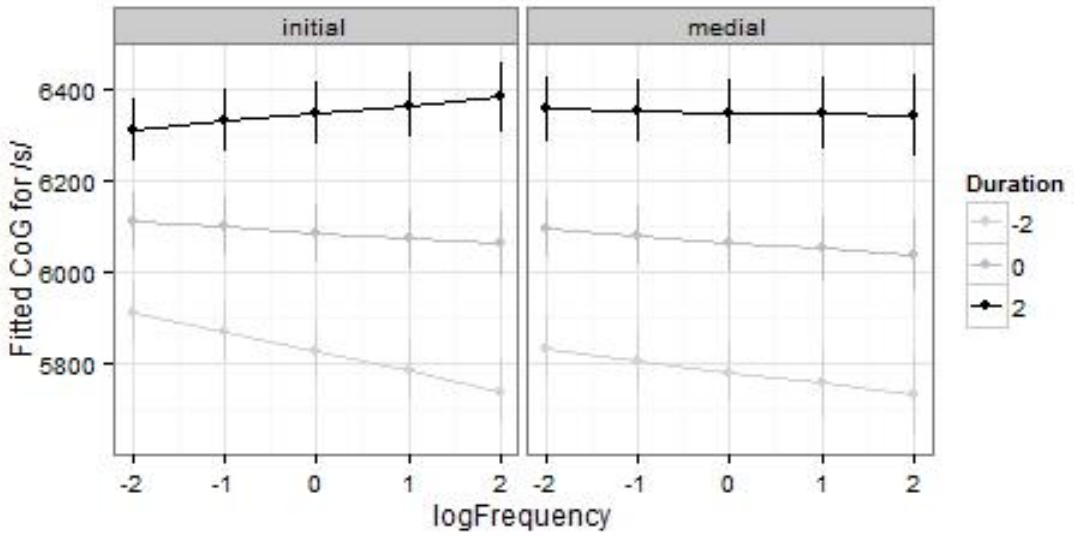

Figure 2: *Fitted centroid frequency values (y-axis) by the interaction of* DURATION *(color: mean duration in dark gray with two standard deviation above (black) and below (light gray) the mean),* WORDFREQUENCY *(x-axis, centered at the mean with two standard deviations above and below the mean), and* POSITION *(left panel: word-initial, right panel: word-medial).*

Clayards and Knowles [15] who found evidence only for enhanced voicing contrasts but not place contrasts in American sibilants in accented positions.

These two distinct prosodic strengthening effects – raising word-initially and in segments with longer durations – are equalized by the interaction of duration and position, in which the contrast between initial and medial positions diminishes as the duration of the segment increases. This suggests that as the duration difference between the initial and medial positions is neutralized, so is the the effect of prosodic position, providing evidence that the duration effect is more than just an indirect effect of prosodic strengthening. Again, one possible explanation is coarticulatory, as the duration of the sibilant increases the potential for coarticulatory influences diminishes, not only word-initially but regardless of prosodic position. The nature of the interaction suggests that the duration effect outweighs the boundary effect, which may speak to the breadth of research examining duration effects and the relative dearth of evidence for centroid frequency dampening effects in sibilants.

Interestingly, the inclusion of the speaker's average centroid frequency of /ʃ/ counteracts the neutralizing effect of the interaction of duration and position, suggesting a reinforcement of the contrast between initial and medial sibilants. This unexpected result suggests that speakers with a high centroid frequency for /ʃ/, most likely correlating with the female speakers in our corpus, are more likely to make a prosodically conditioned contrast between initial and medial sibilants. Similarly, the inclusion of word frequency also counteracts the neutralizing effect of the interaction of duration and positions, reinforcing the contrast between initial and medial instances of /s/ for high frequency words. This effect may suggest that gestural reductions – here perhaps increased coarticulation from subsequent vowel rounding or consonant gestures – are predicted at higher rates in word-medial positions in higher frequency words.

6. Conclusions

Our investigation found evidence for significant centroid frequency raising effects of segment duration and prosodic position. There is also a great degree of individual variability in the effect these factors have on speakers' centroid frequency values. Furthermore, our findings suggest that non-prosodic factors, namely the speakers average centroid frequency of /ʃ/ and lexical frequency, seek to preserve and enhance the prosodic contrast between word-initial and word-medial instances of /s/ in environments in which that contrast may be reduced. Finally, these findings provide evidence supporting the notion of contrast enhancement as a form of prosodic strengthening, with higher, i.e. less /ʃ/-like, centroid frequencies in word-initial positions than word-medially. Contra previous findings [15], these results provide evidence that place of articulation contrast enhancement can be observed in sibilants.

7. Acknowledgements

The authors would like to thank Carissa Abrego-Collier, Betsy Pillion, Kathryn Franich, and other members of the SCOTUS Project at the University of Chicago for their contributions in creating the SCorpus.

