Journal of Phonetics 61 (2017) 30–47

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Research Article

Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English

Eleanor Chodroff*, Colin Wilson

Johns Hopkins University, Department of Cognitive Science, United States

ARTICLE INFO

Article history:
Received 28 June 2016
Received in revised form
1 January 2017
Accepted 4 January 2017

Keywords:
Stop consonants
Voice onset time
Talker variability
Phonetic covariation
Corpus phonetics

ABSTRACT

Variation across talkers in the acoustic-phonetic realization of speech sounds is a pervasive property of spoken language. The present study provides evidence that variation across talkers in the realization of American English stop consonants is highly structured. Positive voice onset time (VOT) was examined for all six word-initial stop categories in isolated productions of CVC syllables and in a multi-talker corpus of connected read speech. The mean VOT for each stop differed considerably across talkers, replicating previous findings, but importantly there were strong and statistically significant linear relations among the means (e.g., the mean VOTs of [pʰ] and [kʰ] were highly correlated across talkers, $r > 0.80$). The pattern of VOT covariation was not reducible to differences in speaking rate or other factors known to affect the realization of stop consonants. These findings support a uniformity constraint on the talker-specific realization of a phonetic property, such as glottal spreading, that is shared by multiple speech sounds. Because uniformity implies mutual predictability, the findings also shed light on listeners' ability to generalize knowledge of a novel talker from one stop consonant to another. More broadly, structured variation of the kind investigated here indicates a relatively low-dimensional encoding of talker-specific phonetic realization in both speech production and speech perception.

1. Introduction

The realization of a phonetic category can vary extensively across languages, phonetic contexts, and talkers. For example, the average voice onset time (VOT) of a voiceless aspirated stop such as [kʰ] is not the same in all languages. Cho and Ladefoged (1999) report a mean VOT of 154 ms for Nahavo [kʰ] (on the basis of recordings described in McDonough & Ladefoged, 1993), but other languages in their survey have far lower means for the same stop (e.g., 84 ms in Hupa) and lower values have been reported in several studies of American English (e.g., Lisker & Abramson, 1964: 80 ms; Klatt, 1975: 70 ms; Byrd, 1993: 52 ms). Even within a language, the average VOT values of [kʰ] and other aspirated stops can vary according to numerous linguistic and sociolinguistic factors (see Section 1.1 below). Differences in the phonetic realization of many other phonetic categories have been documented in the literature on language-specific phonetics (e.g., for vowels: Lindau & Wood, 1977; Disner, 1983; Chung et al., 2012; fricatives: Nartey, 1982; Gordon, Barthmaier, & Sands, 2002; and nasal consonants: Harnsberger, 2000), and these cross-linguistic differences are again paralleled by extensive variation across dialects and speakers of the same language (e.g., for vowels: Peterson & Barney, 1952; Labov, Yaeger, & Steiner, 1972; Jacewicz, Fox, & Salmons, 2007; Escudero, Boersma, Rauber, & Bion, 2009; and fricatives: Newman, Clouse, & Burnham, 2001; Munson, McDonald, DeBoe, & White, 2006).

While the realization of speech sounds is highly variable, it is also highly patterned or *structured*. A simple type of structure involves the means of multiple categories along a single acoustic-phonetic dimension. For example, the mean VOT of [pʰ] varies across languages to an extent similar to that of [kʰ], but the two means *do not vary independently*. In many (if not all) languages that have both [pʰ] and [kʰ], the value for the aspirated labial stop is lower than that of the aspirated velar stop (e.g., Fisher-Jorgensen, 1954; Peterson & Lehiste, 1960; Lisker & Abramson, 1964; Cho & Ladefoged, 1999). Moreover, previous studies have identified a tight positive (linear) correlation between the mean VOTs of [pʰ] and [kʰ] in laboratory productions by individual speakers of American English (e.g., Zlatin, 1974; Koenig, 2000; Newman, 2003; Theodore, Miller, & DeSteno, 2009). For other types of speech sounds, the assumption of linear covariation of acoustic-phonetic means across talkers is built into many models of talker adaptation or 'normalization' (e.g., for vowels: Nearey & Assmann, 2007; and fricatives: McMurray & Jongman, 2011).

* Correspondence to: Johns Hopkins University, Krieger Hall 237, 3400 N. Charles St., Baltimore, MD 21218, United States.
E-mail address: chodroff@cogsci.jhu.edu (E. Chodroff).

0095-4470/$ - see front matter © 2017 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2017.01.001

In this paper, we make several novel contributions to the study of linear relations among American English (AE) stops on the dimension of *positive VOT*. We examine all six stops in word-initial prevocalic position, investigating correlations among their VOT means across talkers both in isolated speech (single words produced in carrier phrases) and, most importantly, in a large corpus of connected read speech. In line with previous results we find high correlations among the means of aspirated stops ([pʰ tʰ kʰ]), and we further establish that weaker relations hold within the voiced stops ([b d g]) and between homorganic stop pairs (e.g., [kʰ g]). Regression analyses indicate that this pattern of covariation across talkers cannot be reduced to differences in overall speaking rate or to contextual factors known to affect positive VOT values (e.g., the quality of the following vowel).[1]

Patterns of covariation such as the one identified here have implications for the theory of phonetic realization as it applies to individual speakers, and for understanding perceptual adaptation on the part of listeners. In particular, VOT covariation among the aspirated stops can be straightforwardly accounted for with a constraint that requires the talker-specific realization of a phonetic property (e.g., glottal spreading) to be *uniform* across speech sounds. The uniformity constraint, which could extend to many other phonetic properties and sound classes, allows talkers to differ but imposes a common relational structure or pattern on their phonetic systems. Listeners could employ prior knowledge of this structure when adapting to a novel talker, as direct experience of the phonetic realization of one sound provides valuable information about how the same talker would realize other related sounds. From the most general perspective, patterns of covariation indicate that talker-specific phonetic systems — which specify means and other parameters on many dimensions for each category — can be accurately represented in a space of relatively low dimensionality.

In the following, we briefly summarize the major sources of variation that influence VOT. These other sources should be controlled, by experimental design or statistical analysis, in order to clearly observe how VOT means covary across talkers. We then review previous research on phonetic covariation, focusing on the studies most similar to our own. Finally, we provide an outline of the rest of the paper.

1.1. Sources of VOT variation

In American English, voiceless stops have systematically longer VOT than voiced stops in word-initial position (e.g., Lisker & Abramson, 1964). Differences in VOT means across place of articulation have been extensively documented in the literature for a variety of languages. For voiceless unaspirated stops, there is a general increase in VOT with more posterior places of articulation (Cho & Ladefoged, 1999). Recall that for voiceless aspirated stops, the VOT of [pʰ] is less than that of [kʰ] (e.g., Peterson & Lehiste, 1960; Klatt, 1975; Zue, 1976). Regarding the relative ranking of [tʰ], the findings are inconsistent: while a few studies report a mean VOT of [tʰ] medial to that of [pʰ] and [kʰ] (e.g., Peterson & Lehiste, 1960; Lisker & Abramson, 1964), many have also found minimal differences between [tʰ] and [kʰ] in both American and British English (Suomi, 1980; Docherty, 1992; Yao, 2009).

In addition to voice and place features, numerous contextual, prosodic, lexical, and global factors also contribute to VOT variability. Longer VOTs are observed before high and tense vowels, particularly [i], for voiceless stops (Klatt, 1975; Port & Rotunno, 1979; Weismer, 1979; Flege, Frieda, Walley, & Randazza, 1998; see also Nearey & Rochet, 1994 for Canadian English). At least for [tʰ], VOT is subject to domain-initial strengthening effects in unaccented words and realized with a slightly longer VOT compared to unaccented utterance-medial [tʰ] (Cho & Keating, 2009; see also Pierrehumbert & Talkin, 1992). The VOT of voiceless stops is also longer in monosyllabic words than in polysyllabic words (Klatt, 1975; Flege et al., 1998). Among lexical properties, more frequent words tend to have shorter VOTs (Yao, 2009), and the VOT of word-initial voiceless stops is longer in words with a voiced-initial neighbor (e.g., Baese-Berk & Goldrick, 2009; Kirov & Wilson, 2012; Buz, Tanenhaus, & Jaeger, 2016). Finally, faster speaking rates result in a significant decrease in VOT (e.g., Miller, Green, & Reeves, 1986; Kessinger & Blumstein, 1997, 1998; Allen & Miller, 1999; Allen, Miller, & DeSteno, 2003; Theodore et al., 2009).

Significant variability in VOT has also been identified across talkers, even after controlling for differences in speaking rate (e.g., Allen et al., 2003; Theodore et al., 2009). Variability across talkers, particularly among the voiceless categories, can span tens of milliseconds, making this source one of the larger factors in VOT variation. Socio-indexical factors, such as differences in dialect (e.g., Scobbie, 2006), gender (e.g., Smith, 1978; Swartz, 1992; Byrd, 1993; Whiteside & Irving, 1998), and age (e.g., Benjamin, 1982; Morris & Brown, 1994; Torre & Barlow, 2009; Kleinschmidt & Jaeger, submitted), as well as physiological factors such as lung volume (Hoit, Solomon, & Hixon, 1993) have all been implicated in talker-specific VOT variation.

1.2. Covariation of acoustic-phonetic properties

Patterns of covariation can hold among multiple acoustic-phonetic dimensions or cues (i.e., 'between-cue' covariation) or among multiple categories on the same dimension (i.e, 'between-category' covariation).[2] For example, fundamental frequency (f0) and vowel height (as indexed by the first formant, F1) are known to be positively correlated in many languages (e.g., Whalen & Levitt, 1995; Assmann, Nearey, & Bharadwaj, 2008). Within stop categories, many studies have examined two cues for voice, VOT and onset f0, to determine whether they combine to enhance the phonological contrast (positive correlation) or participate in a trading relation (negative correlation;

[1] Throughout, 'voiceless' and 'voiced' are used as convenient and traditional terms to refer to the voiceless aspirated (fortis, long-lag) and unaspirated (lenis, short-lag) stops, respectively. We transcribe the latter as [b d g], even though these sounds are known to lack consistent phonetic voicing for many speakers in at least utterance-initial position (e.g., Lisker & Abramson, 1964; Davidson, 2016; but cf. Jacewicz, Fox, & Lyle, 2009; Hunnicutt & Morris, 2016). For discussion of the phonological representation of this contrast in AE and other languages, see for example Kingston and Diehl (1994) and Beckman, Jessen, and Ringen (2013).

We did not measure voicing during stop closure as this can take a variety of context-dependent forms, and need not be contiguous with the release of the stop, making negative (or lead) VOT values difficult to compare with positive (or lag) VOTs (e.g., Docherty, 1992; Möbius, 2004; Davidson, 2016). It could be that the presence, amount, or profile of closure voicing would correlate with positive VOT across talkers, but we leave this for future studies.

[2] Another important type of phonetic covariation exists among mutually-enhancing articulations (e.g., vowel height and height of the soft palate; Kingston, 1992).

e.g., Shultz, Francis, & Llanos, 2012; Dmitrieva, Llanos, Shultz, & Francis, 2015; Kirby & Ladd, 2015; Clayards, in press). A weak positive correlation has been observed between f0 and VOT for [pʰ] while a weak negative correlation has been found for [b] in AE (Dmitrieva et al., 2015; Clayards, in press), but this relation appears to vary considerably by stop and language (e.g., Kirby & Ladd, 2015).

The present study focuses on patterns of talker-specific realization of stops along the positive VOT dimension. Between-category covariation of the type investigated here has long been observed for vowels: talkers have relatively congruent F1 × F2 vowel spaces that can be mapped to one another by (log-)linear translations (e.g., Joos, 1948; Nearey, 1978; Nearey & Assmann, 2007). The consistent relations among spectral and temporal properties of vowels are largely preserved even across different speaking styles (Smiljanić & Bradlow, 2008; DiCanio, Nam, Amith, García, & Whalen, 2015). Similarly, between-category covariation has been found for sibilant fricatives: while the spectral center of gravity (COG) distribution of one talker's [s] may overlap almost entirely with another talker's [ʃ], each talker nonetheless maintains a systematically higher COG for [s] than for [ʃ] (Newman et al., 2001), and the differences among talker's fricative systems on the COG dimension have been modeled with a single linear offset (McMurray & Jongman, 2011).

