Burst spectrum as a cue for the stop voicing contrast in American English

Eleanor Chodroff[a] and Colin Wilson
Department of Cognitive Science, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218

(Received 22 May 2014; revised 2 September 2014; accepted 12 September 2014)

Voicing contrasts in stop consonants are expressed by a constellation of acoustic cues. This study focused on a spectral cue present at burst onset in American English labial and coronal stops. Spectral shape was examined for word-initial, prevocalic stops of all three places of articulation in a laboratory production study and a large corpus of continuous read speech. Voiceless labial and coronal stops were found to have greater energy at higher frequencies in comparison to homorganic voiced stops, a difference that could not be attributed to aspiration in the voiceless stops or modal phonation in the voiced, while no consistent effect was found for dorsal stops. This pattern was found with various methods of spectral estimation (time-averaged and multitaper spectra) and measures of spectral energy concentration (center of gravity and spectral peak) for both linear and auditorily based frequency scales. Perceptual relevance of the spectral cue was tested in laboratory and online experiments with continua created by crossing burst shape and voice onset time. A trading relation was observed such that voiceless identifications were more likely for tokens with higher frequency bursts. Goodness ratings indicated that burst spectrum influences category typicality for voiceless stops even when voice onset time is unambiguous. © *2014 Acoustical Society of America.*
[http://dx.doi.org/10.1121/1.4896470]

PACS number(s): 43.70.Fq, 43.71.Es [SAF] Pages: 2762–2772

I. INTRODUCTION

A central project of phonetic research is to characterize the perceptually significant acoustic correlates, or *cues*, that distinguish speech sounds. A large body of previous work has demonstrated that sounds are typically distinguished by multiple cues (e.g., Lisker, 1986; Jongman *et al.*, 2000), and that speech perception involves mapping complex cue patterns to contrastive categories (e.g., Lisker *et al.*, 1977; Oden and Massaro, 1978; Repp, 1982). The perceptual cues that signal stop voicing contrasts (e.g., /p/ vs /b/) are of particular interest, both because such contrasts are widespread in the languages of the world and they have provided the empirical basis for fundamental claims about the nature of speech perception (e.g., Liberman *et al.*, 1961; Repp, 1984).[1] While voice onset time (VOT) is the primary cue for distinguishing word-initial voiceless and voiced stops in many languages (e.g., Lisker and Abramson, 1964, 1970), previous studies have identified a number of secondary cues that also influence the perception of stop voicing, including F1 onset and transition (Liberman *et al.*, 1958), F0 contour (Haggard *et al.*, 1970), aspiration amplitude (Repp, 1979), length of the following vowel (Summerfield, 1981), and others, depending on the environment in which the stop appears (Lisker, 1986).

The present experiments provide evidence that the spectrum at the beginning of the stop burst serves as an additional secondary cue for the voicing contrast in American English (AE). As reviewed below, many previous studies of stop production and perception have provided suggestive evidence for such a cue. However, such studies have focused primarily on coronal stops rather than the entire AE stop series and have not carefully distinguished burst spectrum from other phonetic properties with which it naturally covaries (e.g., closure voicing, burst amplitude, and VOT). The main goals of this paper are to demonstrate that the shape of the burst spectrum is a correlate of voicing in stop production—with voiceless stops having energy concentrated at higher frequencies in comparison to voiced stops—and to establish that the burst spectrum functions as a perceptual cue for voicing when modal phonation, burst amplitude, and other relevant phonetic properties are controlled. Possible articulatory origins of this correlate are also considered, with differences in peak airflow at the onset of the burst being the most probable source.

The findings reported are relevant for the detailed study of voicing contrasts and the theory of speech perception. At the most general level, many models of speech perception depend upon a proper characterization of perceptual cues in order to make accurate predictions about sound discrimination, identification, typicality rating, and processes, such as lexical activation and recognition. The perception of stop voicing, in particular, has proven important for many theoretical topics, such as the structure of phonetic categories (Miller, 1994) and how multiple cues are integrated in speech perception (Oden and Massaro, 1978; Toscano and McMurray, 2010). Because the initial burst spectrum occurs at a point of rapid spectral change (or landmark; Stevens, 2002), and has already proven important for the perception of place (Blumstein and Stevens, 1979), there are reasons that it could be weighted highly in typicality judgments and

[a]Author to whom correspondence should be addressed. Electronic mail: chodroff@cogsci.jhu.edu

perceptual decisions on the voicing dimension. By establishing that the initial burst spectrum reliably correlates with stop voicing in production, and showing that this correlate is perceptually effective, the present experiments provide new opportunities to evaluate claims about the temporal localization and relative salience of cues for speech perception.

A. Previous production studies

As far as we are aware, Halle *et al*. (1957) were the first to suggest a relationship between burst spectrum and stop voicing. They observed that the bursts of voiced stops have "a strong low-frequency component" with "a significant drop in level in the high frequencies," and suggest that these spectral differences are "a crucial cue," distinguishing voiced from voiceless stops. However, the evidence provided by Halle *et al*. was limited to three speakers, and the observed effect was not investigated quantitatively. Additionally, Halle *et al*. (1957) noted that their voiced productions had considerable vocal fold vibration during the closure phase of the stop. This phonetic property, which is not characteristic of word-initial voiced stops for many AE speakers (e.g., Lisker and Abramson, 1964), is likely to have continued into the burst and excited the lower formants, thus producing a lower energy concentration. The analyses in the present paper separated voiced stops with phonetic voicing, which occurred infrequently in our data set, from those in which vocal fold vibration is not evident immediately before or after the burst.

Zue (1976), in a detailed acoustic analysis of AE stop consonants, also observed that voice is reflected in the spectral peak frequency of the initial 10–15 ms of coronal stop bursts: /t/ had a mean spectral peak of 3600 Hz, higher than the mean of 3300 Hz for /d/. A similar difference was not found for dorsal /k/ vs /g/, and aspects of the recording precluded an analysis of labial stop burst spectra. As in the study of Halle *et al*. (1957), Zue's (1976) findings were due to recordings from a small number of speakers (three, all male), the reliability of the observed differences was not assessed statistically, and the presence of phonetic voicing during the closure or burst was not considered in the analysis of burst spectra.

More recently, Parikh and Loizou (2005) observed that the peak spectral frequency of AE stops, measured over the initial 10 ms of the burst, is higher on average for voiceless than for voiced stops. The average peak frequencies were found to be 1910 Hz for /p/ and 1163 Hz for /b/, with coronals stops showing a similar pattern (average peak of 5649 Hz for /t/ and 5225 Hz for /d/). As in the earlier studies discussed above, the statistical reliability of these differences was not investigated. On the basis of spectral moments calculated from the burst (Forrest *et al*., 1988; see Sec. II for details), Sundara (2005) provided statistical evidence for the same relationship between burst spectrum and voicing in word-initial coronal stops of Canadian English (CE). The voiceless CE coronal stop, /t/, had a higher average spectral mean than its voiced counterpart, /d/; this effect was significant for male CE speakers, but held only numerically for female CE speakers and for speakers of Canadian French. Related results are reported by Kirkham (2011) in a study of British English.

