Quantifying robustness of the /t/-/k/ contrast using a single, static spectral feature

Allison A. Johnson
Department of Hearing and Speech Sciences, University of Maryland—College Park, 0121 Taliaferro Hall, Chapel Drive, College Park, Maryland 20724, USA
ajohns51@umd.edu

Patrick F. Reidy
Callier Center for Communication Disorders, University of Texas at Dallas, 1966 Inwood Road, Dallas, Texas 75235, USA
reidy@utdallas.edu

Jan R. Edwards
Department of Hearing and Speech Sciences, University of Maryland—College Park, 0121 Taliaferro Hall, Chapel Drive, College Park, Maryland 20724, USA
edwards@umd.edu

Abstract: Dynamic spectral shape features accurately classify /t/ and /k/ productions across speakers and contexts. This paper shows that word-initial /t/ and /k/ tokens produced by 21 adults can be differentiated using a single, static spectral feature when spectral energy concentration is considered relative to expectations within a given speaker and vowel context. Centroid and peak frequency—calculated from both acoustic and psychoacoustic spectra—were compared to determine whether one feature could reliably differentiate /t/ and /k/, and, if so, which feature best differentiated them. Centroid frequency from both acoustic and psychoacoustic spectra accurately classified productions of /t/ and /k/.

[BHS]
Date Received: June 2, 2018 **Date Accepted:** July 22, 2018

1. Introduction

Over the past several decades, researchers have focused on classifying stop consonants in order to identify invariant, acoustic cues to place of articulation. This work was initially driven by a theoretical interest in how listeners achieve perceptual constancy despite variability in the acoustic signal (e.g., Blumstein and Stevens, 1979; Kewley-Port, 1983; Kewley-Port and Luce, 1984; Stevens and Blumstein, 1978.) The search for invariant features was further inspired by a practical interest in improving automatic speech recognition systems (e.g., Nossair and Zahorian, 1991). Clinical researchers have also expressed the need for a classification approach to quantify the spectral distance between model token productions and atypical ones (e.g., Forrest *et al.*, 1988). Quantifying an individual speaker's robustness of contrast would support more reliable measures of covert contrasts, improve tracking of developmental changes, and provide more fine-grained descriptions of productions (compared to categorical transcription) upon which to compare groups. Our ultimate goal is to establish a clinically viable method for quantifying the degree of overlap between two sound categories within a speaker. As a first step toward this goal, we focus here on fluent productions of target /t/ and /k/ by healthy adult speakers, which should be differentiable using some feature(s) computed from the spectra of these productions. The purpose of this paper is to determine whether a single, static spectral feature, such as centroid or peak frequency, can sufficiently classify /t/ and /k/ tokens, given knowledge of the speaker and vowel context, and to compare features computed from acoustic versus psychoacoustic spectra.

There is a general consensus in the literature that critical information for differentiating place of articulation in English stop consonants is concentrated in the release burst. Early work by Stevens and Blumstein (1978) (see also Blumstein and Stevens, 1979) identified a unique spectral shape of the burst for each place of articulation, including diffuse-rising for alveolars and a compact, mid-frequency spectral peak for velars. Kewley-Port (1983) expanded on these "templates" with time-varying features, and Kewley-Port and Luce (1984) further improved classification accuracy by demarcating speaker-specific values for "mid-frequency." Limitations of this early work include few speakers, poor generalizability across speakers, and less reliable classification across different vowel contexts. Furthermore, these early classification schemes relied on time-consuming human visual judgments rather than objective methods.

Time-varying spectral features have since been quantified and gained additional support in the automatic classification literature. Nossair and Zahorian (1991) identified stop consonants with 93.7% accuracy from 20 dynamic, global features (discrete cosine transform coefficients) extracted from a 60-ms window around the stop burst. They achieved this high classification accuracy for 30 different speakers (including men, women, and children), across both voiced and voiceless consonants in a variety of vowel contexts.