Previous research has observed that the VOT values of different stops covary across AE speakers in laboratory speech (i.e., single words produced in isolation or in carrier phrases). In the earliest relevant study, Zlatin (1974) reported moderate correlations of talker-specific VOT means among voiceless stops (ranging from $r=0.54$ to 0.57) and among voiced stops ($r=0.46$ to 0.54). Correlations between stops of different voicing specifications and between stops differing in both voice and place were inconsistent in Zlatin's study, most failing to reach significance. Subsequent studies documenting VOT covariation include Koenig (2000) and Newman (2003). Koenig (2000) observed a significant correlation of median VOTs between word-initial [pʰ] and [tʰ] across adult and child talkers ($r=0.78$), and Newman (2003) found significant correlations among voiceless stops ($r=0.88$ to 0.96) and among voiced stops ($r=0.54$ to 0.75), but much weaker relations between stops differing in voice ($r=-0.06$ to 0.37) in CV syllable productions by adults. More recently, Theodore et al. (2009) made the important observation that the difference in VOT means for [pʰ] and [kʰ] was relatively constant across talkers — a clear indicator of covariation between these two stops. Theodore et al. further established that the relationship between [pʰ] and [kʰ] remained even when the potentially confounding factor of utterance-level speaking rate was taken into account (using the method of Allen et al., 2003).[3]

To a large extent, covariation of spectral properties (e.g., vowel formants, fricative spectral shape) can be attributed to talker-specific anatomical properties, such as the length and shape of the vocal tract, that have a direct physical relation to resonant frequencies. While physiological and aerodynamic accounts have been offered for VOT differences across place of articulation in unaspirated stops, extension of such mechanical explanations to aspirated stops has been vexed (see Hoole, 1997 and Cho & Ladefoged, 1999 for extensive reviews). From a strictly articulatory perspective, it would be possible for an AE talker to systematically produce [pʰ] with a VOT that is long relative to the population average but [kʰ] with a VOT that is relatively short. However, the studies just reviewed and the current findings indicate that talker-specific VOT values do not vary independently in this way. Rather, there is an underlying uniformity constraint ensuring that a talker with a relatively long VOT mean for one stop also has relatively long means for the other stops (and similarly for short values; see Section 4.1). The same constraint is plausibly crucial for understanding acoustic-phonetic covariation on other durational and spectral dimensions, as physical accounts of patterns in the output of speech are always conditioned on structure in phonetic inputs or targets (e.g., Keating, 1985: 126–127).

1.3. Current study

The current study investigates VOT covariation among all six stops in both isolated speech (Section 2) and a multi-talker corpus of connected read speech, the Mixer 6 corpus (Section 3), which provides greater insight into VOT variation and patterns in the larger population. Previous studies have been limited to isolated speech, and with the exception of Theodore et al. (2009) have not analyzed talker-specific VOT patterns while taking into account the many other sources of VOT variation discussed earlier. Substantial variability in talker mean VOTs was observed within each stop category, yet in a series of analyses this variation was found to be highly structured across talkers. In Section 4, we discuss the implications of this structure for perceptual adaptation and constraints on the phonetic grammar. Finally, we consider future directions for research on patterned variation across other contexts, segments, and languages, and summarize our findings.

2. Covariation of VOT in isolated speech

The goal of our first study was to replicate and extend previous findings of VOT covariation in isolated speech. Structured variability was explored through the examination of (i) correlations, (ii) ordinal and linear relations among the talker-specific means, and (iii) a mixed-effects model. First, we assessed the strength of mutual predictability through correlations of stop means across talkers. The same analysis was performed on talker-specific means corrected for speaking rate. In addition, we examined whether the means and standard deviations of talker-specific VOT distributions covary.

Previous studies of place effects on VOT have focused primarily on ordinal rankings. We identified the rankings present in our data, but found that simple linear regressions of one stop mean against another to be more revealing. Finally, the VOT data was submitted to a mixed-effects linear regression model that included many of the predictors described in the

[3] The review in the text focuses on American English. Approximately constant VOT differences among aspirated stops have also been observed for speakers of Shetlandic English (Scobbie, 2005) as well as for speakers of Southern British English and Catalan at varying speech rates (Solé & Estebas, 2000; see also Solé, 2007). Solé and Estebas (2000) found that the pattern in English holds most clearly for labial and velar stops, with the VOT of the coronal stop perhaps varying more idiosyncratically across talkers or rates. This is likely related to other findings, including our own, that aspirated coronal stops do not consistently conform to the generalization that VOT increases with more posterior place of articulation (see Sections 1.1 and 2.2).

introduction. The random effect estimates of such a model help to identify the major sources of variation across talkers.

2.1. Methods

2.1.1. Participants

Twenty-four students at Johns Hopkins University (13 female) participated in the experiment and received $10 or partial course credit. All participants were native speakers of American English. Data from 18 of the participants were previously reported in Chodroff and Wilson (2014).

2.1.2. Procedure and measurements

Stop-initial CVC syllables were elicited in the carrier phrase "Say ___ again." The syllables were composed of the six stop consonants [pʰ tʰ kʰ b d g] crossed with ten vowels [i ɪ eɪ ɛ æ ʌ a ɔ oʊ u].[4] The final consonant was always the voiceless coronal stop. One CVC combination was omitted because it formed a taboo word.

Each syllable was assigned an orthographic form according to standard conventions for American English spelling, with the constraint that the consonant and vowel mappings were one-to-one for all stimuli regardless of lexical status. Participants completed five blocks, each syllable occurring once per block. This resulted in a maximum of 50 tokens per stop consonant and participant, except for [tʰ], in which case there was a maximum of 45 tokens per participant. Four participants did not receive the final block due to a programming error.

Stimuli were randomized within each block separately and presented with PsychoPy (Peirce, 2007). Each stimulus was displayed in the frame with a rhyming reference word, used to specify the intended pronunciation of the vowel spelling. The recordings were made in a sound-attenuated booth with a Shure SM58 microphone and Zoom H4n digital recorder at a sampling rate of 48 kHz (16 bit). The experiment was self-paced and participants were given short breaks between blocks. A total of 6776 tokens were analyzed (68 additional tokens were omitted due to pronunciation error).

Initial segmentation of the recordings was performed with the Penn Phonetics Lab Forced Aligner (P2FA; Yuan & Liberman, 2008). VOT boundaries for all word-initial stop consonants were then manually placed on the basis of waveform and spectrogram displays in Praat (Boersma & Weenink, 2015). VOT was defined as the duration of the interval from the beginning of the stop release to the start of periodicity in the waveform or a visible f0 track (whichever came first). This measure did not take into account any closure voicing, and as discussed in the introduction, is therefore more properly called positive (or lag) VOT. No attempt was made to distinguish among components of the release (i.e., transient, frication, and any following aspiration). In addition, local speaking rate was operationalized as the duration of the vowel in each trial (as in Theodore et al., 2009); this was determined from the manually-aligned stop release offset (equivalently, the vowel onset) and the vowel offset as marked by P2FA.

[4] The contrast between /a/ and /ɔ/, represented orthographically in our materials by <O> and <AUGH>, may not have been present in the dialects of all of our speakers (e.g., Kurath & McDavid, 1961).

Table 1
Descriptive statistics of talker-specific VOT (ms) for each stop category in the isolated speech. The mean and standard deviation were calculated from the population sample of talker-specific means. Ranges are reported for talker-specific means and standard deviations.

Stop	Mean	SD	Range of Talker Means	Range of Talker SDs
pʰ	89	27	46–139	12–27
tʰ	98	28	57–156	10–26
kʰ	99	24	67–137	11–20
b	13	5	11–20	2–8
d	21	7	14–32	3–10
g	28	10	19–42	4–13

2.2. Results

Stop VOT means varied substantially across talkers: for example, the difference between the lowest and highest talker-specific values for [tʰ] approached 100 ms (see Table 1). The distributions of talker means are shown as marginal histograms in Fig. 1. The grand means for the voiceless stops were somewhat higher than figures previously reported for AE laboratory speech; we speculate that this reflects an overall slow speaking rate in the current experiment.

2.2.1. Correlation analyses

The key finding was that the means of several stops were highly correlated across talkers. The correlations among voiceless stops were nearly perfect ($r = 0.95$ to 0.96; $ps < 0.006$), and moderate but significant correlations were observed among the voiced stops (rs = [b–d] 0.54, $p = 0.006$; [d–g] 0.56,[g–b] 0.56, $ps < 0.006$). Correlations between homorganic stop pairs failed to reach significance ($r = 0.18$ to 0.33, $ps > 0.006$). All of the correlations are reported in Table 2 and in Fig. 1 together with best-fit linear regression lines.[5,6]

Two additional analyses were performed to estimate the strength of the correlations in the larger population of AE talkers and to control for speaking rate variation. For each pair of stops separately, a confidence interval for the VOT correlation was estimated with a bootstrap procedure. In each of 1000 repetitions, a correlation was computed from a random sample (with replacement) of the talker-specific means for the two stops. The results of the repetitions were then combined to form a 95% confidence interval according to the bias-corrected and accelerated percentile (BCa) method (Efron, 1987). For instance, the bootstrap interval for [pʰ] and [kʰ] ranges from $r = 0.86$ to 0.99, suggesting that the point estimate ($r = 0.95$) did not arise from a handful of outliers (though the correlation in the population may be somewhat smaller).

The second analysis was performed on the residuals of a simple linear regression in which each VOT value was predicted from the corresponding speaking rate (operationalized as vowel

[5] Throughout the paper, the nominal alpha value of 0.05 was Bonferroni-corrected for multiple comparisons. For completeness, we present all relevant correlations even when they are non-independent; this redundancy is eliminated in the mixed-effects analysis reported further below.

[6] Previous work has also modeled VOT on the log scale given the non-linear perception of temporal properties and the large difference in variances between the voiced and voiceless categories (Volaitis & Miller, 1992; Kong, 2009; Sonderegger, 2015). The correlations of talker log VOT means, calculated as the mean of the logged VOTs, resulted in magnitudes comparable to the correlations of talker (linear) VOT means, but the pattern of significance did change (rs = [pʰ–tʰ] 0.96, [tʰ–kʰ] 0.97, [kʰ–pʰ] 0.96, $ps < 0.006$; [b–d] 0.51, [d–g] 0.50, [g–b] 0.51, [pʰ–b] 0.18, [tʰ–d] 0.36, [kʰ–g] 0.17, n.s.).