Phonetic studies on languages other than English have also noted correlations between energy concentration in the initial burst spectrum and stop voicing. Examining labial and coronal stops in Dutch, van Alphen and Smits (2004) reported a mean burst frequency of 830 Hz for /b/ and 1160 Hz for /p/; for coronals, the mean burst frequency was 2140 Hz for /d/ and 3540 Hz for /t/. However, unlike word-initial voiced stops in English, those in Dutch are typically produced with modal phonation during closure (Lisker and Abramson, 1964) and, as noted previously, this may excite low-frequency formants during the burst. Harrington (2010) anecdotally reported a parallel difference for German /t/ vs /d/, which may be more similar to the AE contrast insofar as German /d/ lacks closure voicing. Similarly, in a phonetic analysis of word-initial stops in Georgian, Vicenik (2010) found that voiceless labials and coronals had higher mean burst frequencies than corresponding voiceless unaspirated stops; no significant differences were found in the spectra of voiceless and voiced dorsal stops. As in English and German, voiced stops typically do not have modal phonation at the point of release (mean voicing lag = 11.5 ms, standard deviation = 5.6).

B. Previous perception studies

A small number of experiments on stop perception has provided suggestive evidence that burst spectrum influences voicing categorization. In the first such study, Keating (1979) tested the perception of Polish speakers on two cross-spliced six-step VOT continua, one of which began with the initial 7 ms of a naturally produced /t/ burst and the other with the initial portion of a natural /d/ burst. The resulting identification curves showed that the voiced-to-voiceless category boundary fell at a significantly shorter VOT for the continuum with the /t/ burst; specifically, the boundary was at 16.6 ms of VOT for the /t/ burst continuum, but at 19.9 ms for the /d/ burst continuum.

Nittrouer (1999) obtained similar results by splicing the initial 10 ms of natural /t/ and /d/ bursts onto the beginning of a nine-step VOT continuum. Identification curves obtained from children with normal and poor phonological processing both displayed an earlier voiced-to-voiceless crossover point for the stimuli beginning with the /t/ burst in comparison to those beginning with the /d/ burst; as in Keating (1979), the /t/ burst was perceptually equivalent to approximately 3 ms of additional VOT. This trading relation has also been replicated in a comparison of children with cochlear implants and those with normal hearing (Caldwell and Nittrouer, 2013), presumably using the same materials as in the earlier study.

In describing the stop bursts used to create these VOT continua, Nittrouer (1999) noted that "the spectra of these noises [bursts] did not differ greatly: the /t/ burst simply had a bit more high-frequency energy than the /d/ burst." This impressionistic observation is consistent with the production studies reviewed above, and it is quite plausible that the frequency difference noted by Nittrouer accounts for the reported

shifts in categorization. However, phonetic analysis of the burst spectra was not provided, making it unclear whether the bursts are typical of AE stops. Furthermore, other properties of the naturally produced bursts—such as the presence of modal voicing and burst amplitude—were not reported.

Perception experiments of the type conducted by Keating (1979) and Nittrouer (1999) have never, as far as we know, used continua for non-coronal places of articulation. This is a notable gap in the literature, as findings from the production studies reviewed above suggest that the perceptual effect would be if anything stronger for the labial /p/ vs /b/ contrast, for which the difference in spectral means is numerically larger than in the coronals, and possibly absent for dorsal /k/ vs /g/.

Many other studies of stop voicing perception have been unable to address the role of the burst spectrum due to the method of stimulus construction. It is common to omit the stop burst altogether when creating VOT continua. Even when the burst is included, spectral properties are often confounded with the VOT manipulation; in the widely used method of Ganong (1980), continua are created by splicing progressively longer portions of a naturally produced voiceless stop, beginning at the point of release, into the recording of a syllable beginning with a voiced stop, so that all non-endpoint members of the continuum contain bursts transients from voiceless stops (e.g., Andruski *et al.*, 1994; McMurray *et al.*, 2008). The perceptual experiments reported here adopt the alternative method of stimulus creation pioneered by Keating (1979) and Nittrouer (1999), in which initial burst spectrum is manipulated independently of VOT.

C. Current study

The present study sought, first, to establish that the shape of the initial burst spectrum is significantly different for AE word-initial voiceless and voiced stops. This difference was investigated for all three places of articulation and in a way that attempted to isolate it from other phonetic correlates of the voicing contrast. Speech was collected from a much larger number of participants than in previous production studies. Furthermore, to investigate whether the difference is found outside of isolated word production, analyses were conducted on word-initial stops in a large read corpus (TIMIT; Garofolo *et al.*, 1993). On the basis of the production findings, a laboratory perception experiment evaluated the perceptual relevance of the burst spectrum by collecting identification responses and goodness ratings for cross-spliced VOT continua at the anterior places (labial and coronal). This experiment was replicated with a web-based crowdsourcing service (Amazon.com's Mechanical Turk; MTurk) in order to determine whether the burst spectrum serves as a secondary voicing cue under less controlled listening conditions and for a more diverse population of participants.

II. PRODUCTION STUDY

In this production study, the burst spectrum of word-initial stops was quantified by calculating spectral moments (Forrest *et al.*, 1988) on the average of several short spectral slices (the *smoothed spectrum*; Hanson and Stevens, 2003). The primary questions were whether one or more of the spectral moments differ significantly for voiceless and voiced stops, and whether a similar pattern of spectral difference holds at all three places of articulation. Statistical analyses were performed both on the entire set of stop productions and on the large subset in which modal phonation was not evident immediately before or after the point of stop release. Several alternative measures of spectral shape were also considered.

A. Methods

1. Participants

Eighteen Johns Hopkins University undergraduate students (14 female) participated in this experiment. All participants were native speakers of English, and none reported any speech or hearing impairments. Four additional participants completed the experiment, but were excluded from analysis because of missing or poor-quality recordings.

2. Procedure

Each participant produced consonant-vowel-consonant (CVC) syllables composed of the six English stop consonants /p,b,t,d,k,g/ crossed with the ten vowels /i,ɪ,e,ɛ,æ,ʌ,ɑ,ɔ,o,u/ and the final consonant, /t/. One CVC combination was excluded from the stimulus set due to its sensitive nature. As part of the same experiment, participants also produced CLVC syllables (L = /l/) with initial labial and dorsal consonants, the same vowels, and final /t/; these items were recorded for a different study and are not analyzed here. All syllables were produced within the carrier phrase "Say ___ again."

Participants were recorded in a sound attenuated booth with a Shure SM58 microphone (Niles, IL) and Zoom H4n digital recorder (Tokyo) at a sampling frequency of 48 kHz (16 bit). They were instructed to speak at a normal rate with a slight pause after "Say" and before "again." Stimuli were presented on a computer monitor using PsychoPy (Peirce, 2007). On each trial, a single CVC item was presented visually in the frame sentence; spellings were based on orthographic conventions of AE, with the orthographic form of each consonant and vowel held constant across items regardless of lexical status. Stimulus presentation and recording were self-paced. Each participant completed five blocks, with each block containing all of the CVC (and CLVC) syllables in a different random order and with short breaks allowed between blocks. This procedure resulted in four to five usable recordings of each syllable per participant, for a total of 5047 recorded tokens (ranging from 235 to 295 per participant).

3. Acoustic analyses

The onset and offset of each burst were marked with boundaries in Praat (Boersma and Weenink, 2013). Burst onset was identified from visual inspection of the waveform and wideband spectrogram as a sudden rise in sound energy relative to the preceding stop closure. Burst offset was identified with the onset of the following vowel, as indicated by the f0 track or a periodic waveform (whichever came first).