Forrest *et al.* (1988) also achieved a high classification accuracy (93%) for voiceless stops across 10 speakers using only three spectral features (mean, skew, and kurtosis) extracted from a series of analysis windows spanning 40 ms. Relatively high spectral kurtosis was a defining feature for /k/, and negative spectral skew was the essential feature for /t/. The inclusion of skew and kurtosis possibly improved classification accuracy across speakers because these moments remove differences in spectral means that arise between speakers producing the same target sound, which accomplishes a rough speaker normalization [see Forrest *et al.* (1988), p. 118].

For Nossair and Zahorian's (1991) purpose of developing a speaker-independent, automatic speech classifier, a high-dimensional feature space is well motivated. However, a low-dimensional feature space—or even a single feature—may be more feasible for clinicians to obtain and interpret. A common theme among previous works is that multiple features calculated over relatively long analysis windows have been necessary to achieve accurate and reliable classification of stop consonants across speakers and vowel contexts. However, speaker characteristics and coarticulation influence spectral shapes (and the features computed therefrom). The location of "mid-frequency" varies depending on an individual's vocal tract. Similarly, spectral peak frequency and spectral kurtosis can fluctuate within a speaker for /k/ in front- versus back-vowel contexts. The spectrum for /k/ can present with two spectral peaks due to resonances in both the front and back oral cavities. The spectrum for /t/ can also have a prominent peak near the speaker's F2 locus, and the energy concentration can shift if the tongue dorsum raises in preparation for a high front vowel.

Regardless of vowel context or speaker, /k/ is formed farther back in the mouth than /t/. Thus, theoretically, the overall concentration of energy in the spectrum for /k/ should be lower than that for /t/ for a particular speaker in a given vowel context. McMurray and Jongman (2011) recently conducted a comprehensive review of acoustic features for fricatives and found that none were entirely invariant. They identified fricatives using a compensation model [computing cues relative to expectations (C-CuRE)], which used hierarchical regressions to capitalize on acoustic variability in the signal, and adjusted category expectations relative to known indexical and contextual information (such as speaker identity and vowel context).

Given the current power of mixed-effects modeling, we now have the capacity to process large sets of non-independent observations and examine variability both within and across speakers (for an overview on mixed-effects modeling, see Brauer and Curtin, 2017). It is possible that when speaker and vowel context are statistically controlled, a single static spectral feature calculated over a relatively short analysis window will be sufficient to classify /t/ and /k/ tokens.

Forrest *et al.* (1988) successfully used centroid frequency in combination with skew and kurtosis to classify stop consonants. Because energy concentration should be at higher frequencies for /t/ than /k/ due to different places of articulation, centroid frequency may be suitable to differentiate /t/ from /k/ within a speaker. On the other hand, centroid frequency may not characterize the frequency-location of energy concentration in the spectrum very well if the distribution is bimodal. Thus, the frequency of the most prominent peak (henceforth, "peak frequency") may provide better evidence for the location of the constriction.

With the goal of identifying psychoacoustically relevant features, some researchers have also explored the effect of transforming an acoustic spectrum prior to computing features from it. Typically, these transformations seek to model some process of the auditory system, such as compression of the frequency scale or wider bandwidths at higher frequencies. For example, Forrest *et al.* (1988) transformed the Hertz frequency scale of acoustic spectra to the Bark frequency scale, but did not apply any transformation to model the different bandwidths of auditory filters; classification of voiceless stops was poorer when features were computed from Bark-scale spectra than from Hertz-scale spectra. Kewley-Port and Luce (1984) passed acoustic spectra through a bank of bandpass filters that modeled the different bandwidths of auditory filters, then transformed the Hertz scale to the mel scale; however, classification accuracy from transformed spectra was not compared to untransformed spectra, so it is difficult to assess the utility of these transformations on the classification of voiceless stops.

We address this lacuna by computing two features—centroid and peak frequency—from both acoustic spectra and from transformed spectra (henceforth, "psychoacoustic spectra") that were passed through a gammatone filter bank that models both the frequency-scale compression and the differential frequency selectivity of the auditory system. Our purpose is to determine whether a single, static feature can sufficiently differentiate /t/ and /k/ productions when speaker identity and vowel context are statistically controlled. This research is driven by the need for a standardized approach to quantify robustness of an individual's /t/-/k/ contrast that can be applied quickly, easily, and objectively by researchers and clinicians alike.