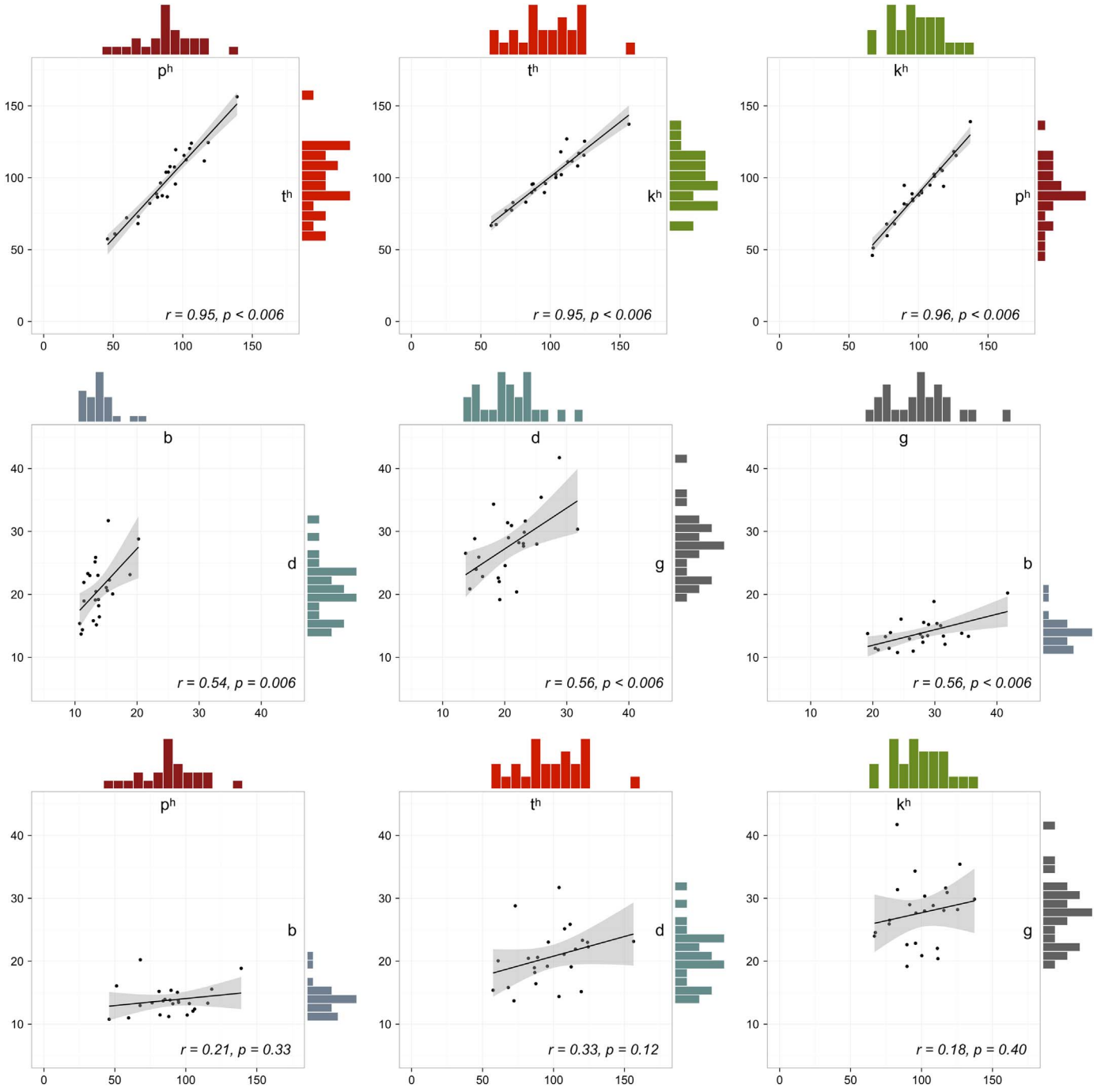

Fig. 1. Variation and covariation of stop VOT means (ms) across talkers in the isolated speech. Marginal histograms show variation in talker means. The top row shows correlations among the voiceless stops, the middle row among the voiced stops (note change of scale), and the bottom row within homorganic pairs. Gray shading reflects the local confidence interval around the best-fit linear regression line.

duration). The residualized VOTs were then averaged by talker and stop category, just as before, and the correlations were recomputed. The magnitudes of the correlations among voiceless stops did not deviate from the original magnitudes, demonstrating that differences among talkers in the realization of these sounds cannot be reduced to talker-specific speaking rates. Among the voiced stops and between homorganic pairs, the correlations increased considerably and reached significance (*voiced*: $r = 0.80$ to 0.89; *homorganic*: $r = 0.60$ to 0.72; $ps < 0.006$). Differences in speaking rate thus appear to have obscured these relationships in the raw data. Bootstrap confidence intervals again indicated that these correlations were consistent in the population from which our speakers were sampled.[7]

The correlations among stop means suggest that variability is highly structured across talkers. Additional structure in phonetic realization may also exist between talker-specific means and standard deviations, and would be expected given previous correlations and relationships found in phonetic temporal measures (e.g., Byrd & Saltzman, 1998; Shaw, Gafos, Hoole, &

[7] This analysis residualized the dependent variable (VOT) against a predictor (speaking rate/vowel duration), and thus was not subject to the pitfalls of residualizing one predictor against another (Wurm & Fisicaro, 2014).

Table 2
Pearson correlation coefficients and 95% BCa bootstrap confidence intervals of talker means for raw and residualized VOT (ms) in the isolated speech.

	Raw VOT			*Residualized VOT*		
	Pearson's *r*	*p*-value	95% CI	Pearson's *r*	*p*-value	95% CI
p^h – t^h	0.95	<0.006	[0.90, 0.98]	0.95	<0.006	[0.89, 0.98]
t^h – k^h	0.95	<0.006	[0.86, 0.98]	0.95	<0.006	[0.88, 0.98]
k^h – p^h	0.96	<0.006	[0.86, 0.99]	0.96	<0.006	[0.88, 0.99]
b – d	0.54	0.006	[0.21, 0.77]	0.89	<0.006	[0.69, 0.95]
d – g	0.56	<0.006	[0.23, 0.78]	0.80	<0.006	[0.54, 0.91]
g – b	0.56	<0.006	[0.21, 0.84]	0.82	<0.006	[0.56, 0.91]
p^h – b	0.21	0.33	[−0.42, 0.76]	0.72	<0.006	[0.44, 0.90]
t^h – d	0.33	0.12	[−0.16, 0.61]	0.64	<0.006	[0.25, 0.85]
k^h – g	0.18	0.40	[−0.25, 0.53]	0.60	<0.006	[0.25, 0.82]

Zeroual, 2009; Turk & Shattuck-Hufnagel, 2014). Indeed, increased temporal durations have been shown to correspond with greater variability throughout human motor behavior (Schmidt, Zelaznik, Hawkins, Frank, & Quinn, 1979; Schöner, 2002). Significant correlations of the talker means and standard deviations were observed for all stops ($r=0.90$), as well as for voiced stops ([b]: $r=0.71$, [d]: $r=0.76$, [g]: $r=0.75$, $ps<0.008$). Moderate correlations were also observed for the voiceless stops; however, these failed to reach significance after correction for multiple comparisons ([p^h]: $r=0.47$, $p=0.02$; [t^h]: $r=0.53$, $p=0.008$; [k^h]: $r=0.43$, $p=0.04$). These correlations likely reflect restricted variation at the lower boundary for each voicing category: a lower bound at 0 ms for voiced stops and at the auditory boundary between the categories for the voiceless stops.

2.2.2. Ordinal and linear relations

Previous studies have generally considered the relationships among VOT means in terms of ordinal rankings (e.g., Peterson & Lehiste, 1960; Cho & Ladefoged, 1999). For comparison with these studies, we also assessed the ranking, and identified three predominant patterns across talkers: [b]<[d]<[g]<[p^h]<[t^h]<[k^h] (11 talkers), [b]<[d]<[g]<[p^h]<[k^h]<[t^h] (8 talkers), and [b]<[g]<[d]<[p^h]<[k^h]<[t^h] (3 talkers); two talkers exhibited other rankings. For all talkers and within both values of [voice], the mean dorsal VOT was longer than the mean labial VOT, consistent with cross-linguistic tendencies (e.g., Cho & Ladefoged, 1999). However, the relative ranking of coronal and dorsal means varied across talkers, with more variation among the voiceless than the voiced stops (see also Docherty, 1992; Yao, 2009).

The preceding correlations and ordinal rankings provide some information about systematic relations among talker-specific stop VOT means, but simple linear regressions can reveal additional structure. While the correlations indicate that stop-specific means are linearly related, this could take the form of a constant difference between means ($y=\beta_0+x$), a constant ratio between means ($y=\beta_1 \cdot x$), or a combination of the two ($y=\beta_0+\beta_1 \cdot x$). We performed a separate simple linear regression for each pair of stops, regressing the talker means of one stop against those of another.

Paralleling the correlation magnitudes, the proportion of variance accounted for by the regressions was largest for the voiceless stop pairs (adjusted R^2s>0.50) and smallest for the voiced stop pairs and homorganic pairs (adjusted R^2s<0.50). We will discuss only the model fits for the voiceless stops, but for completeness all models are reported in Table 3.

Table 3
Additive (β_0) and scalar (β_1) components of simple linear regressions on talker mean VOTs of one stop predicted from another in isolated speech. For each pair, the dependent variable is given first followed by the independent variable.

	β_0	*p*-value	β_1	*p*-value	*Adj.* R^2
t^h ~ p^h	5.15	0.43	1.05	<0.003	0.91
k^h ~ t^h	24.66	<0.003	0.76	<0.003	0.90
k^h ~ p^h	24.37	<0.003	0.85	<0.003	0.92
d ~ b	6.06	0.23	1.06	<0.003	0.28
g ~ d	14.22	0.004	0.65	0.005	0.26
g ~ b	10.00	0.09	1.28	0.004	0.29
p^h ~ b	62.62	0.03	1.89	0.33	0.00
t^h ~ d	63.36	0.009	1.70	0.12	0.07
k^h ~ g	81.83	<0.003	0.64	0.40	−0.01

Table 4
Standard deviations of talker random effects in the maximal mixed-effects model of VOT in isolated speech.

Random effect for talker	SD
intercept	11.17
voice	10.40
poaCor	2.63
poaDor	2.63
speaking rate	2.26
voice × poaCor	2.16
voice × poaDor	1.63

In predicting [k^h] from either [t^h] or [p^h], both the intercept and scaling factors were significant ([k^h~t^h]: $\beta_0=24.66$, $\beta_1=0.76$; [k^h~p^h]: $\beta_0=24.37$, $\beta_1=0.85$; $ps<0.003$). The linear fits inherently account for the ordinal rankings: [k^h]>[t^h], [p^h] is expected over much of the empirical range of VOT values; however, [t^h] and [p^h] also increase faster relative to [k^h], resulting in a point at which the ranking is reversed. For [t^h] and [k^h] in particular, this point is within the reasonable range of values for isolated speech (103 ms). In the model predicting [t^h] from [p^h], only the scaling factor was significant, indicating a straightforwardly proportional relationship ([t^h~p^h]: $\beta_0=5.15$, $p=0.43$, $\beta_1=1.05$, $p<0.003$).

Linear regression models like these have been employed in automatic approaches to speaker adaptation, as pairwise regressions between speech sounds and classes of sound allow for more rapid talker adaptation from limited talker-specific data (Furui, 1980; Cox, 1995). Strong linear relationships among talker-specific realizations of speech sounds could also have implications for cognitive models of adaptation,

accounting for how listeners form expectations about the realization of unheard speech sounds after limited exposure (see Section 4.2).

2.2.3. Mixed-effects analysis

A mixed-effects linear regression model provided further statistical support for the findings reported above while allowing us to investigate additional factors known to influence VOT (Baayen, Davidson, & Bates, 2008). In addition to the factors already considered (i.e., the voice contrast, place of articulation, and speaking rate), the model included properties of the following vowel that are known to condition VOT (i.e., vowel height and tenseness: Klatt, 1975; Port & Rotunno, 1979; Nearey & Rochet, 1994). While the manipulation of vowel properties was balanced across participants in our study, and therefore could not provide an alternative explanation for the speaker differences or the correlations among categories, it is important to identify the signature of phonetic covariation in mixed-effect models. We analyzed the random effect component of the fit model to demonstrate that much of the variability in VOT across participants was due to differences in overall mean (intercept) and in the magnitude of the voicing contrast. Unlike the descriptive analyses reported above, the method of this section is more general: it can be employed for data sets in which vowel and other factors are not balanced across speakers provided there is sufficient data (e.g., in an analysis of spontaneous speech).