Prior to analysis, and for consistency with previous studies of stop acoustics (e.g., Forrest *et al.*, 1988; Sundara, 2005), recordings were resampled at 16 kHz, pre-emphasized above 1000 Hz, and high-pass filtered at 200 Hz to reduce the influence of low-frequency glottal vibration. For each burst, a smoothed spectrum was calculated by averaging the squared amplitude values of seven 64-point FFT spectra taken from 3 ms Hamming windows. The first window was centered on the point of release, and the centers of subsequent windows were advanced in 1 ms steps. If an entire burst (transient and following frication/aspiration) was shorter than the time required to accommodate all seven windows, without including spectral contributions from the following vowel, the number of slices was reduced to the largest number that would fit within the burst. The same procedure was applied to the bursts of voiceless and voiced stops; only 72 tokens (1.4%), all of which were instances of /b/, had bursts too short to accommodate 7 windows.

As our primary measure, we calculated the spectral mean or center of gravity (COG, in Hz) of the smoothed burst. COG is defined as the power-weighted average frequency of a spectrum. This measure and the three higher spectral moments (spectral standard deviation, skewness, kurtosis) were calculated according to the equations in Forrest *et al.* (1988). We also considered several alternative ways of quantifying the energy concentration in the burst. Peak frequency in Hz (Zue, 1976) was measured from the same smoothed spectra used to calculate spectral moments; peak frequency is a coarser, but perhaps perceptually effective, analog of COG. As an alternative to our time-averaged smoothed spectra, we extracted multitaper spectra (Shadle, 2012) from the initial 10 ms of the burst and calculated the COG in Hz. COG was additionally quantified in equivalent rectangular bandwidth (ERB) rate units from the output of an auditorily motivated filter bank (Patterson *et al.*, 1992).

B. Results

Separate Bayesian mixed-effects regression models, as implemented in R (R Development Core Team, 2008) by the MCMCglmm package (Hadfield, 2010), were fit to the acoustic measures. The models included phonological voice (voiceless and voiced), place of articulation (labial, dorsal, and coronal), and talker gender as fully crossed fixed factors; the following vowel category was included as a separate covariate. All fixed factors were effects coded, so that coefficient values indicate deviations from the grand mean. The model also included maximal random-effect structures for participant and item, and the Bayesian priors on fixed coefficients and random covariance matrices were set to their default values (see Hadfield, 2010). The output of MCMCglmm provides average coefficient estimates and associated p-values (calculated from 95% highest posterior density intervals around the coefficients).

1. Spectral measures

a. Smoothed spectra. Figure 1 displays mean smoothed spectra for the six stop consonants. The analysis of COG (Hz) computed from the smoothed spectra revealed significant main effects of all fixed factors. Most importantly, for our purposes, voiceless consonants had a higher concentration of energy than voiced consonants (voice = 129.83 [58.26, 205.73], $p < 0.01$).[2] In agreement with Forrest *et al.* (1988) and other previous work, COG was lower for labial and dorsal stops (labial = −652.56 [−880.87, −404.46], $p < 0.001$; dorsal = −279.58 [−443.73, −79.75], $p < 0.001$), and female talkers had higher COGs than male talkers (gender = 97.27 [15.93, 174.41], $p < 0.05$). Table I provides means and standard deviations for COG, along with values for the higher spectral moments, and peak burst frequency as computed from the same smoothed spectra.

The following vowel also significantly modulated burst COG with the strongest increase before the high front vowel /i/ (441.87 [239.67, 639.47], $p < 0.001$) and the strongest decrease for the mid back vowel /ɔ/ (−440.40 [−629.64, −270.64], $p < 0.001$). A significant interaction between voice and place indicated that the spectral difference between voiceless and voiced stops, in general, is enhanced for labial stops (voice × labial = 113.65 [11.98, 211.49], $p < 0.05$), but essentially nullified for dorsal stops (voice × dorsal = −178.20 [−260.04, −87.15], $p < 0.001$). *Post hoc* analyses of the data subset for each place established a

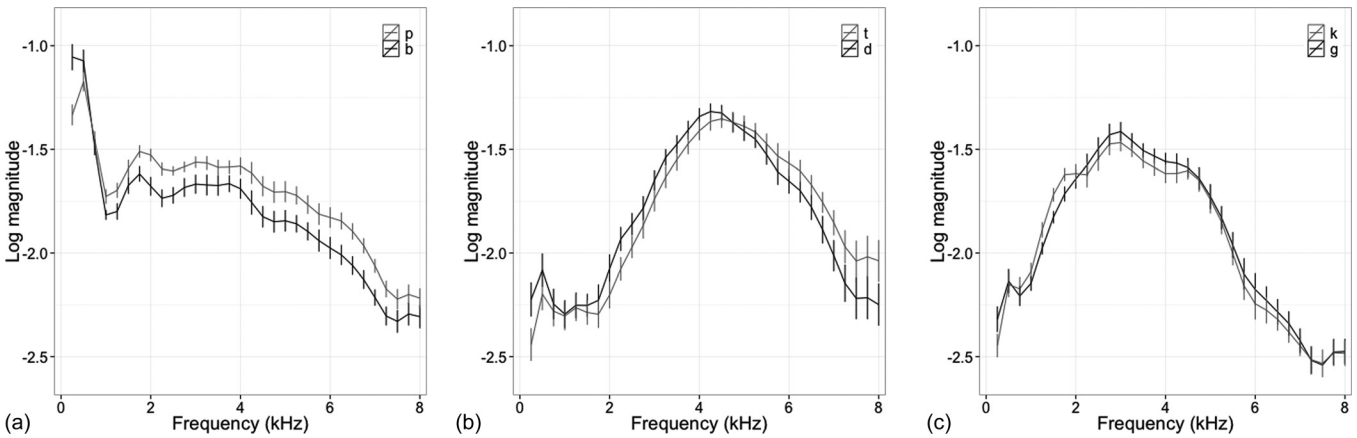

FIG. 1. Mean smoothed spectra for word-initial (a) labial, (b) coronal, and (c) dorsal stops in the laboratory production experiment. Error bars indicate ±1 standard error of the mean. Spectra for each stop were averaged first within participant and then across participants, on a linear (normalized power) scale and finally converted to a logarithmic scale for display.

TABLE I. Spectral moments and spectral peak for word-initial stop consonants in the laboratory study. COG, spectral standard deviation, and peak are in Hz; other spectral moments are unitless.

		COG		Standard deviation (sd)		Skewness		Kurtosis		Peak	
		Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd
Labial	/p/	3318	365	2096	140	0.27	0.23	−1.46	0.47	2236	689
	/b/	2833	464	2079	149	0.40	0.23	−1.29	0.50	1643	719
Coronal	/t/	4967	495	1468	250	−0.14	0.12	−2.33	0.32	5216	757
	/d/	4664	539	1494	233	−0.12	0.11	−2.33	0.30	4878	744
Dorsal	/k/	3450	244	1499	120	0.22	0.10	−2.30	0.24	2868	290
	/g/	3521	265	1511	142	0.19	0.10	−2.29	0.23	3021	353

significant effect of voice on COG for labial stops (voice = 240.39 [170.54, 321.03], $p < 0.001$) and coronal stops (voice = 195.56 [50.54, 316.94], $p < 0.01$); the effect was smaller in magnitude and reversed for dorsal stops (voice = −45.42 [−88.09, −2.40], $p < 0.05$).[3]

To ensure that the voice effect was not solely attributable to modal phonation, we repeated the preceding analysis excluding any tokens with visible glottal pulsing within 10 ms before or after the stop release. Approximately 10% of the dataset was excluded (22% of /b/ tokens, 8% of /d/, 14% of /g/, and 4% of all voiceless stops). The pattern of significance did not change in this analysis; crucially, the effect of phonological voice remained (voice = 105.11 [25.68, 175.20], $p < 0.01$), and the *post hoc* comparisons for each place showed the same pattern.