2. Methods

Twenty-one adult participants (10 women, 11 men; mean age: 21 years, range: 20–29 years) were recruited to participate from Minneapolis, MN. All participants were monolingual, native speakers of Mainstream American English with self-reported normal hearing and no history of speech or language disorders.

Stimuli for the experiment—a picture-prompted, auditory word repetition task—consisted of familiar words presented in isolation. Stimuli were recorded by an adult female speaker in a sound-treated lab setting, and recordings were normalized for amplitude. Words were also represented visually by archetypal, high-quality photographs obtained from online sources and edited for consistency in size and background. Stimuli were organized into two wordlists, each with 16 /t/-initial and 16 /k/-initial tokens. Each wordlist also included either 58 or 94 filler-words that did not begin with /t/ or /k/. Tokens were balanced across front- and back-vowel contexts. Possible front vowels included /i ɪ e ɛ æ/. Possible back vowels included /u ʊ o ʌ ɑ/ and the diphthongs /aɪ aʊ/. Despite known regional variations, our speakers and participants consistently produced these diphthongs with an initial back vowel.

All testing was completed in a sound-treated recording booth. During the experimental task, participants sat in front of a computer screen positioned approximately six inches away from a Shure SM81 cardioid condenser microphone with a custom pop filter. Words were presented in a pseudo-randomized order across participants, with steps taken to ensure target words were not repeated on consecutive trials. Visual stimuli appeared on the screen while auditory stimuli played over loudspeakers. At word-offset, participants repeated the word into the microphone, and an experimenter recorded the session using a Marantz PMD671 solid-state recorder at a sampling frequency of 44 100 Hz.

Five participants completed one wordlist (32 productions per speaker), and sixteen participants completed both wordlists (64 productions per speaker). Productions were excluded from analyses if there was background noise obscuring the release burst, or if voice-onset time was less than 20-ms. The final number of analyzable tokens was 1155.

Coding was done in PRAAT (Boersma and Weenink, 2018). The first author transcribed place of articulation for all /t/ and /k/ tokens. Then, she marked locations on the waveform corresponding to the release burst and the onset of voicing. The release-burst was defined as the first transient-noise spike following a period of silence that coincided perceptually with the release of an oral constriction. The onset of voicing was defined as the first upward swing from the zero-crossing followed by a stable, quasi-periodic pattern of voicing. A second trained phonetician coded a random 20% of the files for reliability purposes. Reliability between the two coders was high: agreement for transcriptions was 100%. Root-mean-square (RMS) values were calculated to determine differences in locations of burst and VOT tags. For burst locations, RMS error was 0.0023 ms, and for VOT locations, RMS error was 0.0039 ms.

Acoustic and statistical analyses were carried out in the R programming environment (R Core Team, 2013), using custom scripts. The method for computing acoustic and psychoacoustic spectra was identical to that reported in Reidy (2016), to which the reader is referred for a comprehensive description with references. For each token, 5 ms prior to the burst tag through 20 ms after the tag defined a 25-ms analysis window. Within this window, the acoustic spectrum of the waveform was estimated with an eighth-order multitaper spectrum. To transform an acoustic spectrum into a psychoacoustic spectrum, it was passed through a filter bank that modeled how the auditory periphery logarithmically compresses the frequency scale and how it differentially resolves frequency components across the audible range (see top panel of Fig. 1). This filter bank comprised 361 fourth-order gammatone filters whose center frequencies were equally spaced every 0.1, from 3 to 39, along the ERB_N number scale (Glasberg and Moore, 1990). The bandwidth of each filter was set to 1.019 times the equivalent rectangular bandwidth of that filter's center frequency in Hz. Each gammatone filter acted on an input spectrum as a bandpass filter. Finally, the psychoacoustic spectrum was constructed by summing the total energy (or "auditory excitation") at the output