The model included fixed effects of phonological voice, place of articulation, speaking rate, vowel height, vowel tenseness, as well as the two-way voice × place, voice × rate, and height × tenseness interactions. All categorical factors were weighted effect coded to correct for slightly unequal sample sizes (Darlington, 1990; p. 246). The coding of the categorical variables was as follows, with contrast weighting reported in the parentheses: phonological voice (*voice*: voiceless = 1, voiced = −0.97), place of articulation (*poaCor*: coronal = 1, dorsal = 0, labial = −0.96; *poaDor*: coronal = 0, dorsal = 1, labial = −1), vowel height (*height*: high = 1, non-high = −0.41); vowel tenseness (*tense*: tense = 1, lax = −1.57). The continuous factor of speaking rate was *z*-scored using the mean and standard deviation ($\mu = 167$ ms, $\sigma = 43$ ms) computed from all vowels collapsed across participants. Similarly, the dependent variable (VOT) was centered at zero by subtracting the grand mean ($\mu = 57$ ms) from each value.

The random effect for speaker included an intercept and slopes for voice, place, rate, and voice × place. While an attempt was made to include random slopes for vowel height and tenseness, these led to non-convergence and were removed. There was also a random intercept for syllable rime (VC portion), which is known to be a salient sublexical unit for English speakers (e.g., De Cara & Goswami, 2002).

The model revealed significant main effects of voice (*voice*: $\beta = 37.30$, $t = 17.50$) and place (*poaCor*: $\beta = 1.48$, $t = 2.56$; *poaDor*: $\beta = 5.50$, $t = 9.55$).[8] The effect of voice was significantly modulated by place, reflecting differences in the rankings of place of articulation across the two voicing categories. Compared to predictions from voice and place alone, coronal stops were significantly longer when voiceless than when voiced (*voice* × *poaCor*: $\beta = 1.38$, $t = 2.79$), whereas voiceless dorsal stops were significantly shorter (*voice* × *poaDor*: $\beta = -1.46$, $t = -3.71$). There was also a main effect of speaking rate, and slightly shorter VOTs were found at faster speaking rates (*rate*: $\beta = -2.46$, $t = -4.55$). (The coefficient for speaking rate can be interpreted as the predicted change in VOT in milliseconds given a one standard deviation change in speaking rate.) The effect of rate was tempered by a significant interaction with voice, in which the effect of rate was enhanced for voiceless stops in comparison to voiced stops (*voice* × *rate*: $\beta = 0.39$, $t = 1.98$). Vowel height, vowel tenseness, and their interaction did not reach significance (*height*: $\beta = 0.37$, $t = 0.29$; *tense*: $\beta = 1.11$, $t = 1.82$; *height* × *tense*: $\beta = 1.36$, $t = 1.39$).

The random effect estimates can provide insight into the major sources of talker variation. As shown in Table 4, the random intercept and the voice slope had the largest standard deviations, indicating substantial differences across talkers in overall mean VOT and in the magnitude of separation between voiced and voiceless stops. In comparison, the variances for the other random talker slopes were much smaller (e.g., the variance of the voice slope was about four times that of either place effect). This is consistent with the finding of Theodore et al. (2009) that there are significant differences across talkers in the intercept, or overall mean, but not in the effect of place of articulation (for [pʰ] and [kʰ]).[9]

It is well known, and confirmed by our data, that there is greater VOT variation for voiceless stops than for voiced stops (see Fig. 1; Dmitrieva et al., 2015). This presumably reflects both a relatively fixed auditory boundary between the voicing categories (e.g., Kuhl, 1981) and, in our study, the lower bound on positive VOT measurements. Therefore, a speaker with a higher overall mean VOT is very likely to have a larger separation between voiced and voiceless stops (thus ensuring that the voiced stops lie below the boundary); and indeed, the random intercept and voice slope were tightly correlated ($r = 0.97$). While this might suggest that voiceless and voiced stops should simply be analyzed separately, the correlations within homorganic pairs reported earlier indicate that some component of talker-specific VOT is shared by all of the stops.

2.3. Discussion

Despite substantial talker variation in VOT values, highly stable relations were observed among categories across talkers. These results are consistent with previous laboratory findings of correlations in talker means, but extend the findings to all six stops while also controlling for other sources of variability such as differences in speaking rate. The correlation and random effect analyses both provide evidence for the existence of strong positive linear relationships in talker VOT. In addition, there were consistent ordinal rankings of stop VOT,

[8] A *t*-value with magnitude greater than 2.0 was considered significant.

[9] The model reported here performed significantly better than models with simpler random effect structures for talker as determined by log-likelihood ratio tests and Bayesian Information Criterion (BIC) comparisons. However, inclusion of additional factors beyond the intercept and voice gave diminishing returns in accounting for VOT variability. In comparison to a model with no talker-specific random effect, the BIC decreased by 5293 for a model with a random intercept and voice slope for talker, but only by a further 150 units for the maximal random effect model reported in the main text.

with talkers predominantly exhibiting a lower mean VOT for labials than for dorsals within each voicing category. Yet, in describing the relation between VOT means, the linear relationships not only captured the ordinal rankings, but also accounted for the variability in the ranking of coronals and dorsals, and critically, quantified the typical magnitude of separation between VOT means.

These results establish that the VOT means of AE stops are highly structured across talkers in isolated speech. However, it remains unclear whether similar patterns would also be observed in the production of known lexical items in connected speech. The following study addressed this question by examining patterns of talker VOT in a large corpus of read speech that contained a greater variety of prosodic and lexical factors, but otherwise matched sentential conditions for each talker. This allowed for analysis of VOT as produced in a more natural and connected speech style, while also ensuring that talkers were producing approximately the same content.

Table 5
Range and median number of tokens per talker and stop category, and total number of tokens per stop category in the connected speech.

Stop	Range	Median	Total
p^h	44–100	77	13,517
t^h	17–77	46	8,218
k^h	46–114	82	14,619
b	42–117	80	14,661
d	58–184	131	23,086
g	52–118	82	14,763

Table 6
Descriptive statistics of talker-specific VOT (ms) for each stop category in the connected speech. The mean and standard deviation were calculated from the population sample of talker-specific means. Ranges are reported for talker-specific means and standard deviations.

Stop	Mean	SD	Range of Talker Means	Range of Talker SDs
p^h	51	9	28–78	11–35
t^h	61	9	40–96	9–34
k^h	56	8	36–79	11–30
b	8	2	6–14	2–8
d	14	3	8–22	4–13
g	17	3	9–28	6–15

3. Covariation of VOT in connected speech

Phonetic research has increasingly employed large connected speech corpora (e.g., Byrd, 1992; Cole, Choi, Kim, & Hasegawa-Johnson, 2003; Yuan & Liberman, 2008). While laboratory conditions ensure a greater degree of control, speech corpora can provide great quantities of naturally-occurring speech. Large-scale corpus studies have been conducted for many aspects of speech, including but not limited to segmental realization (e.g., Byrd, 1992), coarticulatory and contextual effects (Keating, Byrd, Flemming, & Todaka, 1994; Gendrot & Adda-Decker, 2005; Bürki, Ernestus, Gendrot, Fougeron, & Frauenfelder, 2011; Schuppler, Ernestus, Scharenborg, & Boves, 2011; Torreira & Ernestus, 2012; Elvin & Escudero, 2014; Yu, Abrego-Collier, Phillips, Pillion, & Chen, 2015), prosodic structure and speaking rate (e.g., Ostendorf, 2001; Kendall, 2009), and phonetic change over time (e.g., Fruehwald, 2013; Labov, Rosenfelder, & Fruehwald, 2013).

Many techniques originally developed for automatic speech recognition (ASR) have facilitated phonetic analysis of large corpora (e.g., Yuan & Liberman, 2008; Rosenfelder, Fruehwald, Evanini, & Yuan, 2011; Yoon & Kang, 2013). These include algorithms for extracting VOTs values (e.g., Das & Hansen, 2004; Yao, 2007; Sonderegger & Keshet, 2010), vowel formants (Evanini, Isard, & Liberman, 2009; Yao, Tilsen, Sprouse, & Johnson, 2010), and degrees of vowel nasalization (Yuan & Liberman, 2011), as well as for prosodic labeling (e.g., Wightman & Ostendorf, 1994; Hasegawa-Johnson et al., 2005; Gorman, Howell, & Wagner, 2011). With respect to VOT in particular, large-scale analyses have examined population-level VOT distributions (Byrd, 1993), phonetic accommodation over time (Sonderegger, 2015), dialectal differences (Stuart-Smith, Rathcke, Sonderegger, & Macdonald, 2015), and effects of prosodic structure (Cole, Kim, Choi, & Hasegawa-Johnson, 2007), among others.

The corpus employed in our analysis, the Mixer 6 corpus (Brandschain, Graff, Cieri, Walker, & Caruso, 2010; Brandschain, Graff, & Walker, 2013), is well-suited for the study of variation across talkers. The complete corpus contains speech sampled at 16 kHz from approximately 600 AE talkers recorded in one to three separate sessions. In each session, the participant completed an interview (15 min), transcript reading (15 min), and telephone call (10 min), and sessions were separated by at least two days. The corpus was collected primarily to support research in speaker recognition technologies; however, the read transcript portion was specifically added to foster basic scientific research on talker characteristics (for further details, see Brandschain et al., 2010 and Chodroff, Maciejewski, Trmal, Khudanpur, & Godfrey, 2016).

Table 7
Pearson correlation coefficients and 95% BCa bootstrap confidence intervals of talker means for raw and residualized VOT (ms) in the connected speech.

	Raw VOT			***Residualized VOT***		
	Pearson's *r*	*p*-value	95% CI	Pearson's *r*	*p*-value	95% CI
p^h – t^h	0.83	<0.006	[0.77, 0.88]	0.81	<0.006	[0.74, 0.86]
t^h – k^h	0.77	<0.006	[0.71, 0.82]	0.75	<0.006	[0.67, 0.80]
k^h – p^h	0.82	<0.006	[0.77, 0.86]	0.80	<0.006	[0.74, 0.85]
b – d	0.07	0.33	[−0.05, 0.19]	−0.03	0.65	[−0.15, 0.09]
d – g	0.33	<0.006	[0.20, 0.46]	0.20	0.008	[0.05, 0.33]
g – b	0.49	<0.006	[0.36, 0.59]	0.41	<0.006	[0.28, 0.53]
p^h – b	0.15	0.05	[−0.01, 0.30]	-0.11	0.17	[−0.27, 0.18]
t^h – d	0.53	<0.006	[0.43, 0.63]	0.40	<0.006	[0.28, 0.52]
k^h – g	0.40	<0.006	[0.29, 0.50]	0.27	<0.006	[0.14, 0.39]