Spectral peak frequency (Hz), calculated from the same smoothed spectra as above, was highly correlated with COG ($r = 0.85$, $p < 0.001$). Reflecting the similarity between these two measures, the pattern of significant effects for spectral peak was the same as that for COG and, in particular, the effect of voice was present (voice = 140.45 [19.53, 252.56], $p < 0.05$).

b. Multitaper and ERB measures. COG was also calculated from multitaper spectra of the initial 10 ms burst (eight discrete prolate spheroidal sequences), which have been argued to provide a more accurate depiction of aperiodic sounds than traditional periodograms (Shadle, 2012). COG values from the multitaper spectra were nearly perfectly correlated with COGs extracted from the temporally smoothed spectra ($r = 0.97$, $p < 0.001$; see also Reidy and Beckman, 2012) and the pattern of significant effects, including voice, was the same (voice = 143.27 [67.77, 216.91], $p < 0.001$). Finally, to gain a more auditorily plausible model of spectral shape, COGs in ERB-rate were calculated from a gammatone filterbank (Patterson *et al.*, 1992). These values also correlated strongly with those from the smoothed spectra ($r = 0.86$, $p < 0.001$) and, as before, the model showed the same pattern of significance (crucially, voice = 0.19 [0.07, 0.30], $p < 0.01$). Taken together, the results from these alternative measures of spectral energy concentration establish a robust effect of voice on stop burst spectra.

2. Extension to connected speech

To determine whether this effect holds in continuous speech, the same acoustic analyses were performed on the TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo *et al.*, 1993). TIMIT contains samples of read speech from 630 talkers (192 female) of a variety of AE dialects. Word-initial, prevocalic stop bursts were extracted from the training and test subsets of TIMIT using the supplied phonetic transcriptions. All such bursts were included in the analyses regardless of the syllable structure, stress, prosodic position, and other properties of the containing words.[4] Details of the statistical analysis were the same as above, and the data is summarized in Table II.

The effect of voice on COG was in the same direction (and even larger in magnitude) as found for laboratory speech (voice = 325.53 [276.83, 375.24], $p < 0.001$). The main effects of place of articulation, gender, and vowel category—and more importantly the interaction between voice and place—also replicate the findings from our laboratory study (voice × labial = 165.56 [104.84, 225.83], $p < 0.001$; voice × dorsal = −223.69 [−286.69, −153.60], $p < 0.001$). The voice effect remained when, as before, tokens with

TABLE II. Spectral moments and spectral peak for word-initial, prevocalic stop consonants in the TIMIT corpus. COG, spectral standard deviation, and peak are in Hz; other spectral moments are unitless.

		COG		Standard deviation (sd)		Skewness		Kurtosis		Peak	
		Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd
Labial	/p/	3704	870	1999	275	0.12	0.33	−1.72	0.53	3208	2055
	/b/	2672	1022	1844	398	0.33	0.34	−1.60	0.68	1772	1538
Coronal	/t/	4550	842	1560	367	−0.09	0.22	−2.31	0.52	4742	1420
	/d/	3743	1133	1569	464	0.02	0.30	−2.17	0.74	3810	1695
Dorsal	/k/	3157	871	1493	382	0.23	0.27	−2.10	0.56	2710	1376
	/g/	2941	935	1539	425	0.24	0.32	−2.06	0.73	2400	1460

visible glottal pulsing within 10 ms of the burst were excluded (44% of /b/ tokens, 26% of /d/, 21% of /g/, and 3% of voiceless stops; voice = 200.97 [155.16, 250.65], $p < 0.001$). *Post hoc* analyses for each place separately showed that COG was higher for all voiceless stops in comparison to their voiced counterparts (labial: voice = 481.45 [415.79, 546.79], $p < 0.001$; coronal: voice = 397.96 [299.25, 491.26], $p < 0.001$; dorsal: voice = 94.35 [29.12, 165.37], $p < 0.05$). Note that the effect of voice on COG for dorsals is in the opposite direction of that found in laboratory speech, indicating that the previous results may be accidental or perhaps that dorsal bursts are more variable across talkers and dialects.

C. Discussion

Replicating past studies, the voicing of word-initial stops is reflected in the initial burst spectrum. This finding substantiates and extends upon observations going back to Halle *et al.* (1957) that voiceless stop consonant bursts have more energy in higher frequencies (see also Nittrouer, 1999; Sundara, 2005). This difference is limited to labial and coronal stops, with dorsal stops showing weak and inconsistent effects (see also Zue, 1976; Parikh and Loizou, 2005; Vicenik, 2010). The difference is robust to the precise measurement of spectral central tendency, not reducible to (visible) modal phonation, and is to a limited extent also found in higher spectral moments (in particular, skewness).

The finding that voice is reflected in spectral measures is of interest given that such measures have previously been studied with respect to place of articulation. We also found robust effects of place on COG and other moments that agree with prior literature. Importantly, the effect of voice should be understood as "nested" within that of place: the difference between voiceless and voiced stops holds only within (certain) places of articulation, not across places. As discussed further in Sec. IV, this implies that the interpretation of COG as a voice cue must be place dependent.

A potential concern is that the effects observed above are due to inclusion of more aspiration in the measurement window for voiceless stops. If this were correct, the spectral difference would simply reduce to the well-known fact that only voiceless AE stops have significant aspiration. We attempted to mitigate this possibility by measuring COG in short windows targeted at the transient and frication at burst onset. More importantly, aspiration would not uniformly elevate COG in the way required to produce our results; in particular, the energy concentration of aspiration in our recordings tends to be lower than that of the initial bursts of voiceless and voiced coronal stops (3000–4000 Hz vs 4000–5000 Hz). Therefore, the presence of aspiration in our measurement window could only weaken the effect of voice on COG for the coronal place. It is possible that degree of aspiration and concentration of burst energy have a common articulatory origin, as we discuss in Sec. IV, but acoustically these properties are not identical.

III. PERCEPTION STUDY

The production study above demonstrated a significant effect of voice on burst shape for anterior stops, but did not establish whether this difference is perceptually effective. The present experiment aimed to determine whether listeners' voiced–voiceless category boundaries are shifted as a result of a burst manipulation that was crossed with a standard VOT continuum.

A. Methods

1. Participants

Sixteen Johns Hopkins University undergraduate students (11 female) participated in this experiment. The participants were native speakers of English and did not report any speech or hearing impairments.