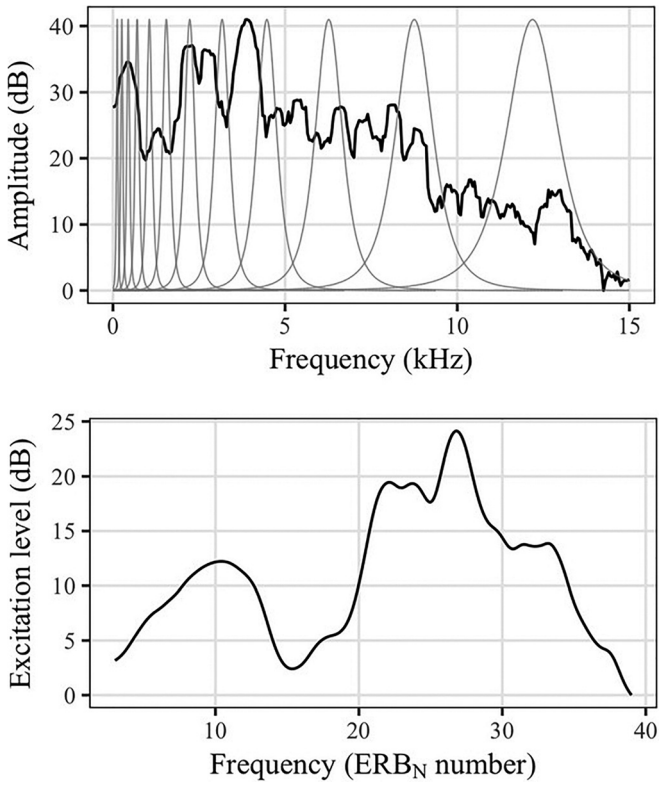

Fig. 1. (Top) Acoustic spectrum of a production of /t/ estimated with an eighth-order multitaper spectrum (centroid = 3.740 kHz, peak = 3.918 kHz). Frequency responses of 12 gammatone filters of different center frequencies are shown, in grey, overlaid on the spectrum. (Bottom) The psychoacoustic spectrum resulting from passing the spectrum in the top panel through a 361-channel gammatone filter bank model of the auditory periphery (centroid = 26.0 ERB_N, peak = 26.8 ERB_N).

of each filter and plotting these excitation levels against the filters' center frequencies in ERB_N (see bottom panel of Fig. 1).

Centroid and Peak frequency were computed from both the acoustic spectra (within the 0.926–9.777 kHz range) and the psychoacoustic spectra (within the 15–35 ERB_N number range). To compute Centroid frequency, the values of a (psycho)acoustic spectrum were normalized so they summed to 1. The normalized (psycho)acoustic spectrum was then treated as a probability mass function over frequency, and the Centroid frequency was the distribution's mean value. Peak frequency was the frequency of the (psycho)acoustic spectrum with the greatest amplitude. Thus, there were four features computed from each token: Centroid (kHz), centroid (ERB_N), peak (kHz), and peak (ERB_N). Prior to statistical analysis, the values for each feature were centered by subtracting the group mean value for that feature.

Our modeling procedure followed that of Holliday *et al.* (2015), which quantifies the degree of category overlap within an individual speaker. We used four mixed-effects logistic regression models—one for each spectral feature—to predict each token's target consonant (either /t/ or /k/). Then, we used two additional mixed-effects logistic regression models to formally compare the accuracy of predictions made by each model. All models were fit using the R LME4 package (Bates *et al.*, 2014).

Logistic regression models are based on the logarithm function, and they are an appropriate statistical choice when the outcome variable is binary or binomially distributed, as in this case where the outcome variable is a prediction of either /t/ or /k/ (for more on analyzing categorical data, see Jaeger, 2008). The dependent variable in a logistic regression model is a log-likelihood ratio, which can be used to determine the probability of one outcome or the other given values of the independent variables. Mixed-effects models, which refer to models that contain at least one conditional random effect, are appropriate to use when data are non-independent, as in this case when multiple tokens are obtained from each speaker. Random effects by-participant produce participant-level adjustments for predictors, which can be used to obtain unique, individually fit models for each participant [for more information on mixed-effects modeling, see Bates *et al.* (2015) or Brauer and Curtain (2017)].