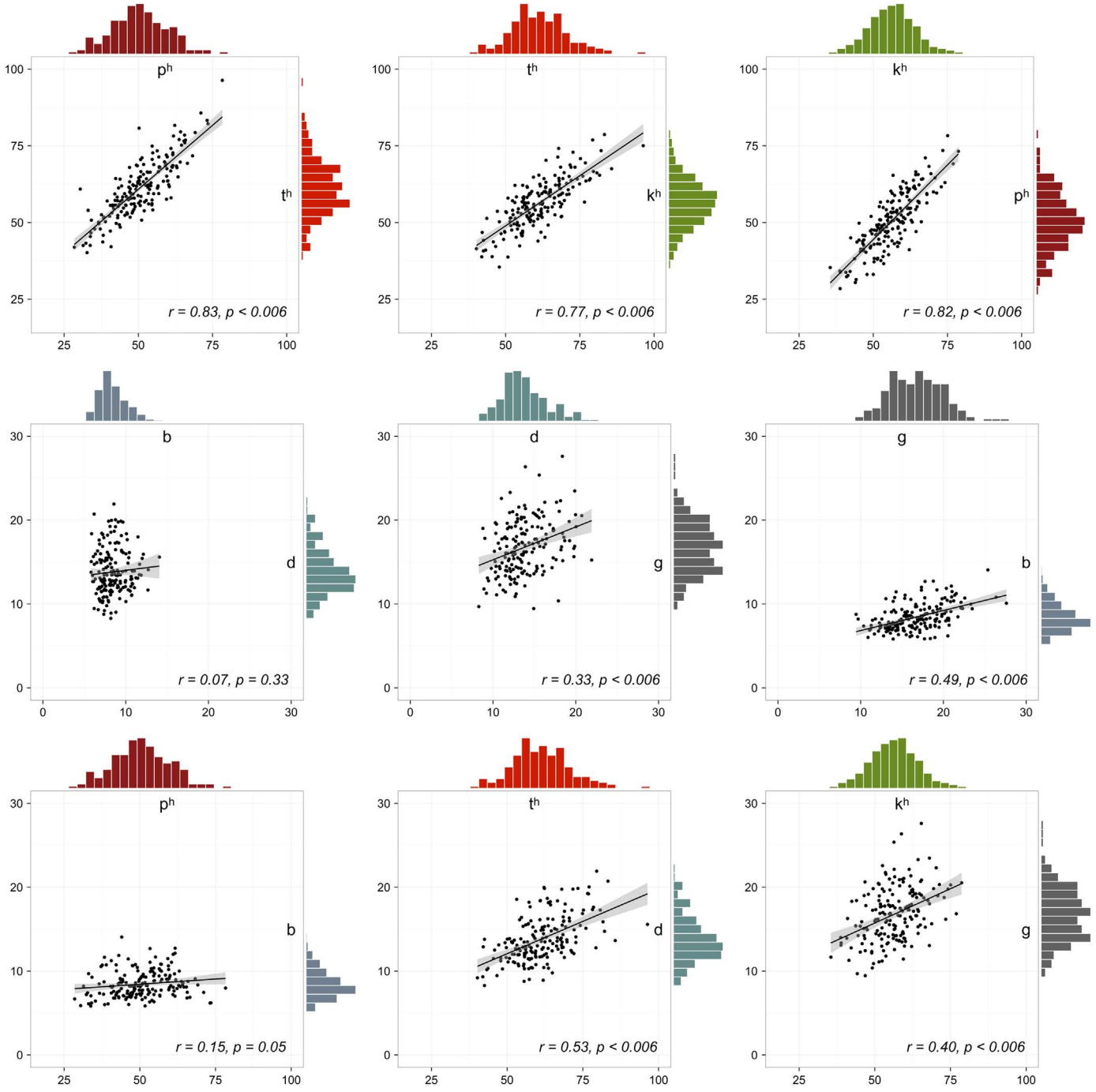

Fig. 2. Variation and covariation of stop VOT means (ms) across talkers in the connected speech. Marginal histograms show variation in talker means. The top row shows correlations among the voiceless stops, the middle row among the voiced stops (note change of scale), and the bottom row within homorganic pairs. Gray shading reflects the local confidence interval around the best-fit linear regression line.

Other transcribed speech corpora can provide a large number of speakers (e.g., Switchboard: Godfrey, Holliman, & McDaniel, 1992; TIMIT: Garofolo et al., 1993), a large number of data points per talker (e.g., Buckeye Corpus: Pitt, Johnson, Hume, Kiesling, & Raymond, 2005), or even the combination of these two (e.g., Wall Street Journal Corpus: Paul & Baker, 1992; LibriSpeech: Panayotov, Chen, Povey, & Khudanpur, 2015). A unique advantage of the Mixer 6 read speech portion is that it provides a large sample for each talker while holding constant prosodic, lexical, and syntactic/semantic factors. This allowed us to investigate talker variation at the level of phonetic categories without a major confound of sentential content.

The same set of analyses as presented in the preceding study were used to assess the extent of structured VOT variation in the connected speech study. Recall from Section 2 that this includes a correlation analysis, an examination of the ordinal and linear relationships among talker means, and finally, an analysis of the talker-specific random effect variances in a linear mixed-effects model.

Table 8
Additive (β_0) and scalar (β_1) components of simple linear regressions on talker mean VOTs of one stop predicted from another in connected speech. For each pair, the dependent variable is given first followed by the independent variable.

	β_0	*p*-value	β_1	*p*-value	*Adj.* R^2
$t^h \sim p^h$	19.09	<0.003	0.83	<0.003	0.69
$k^h \sim t^h$	16.25	<0.003	0.65	<0.003	0.60
$k^h \sim p^h$	21.11	<0.003	0.70	<0.003	0.68
d ∼ b	12.76	<0.003	0.13	0.33	0.00
g ∼ d	11.35	<0.003	0.39	<0.003	0.10
g ∼ b	8.25	<0.003	1.01	<0.003	0.24
$p^h \sim$ b	43.08	<0.003	0.89	0.05	0.02
$t^h \sim$ d	35.84	<0.003	1.84	<0.003	0.28
$k^h \sim$ g	35.57	<0.003	1.00	<0.003	0.16

3.1. Methods

3.1.1. Corpus description

The following analysis employed an audited subset of the Mixer 6 read speech for 180 native AE talkers (102 female, 78 male). Each talker recorded three read speech sessions, resulting in approximately 45 min of speech. The script contained 335 selected sentences randomly drawn from utterances in the Switchboard corpus. The selected sentences were therefore naturally occurring and not selected for the research question at hand. Each selected sentence contained 1 to 17 words with a median of 7 words. Participants read the selected sentences in a fixed order in each session until 15 min had passed. The number of sentences completed and read correctly within each session ranged from 103 to 338 (median: 238; mean: 239).

All talkers in the present analysis were born in the United States: 83 were from Pennsylvania (57 from Philadelphia), 48 from other Northeast states, 18 from the Southeast, 14 from the Midwest, 11 from the West, and 6 from the Southwest. Talkers ranged in age from 18 to 86 years (median: 27 years).

3.1.2. Acoustic measurements

VOT measurements were extracted for all stops that appeared word-initially, in any utterance position, and that were followed immediately by vowels transcribed as bearing primary stress. Prior to measurement, reading and recording errors were removed with a combination of automatic and manual methods (for details see Chodroff et al., 2016). The cleaned transcripts were phonetically aligned to the corresponding audio using P2FA. AutoVOT (Sonderegger & Keshet, 2010, 2012) was then used to locate the onset of each stop release and the onset of the following vowel using pre-trained statistical models. For voiceless stops, the temporal window for this analysis extended 30 ms before and 30 ms after the stop interval as marked by P2FA; for voiced stops, the P2FA interval was extended in both directions by 10 ms. The minimum VOT threshold, required by AutoVOT, was set to 15 ms for voiceless stops and 4 ms for voiced stops.

To estimate the accuracy of AutoVOT for this corpus, and following the same procedure as in Section 2.1.2, we hand-measured the VOTs of a randomly selected subset of the stops (more than 3,000 tokens, or approximately 3% of the data). Comparison of the automatic and manual measurements yielded a root-mean-square deviation of 12.9 ms (somewhat larger than the 7.74 ms reported by Sonderegger & Keshet, 2010 for the Big Brother Corpus).[10] An additional 936 stops with VOTs equal to the minimum threshold, or with exceptionally long values, were hand-corrected.[11] Among the hand-corrected stops, tokens lacking visible stop bursts were excluded from all analyses (209 tokens omitted).

Measurements were taken from the boundaries placed by AutoVOT or, when available, the manually-placed boundaries. Because utterances in this corpus varied considerably in length and structure, we operationalized speaking rate for each one as the average word duration determined from the P2FA boundaries.

All words were retained in the analysis with the exception of *'to'* (which was highly frequent and subject to *wanna*-contraction and other phonetic reductions). VOT values 2.5 standard deviations above or below talker-specific category means were excluded. This left a total of 88,725 measurements for analysis, with a median of 547 per talker (range: 296–741). The range and median number of tokens per talker and stop are given in Table 5 along with the total number of tokens per stop. These tokens are instances of 98 word types: 17 lexical items for [p^h], 14 for [t^h], 21 for [k^h], 18 for [b], 16 for [d], and 12 for [g].

3.2. Results

Talker-specific VOT means varied considerably within each stop category (Table 6). Within the voiceless stops, talker-specific means ranged from 28 ms to 78 ms for [p^h], from 40 ms to 96 ms for [t^h], and from 36 ms to 79 ms for [k^h]. For the voiced stops, the range in talker-specific VOT was limited by the minimum positive VOT and the voicing boundary; however, talker-specific means still differed by up to 19 ms (Table 6). The grand mean VOTs for the voiceless stops were comparable to figures reported in previous studies of read and spontaneous speech (e.g., Byrd, 1993; Yao, 2007), but overall shorter than those observed in isolated speech (e.g., Lisker & Abramson, 1964).

3.2.1. Correlation analyses

As shown in Table 7 and Fig. 2, correlations among the voiceless stop consonants were particularly strong, ranging from $r = 0.77$ to 0.83 ($ps < 0.006$). Among the voiced stops, talker means were significantly correlated between [b] and [g] ($r = 0.49$, $p < 0.006$), as well as [d] and [g] ($r = 0.33$, $p < 0.006$), but not between [b] and [d] ($r = 0.07$, $p = 0.33$). Correlations between homorganic stops were also significant for coronals and dorsals (*coronal*: $r = 0.53$, *dorsal*: $r = 0.43$, $ps < 0.006$; cf. *labial*: $r = 0.15$, $p = 0.05$).[12,13]

[10] The root-mean-square deviation for each stop category was [p^h]: 7.3 ms, [t^h]: 16.3 ms, [k^h]: 6.3 ms, [b]: 2.2 ms, [d]: 2.9 ms, and [g]: 16.9 ms.

[11] AutoVOT provides the capability of training its statistical model on a user-supplied corpus. We trained on two-thirds of our manually-measured stops (1488 voiceless, 990 voiced) and tested on the remaining third (755 voiceless, 489 voiced). The root-mean-square deviation of the resulting model (13.0 ms) was not superior to that of the pre-trained models.

[12] The fact that [p^h–b] showed the lowest correlation among homorganic stops (see Table 7) could reflect a limitation of our method; [b] is the stop most amenable to phonetic voicing, and a higher correlation may emerge when positive and negative VOTs are measured.

[13] The correlations of talker log VOT means had comparable magnitudes and the same pattern of significance as the correlations of linear VOT means (rs = [p^h–t^h] 0.79, [t^h–k^h] 0.71, [k^h–p^h] 0.79, [d–g] 0.39, [g–b] 0.43, [t^h–d] 0.54, [k^h–g] 0.44, $ps < 0.006$; [b–d] 0.07, [p^h–b] 0.18, n.s.).

The same pattern of significance emerged in the correlations of residualized talker means which were obtained after removing the effect of speaking rate on VOT with a simple linear regression (Table 7). While speaking rate was measured here as mean word duration per utterance, the same pattern of significance was also realized when speaking rate was measured as following vowel duration.[14]

Consistent with previously observed temporal patterns in speech (and other motor behaviors), strong positive correlations were also found between talker-specific means and standard deviations. Means and standard deviations were significantly correlated in an analysis of all stops together ($r=0.90$). Moderate correlations were present for each of the voiceless stops ([pʰ]: $r=0.57$, [tʰ]: $r=0.47$, [kʰ]: $r=0.51$, $ps<0.008$). For the voiced stops, strong correlations were observed within [b] and [d], and a moderate correlation was observed for [g] ([b]: $r=0.79$, [d]: $r=0.76$, [g]: $r=0.47$, $ps<0.008$).

3.2.2. Ordinal and linear relations

As in the isolated speech data, three rankings were predominant: [b]<[d]<[g]<[pʰ]<[kʰ]<[tʰ] (113 talkers), [b]<[d]<[g]<[pʰ]<[tʰ]<[kʰ] (31 talkers), or [b]<[g]<[d]<[pʰ]<[kʰ]<[tʰ] (27 talkers). Other patterns were observed for 9 talkers. For five talkers, the order was [b]<[g]<[d]<[pʰ]<[tʰ]<[kʰ], and for four talkers, [tʰ] was marginally shorter than [pʰ]. For all but three talkers and within each voicing category, the mean labial VOT was shorter than the mean dorsal VOT. In all cases, the mean VOTs for the voiceless stops were greater than the mean VOTs for the voiced stops.