2. Materials

a. Continuum creation. Two VOT continua were created for each of the anterior places of articulation (labial and coronal) by cross-splicing tokens of natural speech (Keating, 1979; Ganong, 1980; Andruski *et al.*, 1994; McMurray *et al.*, 2008). A minimal pair of tokens was selected from the production study of Sec. II in the manner described below. For each token, the initial 10 ms of the burst was excised; the intensity of the voiced burst was then rescaled to match that of the voiceless burst. (Original intensities were 47 dB for /b/, 55 dB for /p/, 56 dB for /d/, and 57 dB for /t/.) To create the body of the continua for a given place, the rhyme portion of the voiced token was cross-spliced with aspiration and frication taken from the voiceless token. This was performed in 7 ms increments for a total of seven steps. All segments extracted during the process were taken at zero-crossings to avoid any discontinuities in the waveform. For each step, the vowel was reduced by one pitch period while aspiration was lengthened by 7 ms (Keating, 1979). The initial voiceless and voiced bursts were then appended to each step of the continuum resulting in a pair of continua in which burst was fully crossed with VOT.

b. Token selection. The tokens selected were /bæt/ and /pæt/ ("ba"'-"pat") for the labial continua and /dat/ and /tat/ ("dot"-"tot") for the coronal continua; the talker was a single male speaker whose COG for each place of articulation closely matched the means of Table I. Tokens were selected from the first quartile of the talker's voiced productions for each place of articulation and from the fourth quartile of voiceless productions (labial: /b/ COG = 1513 Hz, /p/ COG = 3494 Hz; coronal: /d/ COG = 3601 Hz, /t/ COG = 5424 Hz). Care was taken to avoid extreme or unnatural COG values. Additionally, the voiced tokens did not have visible glottal pulsing that continued into the burst.

3. Procedure

All testing took place in a quiet room with stimuli presented over Sony MDR-V150 headphones (Tokyo). Place of articulation was counterbalanced within participants such that half the participants received the labial stimuli first. For each place, there were 8 blocks and all 14 stimuli (2 bursts × 7 VOT steps) were randomly presented in each block. Within a trial, the stimulus was presented first for

identification of the initial consonant as "P" or "B" (or "T" or "D"), with the order of response options counterbalanced across participants. On the second presentation, participants were asked to provide a goodness rating for the initial consonant as an instance of their previous identification response. The goodness scale ranged from 1 to 7 with 1 as a "poor" instance of the speech sound, 4 as "ok," and 7 as an "excellent" instance of the speech sound. Three blocks of practice trials preceded each part of the experiment. The four practice stimuli were the natural voiceless and voiced tokens (with aspiration of voiceless stops reduced to 42 ms) and the most extreme members of the critical stimuli (i.e., highest VOT and high COG vs lowest VOT and low COG).

B. Results

1. Categorization responses

Results for both the labial and coronal categorization experiments were analyzed using an MCMC generalized linear mixed-effects model. The two crossed fixed factors were burst type (effects coded as $+1$ = high COG, -1 = low COG) and VOT (coded as a scaled numeric predictor). Random intercepts and slopes were included for participants; note that a continuum member is fully identified by its burst and VOT values so that random item effects cannot be investigated in this study. Burst type was found to have a significant effect in the expected direction (see Fig. 2) with the higher COG burst shifting the category boundary in favor of voiceless responses for both the labial stops (burst = 0.63 [0.42, 0.79], $p < 0.001$) and coronal stops (burst = 0.81 [0.58, 1.11], $p < 0.001$). As expected, VOT also made a significant contribution to identification responses (labial: vot = 4.56 [4.22, 4.85], $p < 0.001$; coronal: vot = 5.07 [4.44, 5.70], $p < 0.001$). No interaction between VOT and burst type was observed for coronals, but there was a significant interaction for labials, possibly indicating a bias in favor of /b/ responses (vot × burst = −0.23 [−0.38, −0.07], $p < 0.01$).

2. Goodness ratings

Goodness ratings were analyzed for each response option separately with VOT and burst as crossed fixed effects and random intercepts and slopes for participants. For stops categorized as "P," higher VOT and higher COG both had positive effects on goodness ratings (vot = 0.56 [0.44, 0.67], $p < 0.001$; burst = 0.15 [0.07, 0.23], $p < 0.01$). Likewise, lower VOT and lower COG significantly increased goodness ratings for stops categorized as "B" (vot = −1.53 [−1.83, −1.22], $p < 0.001$; burst = −0.35 [−0.63, −0.08], $p < 0.05$). The interaction between VOT and burst was not significant for either of these response options.

Goodness ratings were also significantly influenced by VOT and burst type for stops categorized as "T" (vot = 0.86 [0.70, 1.01, $p < 0.001$; burst = 0.23 [0.09, 0.36], $p < 0.001$). Additionally, there was a significant interaction between VOT and burst type, suggesting especially high ratings for stimuli that combine long VOT with high COG (vot × burst = 0.15 [0.04, 0.27], $p < 0.05$). However, for stops categorized as "D," only VOT significantly influenced goodness ratings; there was no significant effect of the burst (vot = −0.78 [−0.92, −0.65], $p < 0.001$; burst = 0.07 [−0.01, 0.18], $p = 0.16$) and no interaction of VOT and burst type.

3. Web-based replication

A perception experiment with the same materials and procedure was also conducted online with Amazon.com's Mechanical Turk (MTurk), a crowdsourcing service that has been increasingly used in psycholinguistic and phonetic research (Cooke *et al.*, 2011). To reduce the length of the experiment, each participant received stimuli for only 1 place of articulation; 16 participants (9 female, ages 20–56, median: 36) were assigned to the labial stimuli, and 16 participants (7 female, ages 20–53, median: 34.5) to the coronal stimuli.[5] All participants reported English as their native language. For the online presentation, there were two practice blocks and eight target blocks with order of response options counterbalanced across participants.

Replicating the laboratory results, burst type and VOT significantly influenced identifications (labial: burst = 0.50 [0.40, 0.58], $p < 0.001$, vot = 3.76 [3.25, 4.24], $p < 0.001$; coronal: burst = 0.52 [0.35, 0.74], $p < 0.001$, vot = 3.94 [3.48, 4.51], $p < 0.001$) with no significant interaction (see Fig. 3). Proportion voiceless responses from this experiment were nearly perfectly correlated with those of the laboratory

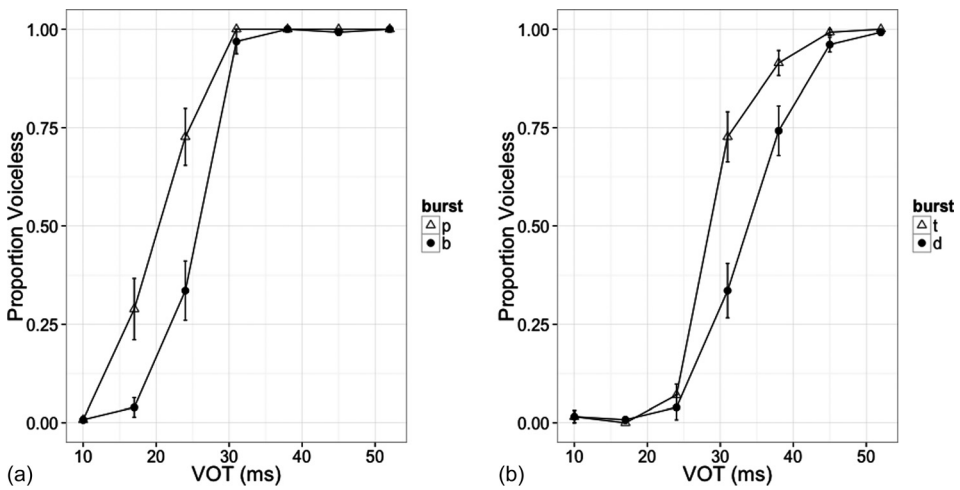

FIG. 2. (a) Categorization results in the laboratory experiment for VOT continua with a high (/p/) and low (/b/) burst. Error bars indicate ±1 standard error of the mean. (b) Categorization results in the laboratory experiment for VOT continua with a high (/t/) and low (/d/) burst. Error bars indicate ±1 standard error of the mean.