An example of the formula used to make predictions is shown in Eq. (1),[1] where the subscripts *i* and *j* range over items and speakers, respectively,

$$\log\left(\frac{/t/}{1-/t/}\right)_{ij} = \beta_0 + \beta_1 \times Centroid(\text{kHz})_{ij} + \beta_2 \times VowelContext_{ij} + \beta_3 \times Centroid(\text{kHz})_{ij} \times VowelContext_{ij} + u_{0j} + u_{1j} \times Centroid(\text{kHz})_{ij} + \varepsilon_{ij}. \quad (1)$$

This model predicted the log-likelihood that the target consonant for a given token was /t/ based on fixed effects of the group-wide intercept (β_0), centroid frequency computed from acoustic spectra (β_1), vowel context (β_2), the interaction between centroid frequency and vowel context (β_3), as well as the speaker-level random intercept (u_{0j}) and slope for centroid frequency (u_{1j}). Three additional models with homologous structures to that in Eq. (1) were fit to make predictions based on the other spectral features of interest. When the predicted log-likelihood was greater than 0, the model predicted the target consonant to be /t/; otherwise, the model classified the token as /k/. If the prediction matched the target consonant, the token was assigned a 1 for *predicted accuracy*. If the model made an incorrect prediction, *predicted accuracy* was 0. Predictions made by each of the four models were highly accurate. Results are shown in Table 1.

To determine which of the four spectral features best differentiated /t/ and /k/, we ran two additional mixed-effects logistic regression models that compared accuracy of predictions. We added two variables to our dataset: *Representation* and *feature*. *Representation* was either "acoustic" or "psychoacoustic," referring, respectively, to whether the spectral feature was computed from an acoustic or psychoacoustic spectrum. *Feature* was either "centroid" or "peak." The formula comparing accuracy of predictions for the spectral features is shown in Eq. (2),[2]

$$\log\left(\frac{PredictedAccuracy}{1-PredictedAccuracy}\right)_{ij} = \beta_0 + \beta_1 \times Representation_{ij} + \beta_2 \times Feature_{ij} + \beta_3 \times Representation_{ij} \times Feature_{ij} + u_{0j} + \varepsilon_{ij}. \quad (2)$$

Models 1 and 2 predicted the log-likelihood that a prediction was accurate based on fixed effects of group-wide intercept (β_0), representation (β_1), feature (β_2), the interaction between representation and feature (β_3), and the speaker-level random intercept (u_{0j}).

The difference between the two models was the reference category. In model 1, the reference level for representation was "psychoacoustic," and the reference level for feature was "centroid." For model 1 with centroid (ERB_N) as the reference category, the main effect of representation characterized the difference in accuracy of predictions based on centroid (ERB_N) versus centroid (kHz), and the main effect of feature characterized the difference in predictions based on centroid (ERB_N) versus peak (ERB_N). In model 2, the reference category was peak (kHz), so the main effect of representation characterized the difference between peak (kHz) and peak (ERB_N), and the main effect of feature characterized the difference between peak (kHz) and centroid (kHz). Because we ran two, re-leveled models testing the same data, we used an adjusted alpha-level, $p = 0.0025$, to denote significance.

3. Results

Model 1 [reference: Centroid (ERB_N)] showed significant main effects of intercept ($\hat{\beta}_0 = 3.44$, $SE = 0.25$, $z = 13.72$, $p < 0.001$) and feature ($\hat{\beta}_2 = -0.75$, $SE = 0.17$, $z = -4.52$, $p < 0.001$). The main effect of representation and the interaction were not significant. The results of model 1 indicate that accuracy of predictions decreased significantly when peak (ERB_N) was used compared to centroid (ERB_N), but there was no difference between centroid (ERB_N) and centroid (kHz).

Table 1. Overall accuracy of predictions made by each model (one model for each spectral feature), and the accuracy of predictions by target consonant and vowel context.