As in the isolated speech study, the linear relationships between stop means were explored with simple regression models predicting the talker mean VOT of one stop from another. The additive and scalar factors for all pairwise linear regression models are provided in Table 8. The best fits, in which the proportion variance accounted for exceeded 0.50, were among the voiceless stops. In each of these models, both the intercept and scaling factor were significant, indicating a combination of additive and proportional factors in the relationship between talker means ([tʰ∼pʰ]: $\beta_0=19.09$, $\beta_1=0.83$; [kʰ∼tʰ]: $\beta_0=16.25$, $\beta_1=0.65$; [kʰ∼pʰ]: $\beta_0=21.11$, $\beta_1=0.70$; $ps<0.001$). These linear fits reflect that the fact that the differences in VOT means for [tʰ] and [pʰ] as well as [kʰ] and [pʰ] become smaller as the mean of [pʰ] increases, and that all but the lowest VOT means for [tʰ] tend to be higher than those of [kʰ] (see Fig. 2). For connected speech, these models provide the best-fitting linear description of how knowledge of one talker-specific mean could be generalized to the other voiceless stops.

3.2.3. Mixed-effects analysis

The model included all of the fixed effects considered for isolated speech (Section 2.2.3): voice, place of articulation, speaking rate, vowel height and tenseness, as well as the two-way voice × place, voice × rate, and height × tenseness interactions. (Recall that all measured stops appeared before vowels bearing primary stress, therefore effects of different stress levels or of following non-syllabic approximants could not be investigated.) In addition, there were fixed effects of the position of the word in the utterance, number of syllables in the word, and word frequency.

To accommodate unequal sample sizes, weighted effect coding was used for the categorical variables (Darlington, 1990). Voice had two levels (*voice*: voiceless = 1, voiced = -0.69), and place of articulation had three levels, corresponding to two contrasts with labial as baseline (*poaCor*: coronal = 1, dorsal = 0, labial = −1.13; *poaDor*: coronal = 0, dorsal = 1, labial = −1.08). As described in the methods, speaking rate was the average word duration per utterance defined by the P2FA boundaries. This predictor was *z*-scored across all talkers ($\mu=242$ ms, $\sigma=59$ ms). Additional binary factors were vowel height (*height*: high [i ɪ u ʊ] = 1, non-high [æ ɛ ei ʌ ɚ a ɔ oʊ oi ai aʊ] = −0.67) and vowel tenseness (*tense*: tense [i ei ʌ a ɔ oʊ u oi ai aʊ] = 1, lax [ɪ ɛ æ ɚ ʊ] = −2.15). Position of the word (utterance position) was coded as one of five categories: initial, medial, final, pre-pausal, or post-pausal. Tokens that were utterance-medial but preceded or followed by a decoded silence were labeled respectively as post-pausal and pre-pausal. For P2FA to decode a segment as silence, the duration of the segment must be at least 30 ms long. Medial position served as the baseline level (*posInit*: initial = 1, medial = −0.10, else = 0; *posFinal*: final = 1, medial = −0.17, else = 0; *posPrePaus*: pre-pausal = 1, medial = −0.05, else = 0; *posPostPaus*: post-pausal = 1, medial = −0.04, else = 0). Number of syllables per word was categorized into three levels: monosyllabic, disyllabic, and polysyllabic (>two syllables), and the monosyllabic level served as baseline (*syllDi*: disyllabic = 1, polysyllabic = 0, monosyllabic = −0.39; *syllPoly*: disyllabic = 0, polysyllabic = 1, monosyllabic = −0.15). Lexical frequency was calculated as the log SUBTLEX frequency (Marian, Bartolotti, Chabal, & Shook, 2012). The dependent variable (VOT) was centered by subtracting the overall mean ($\mu=30$ ms) from each value.

The model also included random effects of talker and word. The random effect structure for talkers included an intercept and slopes for voice, place, speaking rate, and the voice × place interaction. Attempts were made to include additional factors, but this resulted in non-convergence. The random effect of word included an intercept only.

Significant main effects emerged for voice (*voice*: $\beta=24.99$, $t=29.07$) and place (*poaCor*: $\beta=1.95$, $t=2.40$; *poaDor*: $\beta=1.99$, $t=2.55$). The interaction between voice and the first place contrast did not reach significance (*voice* × *poaCor*: $\beta=1.52$, $t=1.74$), but there was a significant interaction between voice and the second place contrast, indicating a smaller difference between the voiced and voiceless dorsals than would have been predicted by voice and place independently (*voice* × *poaDor*: $\beta=-4.00$, $t=-5.00$). These interactions reflect the difference in ranking of places of articulation within each voicing category: while VOT increases with more posterior place among voiced stops, there is little difference in VOT between coronals and dorsals among voiceless stops. Significantly shorter VOTs corresponded to faster speaking rates (*rate:* $\beta=1.40$, $t=16.87$), but this was modulated by a significant interaction between voice and rate (*voice* × *rate*: $\beta=1.20$, $t=16.41$). The effect of rate was augmented for voiceless stops and essentially negated for voiced stops.

[14] The same pattern of significance found for the entire set of talkers was present within the female and male subgroups. Correlations among voiceless stop VOTs ranged from $r=0.80$ to 0.85 for female talkers and from $r=0.74$ to 0.78 for males. Among the voiced stops, correlations ranged from $r=0.25$ to 0.58 for females and from $r=0.36$ to 0.50 for males. Relations between homorganic stops were also similar (female: $r=0.18$ to 0.47; male: $r=0.42$ to 0.47).

Vowel height and tenseness did not reach significance (*height:* β = 1.56, *t* = 1.88; *tense:* β = 0.43, *t* = 0.96); however, there was a significant interaction between height and tenseness indicating that VOT before high tense vowels was significantly longer (β = 1.07, *t* = 2.71). There were significant main effects of utterance position: the VOTs of stops in utterance-initial, pre-pausal, and post-pausal positions were significantly longer than the mean (*posInit*: β = 3.50, *t* = 17.88; *posPrePaus*: β = 0.81, *t* = 3.00; *posPostPaus*: β = 2.78, *t* = 11.84), whereas VOTs of utterance-final stops were significantly shorter (*posFinal*: β = −1.36, *t* = −9.43). The number of syllables was not significant, but the two effects trended in the expected directions. VOTs of disyllabic and polysyllabic words were generally shorter than the VOT of monosyllabic words (*syllDi:* β = −1.15, *t* = −1.04; *syllPoly:* β = −2.72, *t* = −1.39). Finally, higher lexical frequency was associated with a decrease in VOT (β = −1.89, *t* = −2.37).

Analysis of the talker random effects revealed that the intercept and the slope for voice had the largest standard deviations (Table 9). This indicates that the talkers differed most in their overall mean VOT values and in the degree of separation between voiced and voiceless stops. However, the intercept and voice slope were highly correlated (*r* = 0.91), reflecting the fact that the measured means for voiced stops fell between our minimum threshold (4 ms) and the natural voicing category boundary.

3.3. Discussion

The patterns found in the connected speech corpus parallel those of the isolated production study. The mean VOTs of stops were highly correlated, especially within each of the two voicing categories. In addition, there were moderate to strong correlations of talker-specific means and standard deviations for each stop. The magnitudes of the correlations were comparable to those in the previous study, but all reached significance in the connected speech analysis. As the Mixer 6 corpus contained many tokens (n = 88,725) and talkers (n = 180), strong statistical power may have led to an increase in the type I error rate (i.e., false positives). However, this concern was addressed with bootstrap confidence intervals, each of which provides a range of population correlations that is not associated with any null-hypothesis statistical test.

Interestingly, the strength of the correlations in isolated speech increased substantially after correcting for following vowel duration, particularly among the voiced stops and between homorganic stops. No improvement, however, was seen in the connected read speech when either average word duration or following vowel duration were used to estimate speaking rate. Aspects of the isolated speech study, such as the homogeneous repetition of similarly-structured syllables, may have resulted in greater similarity in the realization of stop consonants, and thus stronger correlations after rate correction. Alternatively, it may be that strong correlations are indeed present in connected speech, but harder to estimate given the greater contextual and global variability.[15] In addition, the connected speech study depended on automatic alignment not only of the VOT, but also the individual words in each utterance and the following vowel durations necessary for the speaking rate measurement. Improved precision of these alignments may reveal stronger relations of VOT like those observed in isolated speech after rate correction. These differences notwithstanding, it is striking that a strong pattern of VOT covariation was present among the voiceless stops in spite of the many sources of variation in the connected speech corpus.

Systematic rankings of mean VOT were also observed in both studies, with the notable exception of variation in the talker-specific ranking of [tʰ] and [kʰ]. Specifically, in the connected speech study, there was a strong tendency for talkers to exhibit a slightly greater VOT for [tʰ] in comparison to [kʰ]. Many previous studies have focused on systematic rankings at the population level, and typically report a greater VOT for [kʰ] in comparison to [pʰ], as well as [tʰ]. The present study observed a strong tendency for the ranking of [pʰ] < [kʰ], consistent with previous findings, and little difference between the means of [tʰ] and [kʰ] within or across talkers in both studies. Among the voiced stops, the overwhelming majority of speakers had increasing VOT with more posterior places of articulation ([b] < [d] < [g]). Ordinal rankings are not as informative, however, as linear fits: even consistent ordinal ranking does not entail a linear relation (as any magnitude of separation between VOT means could be consistent with a given ranking), and ordinal rankings are entailed by linear relations (within particular lower and upper limits). In almost all estimated fits between the voiceless stop VOTs, both the additive and scaling factors were significant, indicating that the difference between VOT means varied systematically. The exception in this case was the estimated fit between [tʰ] and [pʰ] in the isolated speech, for which only the scaling factor was significant.

In both studies the mixed-effects linear models revealed greater variation across talkers in the grand mean VOT (intercept) for all six stops and in the degree of separation between voiced and voiceless stops (voice slope). Considerably less variation was observed in the realization of VOT across stop place of articulation. The mixed-effects model also accounted for other important sources of variation in the realization of VOT. In particular, for both isolated and connected speech there were significant effects of speaking rate on the VOT of voiceless stops, and while vowel height and tenseness failed to reach significance individually, a significant interaction was revealed for connected speech, implicating longer VOTs in the context of high tense vowels, [i] and [u] (see also Nearey & Rochet, 1994).

Table 9
Standard deviations of talker random effects in the maximal mixed-effects model of VOT in connected speech.

Random effect for talker	SD
intercept	3.68
voice	4.25
poaCor	1.84
poaDor	1.77
speaking rate	0.74
voice × poaCor	1.45
voice × poaDor	1.62

[15] An additional analysis in which VOT was residualized not only with speaking rate (vowel duration) but also vowel height, vowel tenseness, the interaction between height and tenseness, number of syllables, utterance position of the word, and lexical frequency (described in Section 3.2.3) resulted in significant correlations of talker means across all stop pairs; however, the change in magnitude was less substantial than in the laboratory speech analyses (*r*s = [pʰ–tʰ] 0.83, [tʰ–kʰ] 0.78, [kʰ–pʰ] 0.82, [b–d] 0.25, [d–g] 0.35, [g–b] 0.53, [pʰ–b] 0.29, [tʰ–d] 0.58, [kʰ–g] 0.40, *p*s < 0.006).

In connected speech, utterance position was also a significant factor: compared to the average, stops in utterance-initial, post-pausal, and pre-pausal positions had longer VOTs, whereas stops in utterance-final position had shorter VOTs. The significant VOT lengthening found for utterance-initial stops is consistent with previous findings of domain-initial strengthening at the beginning of the utterance (e.g., Cho & Keating, 2009). VOT tended to be shorter in polysyllabic than in monosyllabic words, however, this effect failed to reach significance. Finally, there was a significant decrease in VOT with higher lexical frequency.