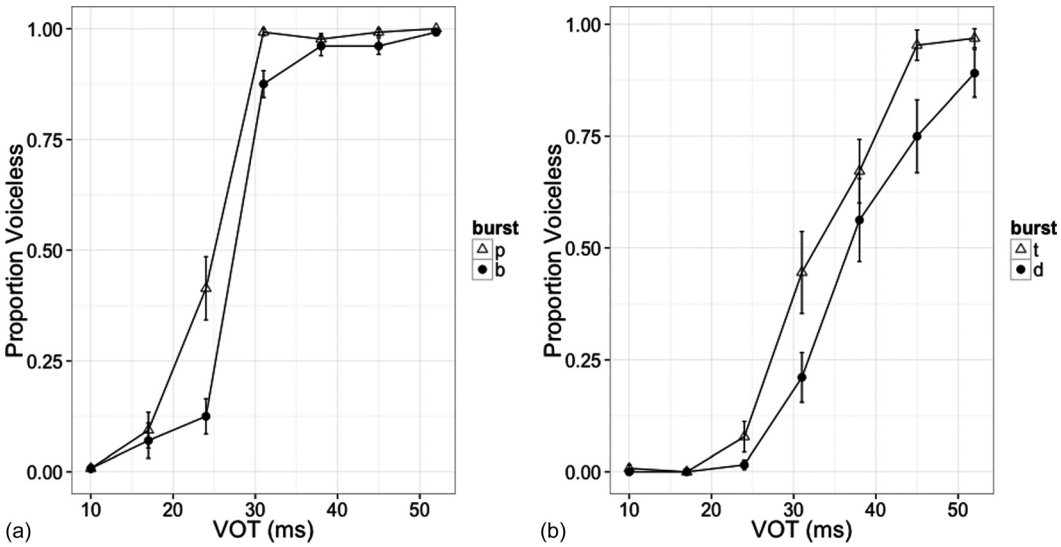

FIG. 3. (a) Categorization results in the MTurk experiment for VOT continua with a high (/p/) and low (/b/) burst. Error bars indicate ±1 standard error of the mean. (b) Categorization results in the MTurk experiment for VOT continua with a high (/t/) and low (/d/) burst. Error bars indicate ±1 standard error of the mean.

study (labial continua: $r = 0.97$, $p < 0.001$; coronal continua: $r = 0.98$, $p < 0.001$).

The pattern of goodness ratings was similar to that found in the laboratory experiment, with burst and VOT affecting the goodness of stops categorized as voiceless ("P": burst = 0.17 [0.02, 0.30], $p < 0.05$, vot = 0.39 [0.28, 0.49], $p < 0.001$; "T": burst = 0.22 [0.02, 0.39], $p < 0.05$, vot = 0.74 [0.58, 0.96], $p < 0.001$). Only VOT had a significant effect on ratings for stops categorized as voiced ("B": burst = −0.09 [−0.27, 0.07], $p = 0.33$, vot = −1.18 [−1.52, −0.91], $p < 0.001$; "D": burst = −0.04 [−0.16, 0.07], $p = 0.44$, vot = −0.77 [−0.87, −0.65], $p < 0.001$).

C. Discussion

These two experiments demonstrate the perceptual significance of the burst spectrum for identifying voicing in labial and coronal stop consonants. Previous studies have observed a trading relation between VOT and the burst of coronal stops only; however, these studies did not control for burst amplitude or the presence of modal phonation (Keating, 1979; Nittrouer, 1999; Caldwell and Nittrouer, 2013) and burst shape was not quantified. The present study controls for possible confounds of amplitude and modal voicing and additionally specifies the difference in spectral shape with COG. Furthermore, the effect of COG on perception of stop consonant voicing is confirmed not only for coronals, as in previous studies, but also for labials.

The results from goodness ratings provide evidence that stops identified as voiceless labials and coronals are better members of their respective phonetic categories when COG is higher. However, the effect of COG was not as straightforward for stops categorized as voiced; only labials received significantly higher ratings when COG was lower, and even this effect was not found in the MTurk study. The asymmetry in goodness ratings may be attributable to differences in VOT typicality. While the low VOT values in the experiment are typical of voiced stops, the higher VOT values are still somewhat low for voiceless stops. It may be that burst spectrum has a strong impact on goodness only when VOT is non-prototypical for a phonetic category.

IV. GENERAL DISCUSSION

Stop bursts are rich sources of information about phonetic identity. We have shown that the spectrum at the beginning of the burst, in addition to providing information about place of articulation, also reflects voicing. Voiceless stops have higher spectral means than voiced stops, a difference that is particularly strong for labials and also significant for coronal stops, but not consistently found for dorsals. This effect is not limited to laboratory speech, but also established in a widely used corpus of read speech. Additionally, it is robust to the precise quantification of the burst spectrum mean or peak (e.g., with temporal vs multitaper spectral smoothing, in Hz and ERB-rate units). In voice perception, a cue-trading relationship emerges when burst spectrum is manipulated independently of VOT. Furthermore, the burst spectrum contributes to phonetic category goodness ratings, at least for voiceless stops with relatively low (but unambiguous) VOT values.

Previous research on the burst-spectrum cue to stop voicing has not clearly distinguished differences in spectral shape from amplitude, modal phonation, and other phonetic properties and has been limited to the coronal place of articulation, whereas we have found the largest effect for labials. Our results are similar to those obtained for AE non-sibilant fricatives by Jongman *et al.* (2000), suggesting that a higher concentration of spectral energy is characteristic of all anterior voiceless sounds other than /s/ and /ʃ/. Parallel differences have also been found in other varieties of English, in Georgian (which also has a short- vs long-lag stop voicing contrast), and in Dutch (which opposes modally voiced vs voiceless unaspirated stops). It remains to be determined

whether a difference in burst energy concentration holds for stop voice contrasts cross-linguistically.

We have focused on the acoustic–phonetic and perceptual aspects of this effect, and its precise articulatory origin remains to be identified. One potential source may arise from the difference in intraoral pressure between voiceless and voiced stops. Voiceless stops are characterized by greater intraoral pressure than voiced stops (Malécot, 1966; Lubker and Parris, 1970; Rodgers and Fuchs, 2010; but cf. Lisker, 1970, on utterance-initial AE stops). This pressure difference, if it results in higher volume velocity during the burst, would give rise to greater energy at higher frequencies (Zue, 1976).