Spectral feature	Target /t/		Target /k/	
Centroid (kHz)	Front	Back	Front	Back
95%	94%	95%	95%	94%
Centroid (ERB_N)	Front	Back	Front	Back
95%	91%	98%	94%	96%
Peak (kHz)	Front	Back	Front	Back
89%	83%	87%	93%	93%
Peak (ERB_N)	Front	Back	Front	Back
90%	84%	94%	91%	90%

Model 2 [reference: Peak (kHz)] showed significant main effects of intercept ($\hat{\beta}_0 = 2.58$, $SE = 0.23$, $z = 11.13$, $p < 0.001$) and feature ($\hat{\beta}_2 = 0.83$, $SE = 0.16$, $z = 5.09$, $p < 0.001$). The main effect of representation and the interaction were not significant. The results of model 2 indicate that accuracy of predictions increased significantly when centroid (kHz) was used compared to peak (kHz), but there was no difference between peak (kHz) and peak (ERB_N).

Taken together, these results suggest that centroid frequency better differentiated /t/ and /k/ than peak frequency, but there was no difference between spectral features computed from acoustic spectra versus psychoacoustic spectra.

4. Discussion

The central finding of this paper is that word-initial /t/ and /k/ tokens in the context of 12 different vowels produced by 21 different speakers were differentiated with 95% accuracy using a single, static, spectral feature when vowel context and speaker identity were statistically controlled. Centroid frequency yielded higher classification accuracy than peak frequency, and features computed from psychoacoustic spectra were equally successful as those from acoustic spectra.

We used a mixed-effects logistic regression model with spectral feature and vowel context as fixed effects, and speaker identity as a grouping factor to differentiate /t/ and /k/ productions. This approach was described by Holliday *et al.* (2015) as a way to quantify robustness of an individual's /s/-/ʃ/ contrast. The primary objective of this type of model—one that is not independent of speaker or vowel context—is to quantify the relationship between productions of two target categories within a speaker. The model uses by-participant random effects to make individualized predictions, and ultimately the variable of interest derived from the model is the percentage of tokens correctly predicted for each speaker. This variable indexes one notion of distance between sets of productions (cf. the mean Mahalanobis distance between two sets of points, the distance between the means of two sets of points, or the discriminability between the sets of points).

Researchers and clinicians alike can use this approach to determine the extent to which two sets of productions are separable within a speaker and make comparisons over time or across groups. For example, using a similar logistic regression classifier, Nicholson *et al.* (2015) found that robustness of the /s/-/ʃ/ contrast increases with age and vocabulary. Todd *et al.* (2011) showed that children with cochlear implants produce less robust /s/-/ʃ/ contrasts than peers with normal hearing, even when tokens are transcribed as correct. These studies speak both to the importance of using continuous, acoustic measures to characterize productions, and to the utility of a within-participant measure of robustness of contrast. The aim of this study was to empirically compare spectral features that could differentiate /t/ and /k/ productions within and across adult speakers, and that would be accessible to a range of professionals.

There are some limitations to this classification approach. First, it requires several tokens per category per context per speaker for the mixed-effects logistic regression model to work reliably. This increases the time required to collect and code production data. However, the coding procedure was streamlined and largely automated, so each token was transcribed and tagged in approximately one minute. Second, our method of coding vowel context qualitatively as "front" or "back" based on the target word is not always feasible. The vowel could be centralized in some productions or dialects, misarticulated by children, or the target word may not be known. In these cases, an on-line coding procedure to label the vowel for each token may be necessary. Quantitative representations of formant transitions could also be incorporated, but would substantially increase the amount of time and expertise required to obtain reliable measurements. Finally, using a logistic regression model that includes an indexical grouping factor is not well suited for all classification purposes. This approach would not translate easily to applications with the goal of low-resource, fully automatic recognition and classification of stop consonants.

Previous work (e.g., Kewley-Port, 1983; Nossair and Zahorian, 1991) suggested that dynamic, global spectral shape features are superior to static features for identifying stop consonants across speakers and contexts. McMurray and Jongman (2011) found that no acoustic parameters were unaffected by context for fricatives. Perhaps dynamic features calculated over relatively long analysis windows, especially ones that overlap with following vowel, serve as a compensation mechanism for variability. We acknowledge the importance of dynamic cues in differentiating stop consonants, and we submit that including a categorical variable of vowel context provides a sufficient, yet simpler approach for encoding dynamic features. The success of our single, static feature (calculated purposefully from a window excluding any voicing)

supports the idea that differences in energy concentration in the burst alone provide sufficient information to differentiate /t/ from /k/, when those values are considered relative to expectations for an individual speaker within a given vowel context.