The large-scale analysis implemented here contributes to the understanding of VOT variation and covariation in a speech corpus with a greater number of talkers, larger variety of contextual, prosodic, and lexical factors, and greater amount of data than is typically collected in a laboratory experiment. Despite some measurement error, the automated alignment with P2FA and AutoVOT yielded a pattern that corresponded closely with the findings for isolated speech. Overall, the methods and analyses employed in this section extend our understanding of structured VOT realization to a connected speech style, and more generally advance research in corpus-based phonetics.

4. General discussion

Highly systematic and linear relationships of stop VOT were observed across talkers, particularly within each voicing category, in two quite different speech styles. Multiple methodologies (e.g., correlational structure, ordinal rankings, simple regression, and relative variance in the random effect structure) lend support to the notion that there is *structured* variation in talker mean VOT. Furthermore, these findings hold across isolated and connected speech styles, and preliminary evidence from the Buckeye Corpus (Pitt et al., 2005) indicates that the same pattern of variation also occurs in spontaneous speech (see Section 4.3).

In particular, strong correlations emerged among the voiceless stops, but were also present to a moderate degree among the voiced stops. The present study also corrected for speaking rate and revealed moderately strong correlations among voiced stops and between homorganic stops in isolated speech (cf. Zlatin, 1974; Newman, 2003). The mixed-effects model further corrected for other sources of contextual variation; the random effect structure for the talker revealed that the greatest amount of variability across talkers was explained by identifying the talker's grand mean and the talker-specific difference between the voicing categories. While talkers nonetheless vary along many dimensions (e.g., place of articulation, interaction of voice and place), a large portion of talker variability can be defined in a lower dimensional space. Returning to the points raised in the introduction, the acoustic-phonetic covariation and implied low dimensional variation have strong implications for the structure of the phonetic grammar and for adaptation to novel talkers.

4.1. Implications of VOT covariation for constraints on phonetic systems

The patterns of VOT covariation documented above have two main, and we believe closely connected, theoretical implications. The first implication, which relates to the constraints that restrict phonetic systems, is discussed here. The second implication, concerning how listeners could use implicit knowledge of the covariation pattern to efficiently adapt to novel talkers, is discussed in the following subsection.

Perhaps the most widely invoked constraint on phonetic systems (aside from anatomical limitations on possible speech sounds) is that of *perceptual dispersion* (e.g., Liljencrants & Lindblom, 1972; Lindblom, 1986). The pressure to maintain sufficient perceptual distance between contrasting categories could potentially account for some of our findings. In particular, the VOT covariation of homorganic stops (e.g., [k^h–g]) could be due to dispersion: as the VOT of a voiceless stop becomes shorter across talkers, the VOT of its voiced counterpart could also shorten in order to maintain a clear contrast on this acoustic-phonetic dimension. However, correlations between homorganic stops were much smaller in magnitude than those within the set of voiceless aspirated stops, and often failed to reach significance (see Table 2 and Table 7).

Dispersion could also underlie the VOT covariation of some voiceless stop pairs, such as [p^h] and [k^h], because VOT potentially serves as a secondary perceptual cue for place of articulation (as suggested by Cho & Ladefoged, 1999: 220). Talkers who have relatively long means for [p^h] could ensure that the putative VOT cue for place remains reliable by also having longer means for [k^h]. Critically, however, not all observed correlations among voiceless stops correspond to consistent differences: [t^h] and [k^h] are highly correlated but have similar (and to a certain extent inconsistently ordered) VOT means across talkers; this covariation arguably shows that the two stops are *less dispersed* within each talker than would be expected from contrast preservation alone. Thus a dispersion-theoretic approach to our findings has significant limitations.

Instead, the strong correlations among aspirated stops suggest a principle or constraint of *uniformity*, in the sense of "uniform or parallel behavior of members of a class" (Keating, 2003). The uniformity constraint could in principle apply to articulatory or acoustic targets of phonetic realization. All three stops [p^h t^h k^h] share a feature [+spread glottis] (e.g., Halle & Stevens, 1971) and the associated glottal abduction gesture (e.g., Löfqvist, 1980; Löfqvist & Yoshioka, 1984). An articulatory uniformity constraint would require the glottal targets of these stops to be similar in magnitude, duration, and timing relationship with respect to the oral constriction (e.g., Weismer, 1980; Löfqvist, 1980; Löfqvist & Yoshioka, 1984; Hoole & Pouplier, 2015). Alternatively, the uniformity constraint could apply directly to the acoustic-phonetic results of glottal spreading, requiring similar VOT values within the class of aspirated stops.

This line of thinking reverses the traditional perspective on VOT differences across place of articulation. A large body of research has been devoted to understanding why VOT *varies* across place within an aspirated or unaspirated stop class, with more posterior articulations generally having longer values (e.g., Maddieson, 1997; Cho & Ladefoged, 1999). Phoneticians have been intrigued by such differences, we believe, because of an assumption that laryngeal targets and the resulting VOT values should be the *same* — or at least highly *similar* — across the members of a class within each language. Indeed, Cho and Ladefoged (1999) conclude that "[i]n general, speakers do not deliberately produce different values of the feature VOT across different places of articulation" (225) but that "[s]pecific values

for each place or articulation might be required in the grammar for aspirated stops" (227). The uniformity constraint codifies the assumption of underlying similarity across place of articulation. It need not enforce strict identity of glottal or VOT targets within a stop class, but it does limit their differences. The uniformity constraint helps to shape the phonetic systems of individual speakers, who may differ from one another extensively in absolute but not relative target values within a class.

How broadly the uniformity constraint extends, and how it is best formalized, are important outstanding questions. It is tempting to attribute the correlations among voiced stops, and between homorganic stop pairs, to a phonetic component that all released plosives share: the initial transient and following frication (e.g., Hanson & Stevens, 2003). It is plausible that the target durations for these early portions of stop release are systematic within each talker, though direct evidence on this point would require accurate segmentation of the transient/frication from the following aspiration in voiceless stops. More generally, a constraint favoring "uniform or parallel" phonetic targets within a class should not be limited to stop consonants or to the VOT dimension. It may apply to all or many other categories and dimensions, constraining phonetic realization at the level of individual speakers and of language communities.

There are two straightforward ways of formalizing the uniformity constraint. The first would be essentially identical to the mixed-effects models reported earlier, except the models would now be considered as hierarchical generative descriptions of phonetic targets (e.g., Nielsen & Wilson, 2008; Pajak, Bicknell, & Levy, 2013; Kleinschmidt & Jaeger, 2015). The relative variance of talker-level effects would be established by the constraint: high variance would be allowed for the intercept (talker grand mean) and for the voice effect (reflecting the voicing contrast); but the variance of the place effect would be constrained, ensuring that place differences within a voicing class are limited.

The second formalization would involve reducing VOT means and other measurements to *decorrelated* variables in a lower-dimensional space. This could be accomplished with principal component analysis (PCA) or other methods available for dimensionality reduction (e.g., factor analysis; Murphy, 2012). For example, when PCA is applied to the data from our connected speech study (Section 3), two uncorrelated dimensions or 'components' suffice to account for 90% of the variance in talker means for all six stops. The idea that talker differences can be accurately described with a relatively small number of 'latent' variables is closely related to the central linguistic notions of symmetry and economy, which have been invoked in theories of phoneme inventories (e.g., Maddieson, 1997; Clements, 2003), phonetically-grounded phonology (e.g., Hayes, 1999; Gordon, 2006), and sound change (e.g., Fruehwald, Gress-Wright, & Wallenberg, 2009). Just as symmetric inventories can be derived by combining a small number of independent features, uniformities in phonetic realization can be accounted for by deriving the targets for many sound categories from a small number of talker-specific parameters.

4.2. Implications of VOT covariation for perceptual adaptation

Acoustic-phonetic covariation across speech sounds may also facilitate perceptual adaptation to novel speakers and dialects, and more specifically, perceptual generalization. Consider a scenario in which a listener hears a novel talker producing instances of [p^h] but not of [k^h]. If a listener can estimate the mean VOT of [p^h] directly from exposure to a new talker, prior knowledge about how the means of [p^h] and [k^h] covary may allow the listener to form reasonable expectations about the talker's mean VOT for [k^h]. In essence, the means of stops for which the listener has little talker-specific evidence can be 'read off' the regression lines, as depicted in Figs. 1 and 2. More broadly, prior knowledge of relations among phonetic categories at the talker-specific level allows evidence about the idiosyncratic realization of one category to inform rational expectations about the realizations of other categories.

Evidence that listeners actively predict cross-category VOT has been demonstrated in a variety of studies on perceptual learning and generalization. Theodore and Miller (2010) determined that listeners transfer acoustic-phonetic detail from one place of articulation to another at a talker-specific level. Listeners were trained on two different talkers who differed only in their mean VOT for [p^h]. After exposure, listeners were able to identify that a long [k^h] VOT was more characteristic of the talker with the long [p^h] VOT, and vice versa.

Similarly, listeners generalized a talker's characteristically long VOT from [p^h] to [k^h] in phonetic imitation, without any prior exposure to that talker's [k^h] (Nielsen, 2007, 2011). Interestingly, no effects of imitation were observed for a reduced VOT; however, in these cases, imitation may have been inhibited by the natural lower VOT boundary for voiceless stop consonants (see also Clarke & Luce, 2005). Finally, VOT generalization has also been observed in Eimas and Corbit (1973) in selective adaptation and Kraljic and Samuel (2006) with lexically-induced perceptual learning.

In an analysis of the VOT generalization found in Nielsen (2007), Nielsen and Wilson (2008) proposed a Bayesian model that implicitly encodes VOT covariation between [p^h] and [k^h] through the laryngeal and place feature values. The original model was designed to predict effects of phonetic imitation, but implicit in imitation is talker-specific learning. The model thus adapts to the talker by inferring mean VOT values from the linear combination of the spread glottis and dorsal features. The effect of spread glottis is shared by both [p^h] and [k^h], and the dorsal feature is always 0 for [p^h] and positive for [k^h]. The positive offset from the dorsal feature maintains the ordinal relationship between [p^h] and [k^h], while uniformity in VOT is expressed in the shared spread glottis effect.

To account for differences among talkers, other models have used extrinsic relationships among speech sounds that assume structured variation. Extrinsic normalization techniques generally apply uniform transformations, such as mean subtraction or *z*-scoring, to all members of a class of speech sounds (e.g., vowels or fricatives; Gerstman, 1968; Lobanov, 1971; Nearey, 1978; McMurray & Jongman, 2011). While many of these approaches perform reasonably well in off-line talker normalization, it is unclear how they could be applied in on-line speech perception. Taken literally, these methods imply that a listener would have to hear all members of a class of speech sounds before any adaptation can occur.

In contrast, the present findings together with evidence from perceptual generalization support the idea that listeners may rapidly employ prior perceptual knowledge of talker covariation.

Joint estimation of talker parameters for many speech sounds could proceed more efficiently (and with greater precision). Information about covariation and linear relations has already proved fruitful in on-line automatic speaker adaptation (e.g., Lasry & Stern, 1984; Cox, 1995; Zavaliagkos, Schwartz, & Makhoul, 1995) and has been incorporated to a certain extent in other cognitive models of talker adaptation and perceptual generalization similar to that of Nielsen and Wilson (2008) (e.g., McMurray & Jongman, 2011; Pajak et al., 2013).

In addition to covariation among category means, we also identified strong correlations between talker-specific VOT means and standard deviations (equivalently, variances) within categories. Knowledge of such relations could be used in perceptual adaptation: listeners could infer a novel talker's mean and variance jointly (rather than independently). However, previous work indicates that prior expectations about the relation between category means and variances can be overridden with sufficient evidence (e.g., Clayards, Tanenhaus, Aslin, & Jacobs, 2008). Furthermore, perceptual adaptation to a new talker can in some cases be modeled by recalibration of either the mean or the variance of a category (Kleinschmidt & Jaeger, 2015). Additional research is clearly needed to investigate whether listeners exploit knowledge of typical mean-variance relations for VOT and other acoustic-phonetic properties.