In order for this account to explain the finding of a strong effect for labials, but not dorsals, it may be crucial to consider the time of peak airflow after the release. Peak airflow for labial stops is estimated to occur within the first 10 ms, whereas airflow for dorsal stops increases slowly until it reaches a peak later on in the burst (Stevens, 1998). Thus, the difference in peak airflow between the voiceless and voiced stops at burst onset should be greater for labials than for dorsals. Assuming that the timing of the airflow peak is related to rate of release, coronal stops should fall between labials and dorsals, in agreement with our results. Alternatively, it is possible that the spectral difference between voiceless and voiced coronals reflects a difference in the length of the cavity anterior to the place constriction (see van Alphen and Smits, 2004, on Dutch); however, it is not clear how such an account could extend to the labial stops.

Our results are relevant for the general theory of how phonetic contrasts are encoded in the acoustic signal, theories of cue weighting, and the timecourse of phonetic categorization. Because differences in spectral shape due to place are much larger than those attributable to voice, burst spectrum is interpretable as a voice cue only subsequent to (or simultaneously with) the perception of place. The perceptual interaction between voice and place has been noted in a variety of studies, including Lisker and Abramson (1970), Kuhl and Miller (1975), Miller (1977), and, more recently, Benkí (2001). More broadly, the present findings provide further support for a many-to-many mapping between acoustic properties and phonetic features (e.g., Nearey, 1997).

Having empirically established an additional cue to stop voicing, we look forward to future research on how this cue is weighted relative to others. It is possible that burst spectrum will receive a weight that is determined entirely by its reliability (cf. Toscano and McMurray, 2010). However, other considerations, such as presence at a landmark (Stevens, 2002), robustness to noise, and temporal order, may also impact cue integration. Given that the initial burst spectrum is one of the earliest cues to voicing in word-initial position, it may receive a weight that is greater than its reliability would warrant. Relatedly, the burst spectrum may be used very early in the incremental identification of stop voice contrasts (Allopenna *et al.*, 1998). Furthermore, the burst spectrum provides a clear case in which the weight of a cue to a phonetic distinction clearly differs across contrasts (labial, coronal vs dorsal).[6]

V. CONCLUSION

This study verified that the initial burst spectrum of word-initial labial and coronals stops is a secondary cue to voicing. Various measures of spectral central tendency confirmed a greater concentration of energy in the higher frequencies for voiceless anterior stops in comparison to their voiced counterparts. This difference was established for word-initial, prevocalic stop consonants in both a laboratory production study and the TIMIT corpus. No consistent effect of voicing on spectral shape was found for dorsal stops. This difference in spectral shape was also found to be perceptually relevant in voicing identification and goodness rating experiments; a higher COG burst required less VOT for a voiceless response and received higher goodness ratings when identified as voiceless. These findings were replicated in a web-based experiment, suggesting that online methods are reliable for investigating speech perception.

While this study established an acoustic and perceptual relationship between initial burst spectrum and stop voicing, further research is necessary to identify the articulatory source of this relationship. Additionally, the perception experiments employed bursts with naturally-produced, albeit exaggerated, spectral values; whether finer differences in the burst spectrum would be reflected in listeners' voicing perceptions remains open to further investigation. Implications for cue weighting were discussed, but a computational model that captures the trading relation between VOT and burst spectrum must await future research. Finally, some previous research has suggested that spectral properties of the burst covary with voicing distinctions in other languages, but no systematic typological survey has been conducted.

ACKNOWLEDGMENTS

The authors would like to thank Anthony Arnette and Samhita Ilango for their assistance in data collection and processing. We would also like to thank Mary Beckman, Jack Godfrey, and, especially, Edward Flemming for input on acoustic analysis. Portions of this work were presented at the 2014 Linguistic Society of America meeting and to the Johns Hopkins University Cognitive Science Department, and we thank those audiences for their insightful questions and comments.

[1]We use the traditional terms *voiceless* and *voiced* to describe stops even when modal phonation during the closure is not typical of their realization, as is the case in our production study. The contrast studied here is most often *voiceless aspirated* (spread glottis) vs *voiceless unaspirated*, and has also been characterized as *tense* vs *lax* (e.g., Halle *et al.*, 1957).

[2]The results of statistical analyses are reported as means of regression coefficients, followed by 95% highest posterior density intervals in square brackets and associated *p*-values. Binary fixed factors were effects coded (voicing: +1 voiceless, −1 voiced; gender: +1 female, −1 male). Fixed factors with multiple levels were coded as effects relative to the reference level (e.g., coronal was the reference level for place of articulation).

[3]The analyses of the remaining three spectral moments (standard deviation, skewness, and kurtosis) showed significant main effects of place (and, in some cases, gender and vowel as well), but not of voice (variance: $p = 0.65$; skewness: $p = 0.09$; kurtosis: $p = 0.17$). As indexed by standard deviation, the distribution of energy was more diffuse in the spectra of labial stops (labial = 412.14 [319.78, 496.33], $p < 0.001$) and more

concentrated in those of dorsal stops (dorsal = −176.12 [−242.35, −101.05], $p < 0.001$). The spectra of both labial and dorsal stops had significantly more positive skew (labial = 0.17 [0.09, 0.25], $p < 0.001$; dorsal = 0.07 [0.02, 0.12], $p < 0.05$). There was also a significant interaction between voice and place in the analysis of skewness; voiceless labials exhibited more negative skew than voiced labials (−0.04 [−0.07, −0.01], $p < 0.05$)—another reflection of the greater concentration of energy in higher frequencies for the voiceless stops. Finally, labial stops indicated a significantly higher kurtosis (0.63 [0.43, 0.81], $p < 0.001$), whereas dorsals exhibited a significantly lower kurtosis (−0.27 [−0.37, −0.16], $p < 0.001$), relative to the average kurtosis across stops; there was no interaction with voice for this spectral moment.

[4]A few words were excluded from the analysis due to their disproportionately high frequency within the corpus ("to," "too," "do," "carry," and "dark").

[5]Participants reported the audio devices they used to listen to the stimuli. For the labial conditions, participants used headphones (9), earbuds (3), external speakers (3), and internal speakers (1). For the coronal conditions, the devices were headphones (8), earbuds (1), external speakers (4), and internal speakers (3).

[6]An additional perception study with dorsal stops, performed in the same manner as that of Sec. III, failed to find a significant effect of burst spectrum on /k/ vs /g/ identifications (vot = 4.01 [3.68, 4.33], $p < 0.001$; burst = 0.12 [−0.02, 0.23], $p = 0.11$). While listeners might have extrapolated from the voicing contrast in the labial and coronal stops, it appears that in this case the contrast-specific acoustic distribution takes priority over contrast-general patterning. It remains possible that a spectral cue to voicing could be present, later in dorsal stop bursts, at the point of peak airflow.

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (**1998**). "Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models," J. Mem. Lang. **38**(4), 419–439.

Andruski, J. E., Blumstein, S. E., and Burton, M. (**1994**). "The effect of subphonetic differences on lexical access," Cognition **52**(3), 163–187.

Benkí, J. (**2001**). "Place of articulation and first formant transition pattern both affect perception of voicing in English," J. Phon. **29**(1), 1–22.

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. (**2013**) "Praat: Doing phonetics by computer" (version 5.3.43) [computer program] (Last viewed September 18, 2014).

Caldwell, A., and Nittrouer, S. (**2013**). "Speech perception in noise by children with cochlear implants," J. Speech Lang. Hear. Res. **56**(1), 13–30.

Cooke, M., Barker, J., Lecumberri, M. L. G., and Wasilewski, K. (**2011**). "Crowdsourcing for word recognition in noise," in *Proceedings of INTERSPEECH*, ISCA, Florence, Italy, pp. 3049–3052.