Future work could compare traditional dynamic features to the current approach and determine whether there are significant performance benefits; assess clinicians' ability and willingness to implement our classification approach compared to one that relies on dynamic features; and determine whether centroid frequency is also sufficient for classifying children's /t/ and /k/ productions, which are notoriously more variable. Finally, it will be important to validate the current findings with a perceptual measure, as in Holliday *et al.* (2015). Perhaps acoustic features computed from acoustic and psychoacoustic spectra yield equivalent classification accuracy, but features from one representation better align with listeners' perceptual ratings.

Acknowledgments

Thanks to all of the participants and the Learning to Talk lab members who made this work possible. This work was supported by Grant No. NIDCD R01 02932 to J.R.E., Mary E. Beckman, and Benjamin Munson; Grant No. NICHD P30 HD03352 to the Waisman Center; T32 Training Grant No. DC05359-10 to Susan Ellis Weismer; and NSF Grant No. 1449815 to Colin Phillips. We are also thankful to the anonymous reviewers for their thoughtful comments, which led to enhanced clarity and quality of this paper.

References and links

[1]R code for implementing Eq. (1): `glmer(formula = Target Consonant ~ Centroid.kHz * Vowel Context + (1 + Centroid.kHz | Participant), data = Adult Productions, family = 'binomial').`

[2]R code for implementing Eq. (2): `glmer(formula = Predicted Accuracy ~ Representation * Feature + (1 | Participant), data = Adult Productions, family = 'binomial').`

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2014**). "lme4: Linear mixed-effects models using Eigen and S4," R package version 1.1-11.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**(1), 1–48.

Blumstein, S. E., and Stevens, K. N. (**1979**). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. **66**(4), 1001–1017.

Boersma, P., and Weenink, D. (**2018**). "Praat: Doing phonetics by computer" [computer program], version 6.0.30, http://www.praat.org/ (Last viewed June 1, 2018).

Brauer, M., and Curtin, J. (**2017**). "Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items," Psychol. Meth. (published online).

Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. N. (**1988**). "Statistical analysis of word-initial voiceless obstruents: Preliminary data," J. Acoust. Soc. Am. **84**(1), 115–123.

Glasberg, B., and Moore, B. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**(1), 103–138.

Holliday, J., Reidy, P., Beckman, M., and Edwards, J. (**2015**). "Quantifying the robustness of the English sibilant fricative contrast in children," J. Speech Lang. Hear. Res. **58**(3), 622–637.

Jaeger, T. F. (**2008**). "Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models," J. Mem. Lang. **59**(4), 434–446.

Kewley-Port, D. (**1983**). "Time-varying features as correlates of place of articulation in stop consonants," J. Acoust. Soc. Am. **73**(1), 322–335.

Kewley-Port, D., and Luce, P. A. (**1984**). "Time-varying features of initial stop consonants in auditory running spectra: A first report," Atten. Percept. Psychophys. **35**(4), 353–360.

McMurray, B., and Jongman, A. (**2011**). "What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations," Psychol. Rev. **118**(2), 219–246.

Nicholson, N., Reidy, P., Munson, B., Beckman, M. E., and Edwards, J. (**2015**). "The acquisition of English lingual sibilant fricatives in very young children: Effects of age and vocabulary size on transcribed accuracy and acoustic differentiation," in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow, Scotland.

Nossair, Z., and Zahorian, S. (**1991**). "Dynamic spectral shape features as acoustic correlates for initial stop consonants," J. Acoust. Soc. Am. **89**(6), 2978–2991.

R Core Team (**2013**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/ (Last viewed June 1, 2018).

Reidy, P. F. (**2016**). "Spectral dynamics of sibilant fricatives are contrastive and language specific," J. Acoust. Soc. Am. **140**(4), 2518–2529.

Stevens, K. N., and Blumstein, S. E. (**1978**). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**(5), 1358–1368.

Todd, A. E., Edwards, J. R., and Litovsky, R. Y. (**2011**). "Production of contrast between sibilant fricatives by children with cochlear implants," J. Acoust. Soc. Am. **130**(6), 3969–3979.