8. References

[1] J. Stuart-Smith, "Empirical evidence for gendered speech production: /s/ in Glaswegian," *Laboratory Phonology*, vol. 9, pp. 65–86, 2007.

[2] D. Byrd, "Preliminary results on speaker-dependent variation in the TIMIT database," *Journal of the Acoustical Society of America*, vol. 92, no. 1, pp. 593–596, 1992.

[3] R. S. Newman, S. A. Clouse, and J. L. Burnham, "The perceptual consequences of within-talker variability in fricative production," *Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1181–1196, 2001.

[4] V. Mann and B. H. Repp, "Influence of vocalic context on per-

ception of the [ʃ]-[s] distinction," *Perception & Psychophysics*, vol. 28, no. 3, pp. 212–228, 1980.

[5] J. Mielke, A. Baker, and D. Archangeli, "Variability and homogeneity in American English /r/ allophony and /s/ retraction," *Laboratory Phonology*, vol. 10, pp. 699–719, 2010.

[6] A. Baker, D. Archangeli, and J. Mielke, "Variability in American English s-retraction suggests a solution to the actuation problem," *Language Variation and Change*, vol. 23, pp. 347–374, 10 2011.

[7] B. S. Phillips, "Lexical diffusion, lexical frequency, and lexical analysis," in *Frequency and the Emergence of Linguistic Structure*, J. Bybee and P. Hopper, Eds. Amsterdam: John Benjamins, 2001, pp. 123–126.

[8] S. Lin, P. S. Beddor, and A. W. Coetzee, "Gestural reduction, lexical frequency, and sound change: A study of post-vocalic /l/," *Laboratory Phonology*, vol. 5, no. 1, pp. 9–26, 2014.

[9] C. Fougeron and P. Keating, "Articulatory strengthening at edges of prosodic domains," *The Journal of the Acoustical Society of America*, vol. 101, no. 6, pp. 3728–3740, 1997.

[10] T. Cho and P. Keating, "Articulatory and acoustic studies on domain-initial strengthening in Korean," *Journal of Phonetics*, vol. 29, pp. 155–190, 2001.

[11] T. Cho, "Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /ɑ,i/ in English," *Journal of the Acoustical Society of America*, vol. 117, no. 6, pp. 3867–3878, 2005.

[12] J. Cole, H. Kim, H. Choi, and M. Hasegawa-Johnson, "Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech," *Journal of Phonetics*, vol. 35, no. 180–209, 2007.

[13] T. Cho and P. Keating, "Effects of initial position versus prominence in English," *Journal of Phonetics*, vol. 37, no. 4, pp. 466–485, 2009.

[14] P. Keating, R. Wright, and J. Zhang, "Word-level asymmetries in consonant articulation," in *UCLA Working Papers in Phonetics*, 1999, pp. 157–173.

[15] M. Clayards and T. Knowles, "Prominence enhances voicelessness and not place distinction in English voiceless sibilants," in *Proceedings of the 18th International Congress of Phonetic Sciences*, 2015. [Online]. Available: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0934.pdf

[16] S. Manuel, "The role of contrast in limit vowel-to-vowel coarticulation in different languages," *Journal of the Acoustical Society of America*, vol. 88, pp. 1286–1298, 1990.

[17] P. P. K. Mok, "Does vowel inventory density affect vowel-to-vowel coarticulation?" *Language and Speech*, vol. 56, no. 2, pp. 191–209, 2013.

[18] E. Chodroff and C. Wilson, "Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English," *Journal of Phonetics*, vol. 61, pp. 30–47, 2017.

[19] D. Chen, Y. Halberstam, and A. C. L. Yu, "Perceived masculinity predicts U.S. Supreme Court outcomes," *PLoS ONE*, vol. 11, no. 10, p. e0164324, 2016.

[20] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus." *Journal of the Acoustical Society of America*, vol. 123, no. 5, 2008.

[21] S. J. Young, *The HTK Hidden Markov Model Toolkit: Design and philosophy*, Entropic Cambridge Research Laboratory, Ltd., 1994.

[22] *CMUdict: The CMU Pronouncing Dictionary*. Carnegie Mellon University, 2008.

[23] C. Shadle and S. Mair, "Quantifying spectral characteristics of fricatives," in *Proceedings of the 4th International Conference on Spoken Language Processing*, 1996.

[24] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[25] C. DiCanio, *Time averaging for fricatives*, Haskins Laboratories and SUNY Buffalo, 2013.

[26] C. H. Shadle, "On the acoustics and aerodynamics of fricatives," in *The Oxford Handbook of Laboratory Phonology*, A. C. Cohn, C. Fougeron, and M. K. Huffman, Eds., 2012, pp. 511–526.

[27] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, 2015.

[28] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Stastical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[29] M. Brysbaert and B. New, "Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English," *Behavior Research Methods*, vol. 41, no. 4, pp. 977–990, 2009.