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.

Ganong, W. F. (**1980**). "Phonetic categorization in auditory word perception," J. Exp. Psychol.: Hum. Percep. Perf. **6**(1), 110–125.

Garofolo, J., Lamel, L., Fisher, M., Fiscus, J., Pallett, D., and Dahlgren, N. (**1993**). DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus, Technical Report NISTIR4930 (National Institute of Standards and Technology, Gaithersburg, MD).

Hadfield, J. D. (**2010**). "MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package," J. Stat. Soft. **33**(2), 1–22.

Haggard, M., Ambler, S., and Callow, M. (**1970**). "Pitch as a voicing cue," J. Acoust. Soc. Am. **47**(2), 613–617.

Halle, M., Hughes, G. W., and Radley, J.-P. (**1957**). "Acoustic properties of stop consonants," J. Acoust. Soc. Am. **29**(1), 107–116.

Hanson, M., and Stevens, K. N. (**2003**). "Models of aspirated stops in English," in *Proceedings of the 15th ICPhS*, Barcelona, Spain, pp. 783-786.

Harrington, J. (**2010**). *Phonetic Analysis of Speech Corpora* (Wiley-Blackwell, Malden, MA), p. 192.

Jongman, A., Wayland, R., and Wong, S. (**2000**). "Acoustic characteristics of English fricatives," J. Acoust. Soc. Am. **108**(3), 1252–1263.

Keating, P. A. (**1979**). "A phonetic study of a voicing contrast in Polish," Ph.D. dissertation, Brown University, Providence, RI.

Kirkham, S. (**2011**). "The acoustics of coronal stops in British Asian English," in *Proceedings of the 17th ICPhS*, Hong Kong, China, pp. 1102–1105.

Kuhl, P. K., and Miller, J. D. (**1975**). "Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants," Science **190**(4209), 69–72.

Liberman, A. M., Delattre, P. C., and Cooper, F. S. (**1958**). "Some cues for the distinction between voiced and voiceless stops in initial position," Lang. Speech **1**(3), 153–167.

Liberman, A. M., Harris, K. S., Kinney, J. A., and Lane, H. (**1961**). "The discrimination of relative onset-time of the components of certain speech and nonspeech patterns," J. Exp. Psychol. **61**(5), 379–388.

Lisker, L. (**1970**). "Supraglottal air pressure in the production of English stops," Lang. Speech **13**(4), 215–230.

Lisker, L. (**1986**). "'Voicing' in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees," Lang. Speech **29**(1), 3–11.

Lisker, L., and Abramson, A. S. (**1964**). "A cross language study of voicing in initial stops: Acoustical measurements," Word **20**(3), 384–422.

Lisker, L., and Abramson, A. S. (**1970**). "The voicing dimension: Some experiments in comparative phonetics," in *Proceedings of the 6th ICPhS*, Prague, Czech Republic, pp. 563–567.

Lisker, L., Liberman, A. M., Erickson, D., Dechovitz, D., and Mandler, R. (**1977**). "On pushing the voice-onset-time (VOT) boundary about," Lang. Speech **20**(3), 209–216.

Lubker, J. F., and Parris, P. J. (**1970**). "Simultaneous measurements of intraoral pressure, force of labial contact, and labial electromyographic activity during production of the stop consonant cognates /p/ and /b/," J. Acoust. Soc. Am. **47**(2), 625–633.

Malécot, A. (**1966**). "The effectiveness of intra-oral air-pressure-pulse parameters in distinguishing between stop cognates," Phonetica **14**(2), 65–81.

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., and Subik, D. (**2008**). "Gradient sensitivity to within-category variation in words and syllables," J. Exp. Psychol.: Hum. Percep. Perf. **34**(6), 1609–1631.

Miller, J. L. (**1977**). "Nonindependence of feature processing in initial consonants," J. Speech Hear. Res. **20**(3), 519–528.

Miller, J. L. (**1994**). "On the internal structure of phonetic categories: A progress report," Cognition **50**(1), 271–285.

Nearey, T. M. (**1997**). "Speech perception as pattern recognition," J. Acoust. Soc. Am. **101**(6), 3241–3254.

Nittrouer, S. (**1999**). "Do temporal processing deficits cause phonological processing problems?," J. Speech Lang. Hear. Res. **42**(4), 925–942.

Oden, G. C., and Massaro, D. W. (**1978**). "Integration of featural information in speech perception," Psych. Rev. **85**(3), 172–191.

Parikh, G., and Loizou, P. C. (**2005**). "The influence of noise on vowel and consonant cues," J. Acoust. Soc. Am. **118**(6), 3874–3888.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (**1992**). "Complex sounds and auditory images," in *Auditory Physiology and Perception, Proceedings of the 9th International Symposium on Hearing*, edited by Y. Cazals, L. Demany, and K. Horner (Pergamon, Oxford, UK), Vol. 83, pp. 429–446.

Peirce, J. W. (**2007**). "Psychopy: Psychophysics software in Python," J. Neuro. Meth. **162**(1), 8–13.

R Development Core Team (**2013**). *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).

Reidy, P., and Beckman, M. (**2012**). "The effect of spectral estimator on common spectral measures for sibilant fricatives," in *Proceeedings of INTERSPEECH*, ISCA, Portland, Oregon, pp. 1516–1519.

Repp, B. H. (**1979**). "Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants," Lang. Speech **22**(2), 173–189.

Repp, B. H. (**1982**). "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception," Psych. Bull. **92**(1), 81–110.

Repp, B. H. (**1984**). "Categorical perception: Issues, methods, findings," in *Speech and Language: Advances in Basic Research and Practice*, edited by N. J. Lass (Academic, Orlando, FL), pp. 243–335.

Rodgers, B., and Fuchs, S. (**2010**). "How intraoral pressure shapes the voicing contrast in American English and German," Zentr. Allg. Sprach. Pap. Ling. **52**(2010), 63–82.

Shadle, C. H. (**2012**). "Acoustics and aerodynamics of fricatives," in *Handbook of Laboratory Phonology*, edited by A. Cohn, C. Fougeron, and M. Huffman (Oxford University Press, New York), pp. 511–526.

Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT Press, Cambridge, MA), p. 453.
Stevens, K. N. (**2002**). "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am. **111**(4), 1872–1891.
Summerfield, Q. (**1981**). "Articulatory rate and perceptual constancy in phonetic perception," J. Exp. Psychol.: Hum. Percep. Perf. **7**(5), 1074–1095.
Sundara, M. (**2005**). "Acoustic-phonetics of coronal stops: A cross-language study of Canadian English and Canadian French," J. Acoust. Soc. Am. **118**(2), 1026–1037.
Toscano, J. C., and McMurray, B. (**2010**). "Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics," Cog. Sci. **34**(3), 434–464.
van Alphen, P. M., and Smits, R. (**2004**). "Acoustical and perceptual analysis of the voicing distinction in Dutch initial plosives: The role of prevoicing," J. Phon. **32**(4), 455–491.
Vicenik, C. (**2010**). "An acoustic study of Georgian stop consonants," J. Int. Phon. Assoc. **40**(1), 59–92.
Zue, V. W. (**1976**). "Acoustic characteristics of stop consonants: A controlled study," Ph.D. dissertation, MIT, Cambridge, MA.