Finally, adaptation to talker-specific VOT means will be possible and beneficial only if within-talker variation on this dimension is highly structured. If listeners could not reliably 'factor out' the effects of speaking rate, prosodic context, lexical frequency, and other predictors (e.g., because these effects were highly variable across talkers), they could not accurately estimate talker means during adaptation or generate useful expectations during further processing. However, the perceptual experiments discussed above indicate that listeners do become attuned to talker-specific VOT parameters (e.g., Clayards et al., 2008; Nielsen, 2011; Kleinschmidt, Weatherholtz, & Jaeger, submitted). Furthermore, in our mixed-effects analysis there were diminishing returns for talker-specific random effects beyond the intercept and voice slope, suggesting that additional sources of variability have relatively constant effects across speakers. These findings lend support to models in which adapting to a novel talker mainly involves estimating a relatively small number of parameters, such as the mean and variance along relevant acoustic-phonetic dimensions (e.g., Nearey, 1978; McMurray & Jongman, 2011).

4.3. Future directions

One clear extension of our study would be to carefully investigate VOT covariation in spontaneous speech. As a preliminary step, we used AutoVOT to extract measurements for all of the word-initial prevocalic stops of 38 talkers from the Buckeye corpus (Pitt et al., 2005). Unlike the Mixer 6 corpus, the content of the Buckeye corpus is not matched across talkers. In spite of the much greater variation in prosodic, lexical, and syntactic contexts, talker means were again found to be highly correlated after removal of outliers across all voiceless stop pairs and between [b] and [g] (e.g., [p^h–t^h]: $r = 0.82$, [t^h–k^h]: $r = 0.83$, [k^h–p^h]: $r = 0.81$, $ps < 0.006$; [b–g]: $r = 0.43$, $p < 0.01$). While further examination of the patterns in this and other spontaneous speech corpora is certainly warranted, we tentatively conclude that strong correlations, at least for aspirated stops, will be found in essentially any speech style.

The systematic relations observed among stop categories may be present not only for VOT, but also for other acoustic-phonetic cues to stop consonant voice and place. Research is currently underway to investigate talker systematicity in stop consonant spectral COG (Blumstein & Stevens, 1979; Chodroff & Wilson, 2014), f0 (Haggard, Ambler, & Callow, 1970; Ohde, 1984; Whalen, Abramson, Lisker, & Mody, 1990; Kong & Edwards, 2016), relative amplitude (Repp, 1979; Ohde & Stevens, 1983), and following vowel duration (Summerfield, 1981; Allen & Miller, 1999). Systematicity in closure duration and prevoicing, and their respective relations to positive VOT, also warrant further investigation. Additional research is necessary to determine whether these relations exist for acoustic-phonetic cues among other natural class such as fricatives, nasals, and liquids.

Structured VOT variation could potentially be one reflection of talker differences in domain-initial strengthening (Fougeron & Keating, 1997; Cho & Keating, 2001) or other types of hyperarticulation (e.g., Lindblom, 1990). If talkers vary in the degree of strengthening due to prosodic boundaries, and the effect of strengthening on VOT is similar for all stops that have the same laryngeal specification, the correlations observed here would be predicted. Note that this analysis crucially assumes a uniformity constraint similar to that discussed in Section 4.1 (i.e., talker-specific prosodic effects would have to apply uniformly to all stops within each voicing category). Talker-specific VOT values would then reflect the talker's degree of hyperarticulation and be expected to correlate with other measures of domain-initial strengthening (see Bang & Clayards, 2016 for related research). In this way, a small number of prosodic (or hyperarticulation) variables would account for many idiosyncratic aspects of a talker's phonetic system. Listeners could then adapt to a talker by estimating these higher-level variables, jointly inferring the means and other parameters of many phonetic categories along multiple dimensions.

While this study focuses only on American English, similar patterns for VOT and other acoustic-phonetic dimensions may also hold among talkers of other languages. For VOT in particular, the strength of talker covariation could depend on the presence of a laryngeal contrast among stops, the phonetic realization of each voicing category (e.g., voiceless aspirated vs. voiceless unaspirated vs. phonetically voiced), and the particular language (e.g., strong for AE aspirated stops, weak for Navajo aspirated stops). The nature of talker covariation across languages may shed further light on universal and language-specific aspects of the phonetic grammar.

Finally, studies of perceptual adaptation provide support for knowledge of VOT covariation, but have examined generalization only after considerable exposure to a new talker. Yet, perceptual knowledge of extrinsic relations would seemingly be most beneficial in early stages of adaptation when talker-specific evidence is minimal. These limitations warrant further investigation with regards to the cognitive status of VOT covariation, and of phonetic covariation more generally.

5. Conclusion

With converging evidence from multiple statistical methods, we have established highly systematic relationships of VOT distributions across AE talkers. While previous studies have largely focused on ordinal rankings, the present study extends and strengthens our understanding of VOT variation by demonstrating linear relelations among stop categories. Many of these relations involve both additive and scalar factors, indicating that there is not necessarily a constant difference between stop means across talkers (cf. Theodore et al., 2009). Finally, the findings from isolated speech generalize well to connected speech data from more than one hundred talkers. Structured VOT variation provides support for a uniformity constraint that restricts the phonetic systems of individual speakers and that listeners could use to generalize from limited experience with a new talker. The VOT pattern documented here may be one part of a much larger system of covariation that encompasses multiple phonetic dimensions and sound classes.

Acknowledgments

The authors would like to thank Alessandra Golden, Chloe Haviland, Spandana Mandaloju, and Benjamin Wang for assistance in data processing. Thanks also go to Emily Atkinson, Meghan Clayards, Lisa Davidson, Florian Jaeger, Christo Kirov, Paul Smolensky, Morgan Sonderegger, Douglas Whalen, and Mackenzie Young for helpful comments and discussion. Finally, we thank Jack Godfrey and Sanjeev Khudanpur for making this project possible. This research was partially supported by a JHU Distinguished Science of Learning Pre-Doctoral Fellowship, the Dolores Zohrab Liebmann Fund, and the DHS-USSS Forensic Services Division.

References

Allen, J. S., & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *The Journal of the Acoustical Society of America*, *106*(4), 2031–2039, http://dx.doi.org/10.1121/1.427949.

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544–552 http://dx.doi.org/10.1121/1.1528172.

Assmann, P. F., Nearey, T. M., & Bharadwaj, S. (2008). Analysis of a vowel database. *Canadian Acoustics*, *36*(3), 148–149.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baese-Berk, M., & Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and Cognitive Processes*, *24*(4), 527–554, http://dx.doi.org/10.1080/01690960802299378.

Bang, H.-Y., & Clayards, M. (2016). Structured variation across sound contrasts, talkers, and speech styles. Poster presented at LabPhon 15: Speech Dynamics and Phonological Representation. Ithaca, NY.

Beckman, J., Jessen, M., & Ringen, C. (2013). Empirical evidence for laryngeal features: Aspirating vs. true voice languages. *Journal of Linguistics*, *49*(2), 259–284, http://dx.doi.org/10.1017/S0022226712000424.

Benjamin, B. J. (1982). Phonological performance in gerontological speech. *Journal of Psycholinguistic Research*, *11*(2), 159–167, http://dx.doi.org/10.1007/BF010682.

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, *66*(4), 1001–1017.

Boersma, P. & Weenink, D. (2015). Praat: Doing Phonetics by Computer [Computer program]. Version 6.0.05, retrieved 06 November 2015 from ⟨http://www.praat.org/⟩.

Brandschain, L., Graff, D., Cieri, C., Walker, K., & Caruso, C. (2010). The Mixer 6 corpus: Resources for cross-channel and text independent speaker recognition. *In Proceedings of the seventh international conference on language resources and evaluation (LREC 2010)*, 2441–2444. Malta.

Brandschain, L., Graff D., and Walker K. (2013). *Mixer 6 Speech LDC2013S03*. Hard Drive. Philadelphia: Linguistic Data, Consortium.

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language*, *89*, 68–86.

Bürki, A., Ernestus, M., Gendrot, C., Fougeron, C., & Frauenfelder, U. H. (2011). What affects the presence versus absence of schwa and its duration: a corpus analysis of French connected speech. *The Journal of the Acoustical Society of America*, *130*(6), 3980–3991, http://dx.doi.org/10.1121/1.3658386.

Byrd, D. (1992). Preliminary results on speaker-dependent variation in the TIMIT database. *The Journal of the Acoustical Society of America*, *92*(1), 593–596, http://dx.doi.org/10.1121/1.404271.

Byrd, D. (1993). 54,000 American stops. *UCLA Working Papers in Phonetics*, *83*, 97–116.

Byrd, D., & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics*, *26*, 173–199, http://dx.doi.org/10.1006/jpho.1998.0071.

Cho, T., & Keating, P. A. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics*, *29*, 155–190, http://dx.doi.org/10.1006/jpho.2001.0131.

Cho, T., & Keating, P. A. (2009). Effects of initial position versus prominence in English. *Journal of Phonetics*, *37*, 466–485, http://dx.doi.org/10.1016/j.wocn.2009.08.001.

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics*, *27*, 207–229, http://dx.doi.org/10.1006/jpho.1999.0094.

Chodroff, E., Maciejewski, M., Trmal, J., Khudanpur, S., & Godfrey, J. (2016). New release of Mixer-6: improved validity for phonetic study of speaker variation and identification. *In Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*, 1323–1327. Portorož, Slovenia.

Chodroff, E., & Wilson, C. (2014). Burst spectrum as a cue for the stop voicing contrast in American English. *The Journal of the Acoustical Society of America*, *136*(5), 2762–2772, http://dx.doi.org/10.1121/1.4896470.

Chung, H., Kong, E. J., Edwards, J., Weismer, G., Fourakis, M., & Hwang, Y. (2012). Cross-linguistic studies of children's and adults' vowel spaces. *The Journal of the Acoustical Society of America*, *131*(1), 442–454, http://dx.doi.org/10.1121/1.3651823.

Clarke, C.M., & Luce, P.A. (2005). Perceptual adaptation to speaker characteristics: VOT boundaries in stop voicing categorization. *In* Hazan, V., & Iverson, P. (Eds.), *Proceedings of ISCA workshop on plasticity in speech perception*, 23–26. London, UK.

Clayards, M.A. (in press). Individual talker and token variability in multiple cues to stop voicing. *Phonetica*.

Clayards, M. A., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*, 804–809, http://dx.doi.org/10.1016/j.cognition.2008.04.004.

Clements, G.N. (2003). Feature economy as a phonological universal. *In* Solé, M., Recasens, D., & Romero, J. (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences*, 371–374. Barcelona, Spain.

Cole, J.S., Choi, H., Kim, H., & Hasegawa-Johnson, M. (2003). The effect of accent on the acoustic cues to stop voicing in Radio News speech. *In* Solé, M., Recasens, D., & Romero, J. (Eds.), *Proceedings of the 15th international congress of phonetic sciences*, 15–18. Barcelona, Spain.

Cole, J. S., Kim, H., Choi, H., & Hasegawa-Johnson, M. (2007). Prosodic effects on acoustic cues to stop voicing and place of articulation: evidence from Radio News speech. *Journal of Phonetics*, *35*, 180–209, http://dx.doi.org/10.1016/j.wocn.2006.03.004.

Cox, S. (1995). Predictive speaker adaptation in speech recognition. *Computer Speech and Language*, *9*, 1–17, http://dx.doi.org/10.1006/csla.1995.0001.

Darlington, R. B. (1990). In J. D. Anker, & B. Boylan (Eds.), *Regression and linear models*. New York: McGraw-Hill Publishing Company.

Das, S., & Hansen, J.H.L. (2004). Detection of voice onset time (VOT) for unvoiced stops (/p/, /t/, /k/) using the Teager energy operator (TEO) for automatic detection of accented English. *In* Tanskanen, J. M. A. (Ed.) *Proceedings of the 6th Nordic signal processing symposium*, 344–347. Espoo, Finland.

Davidson, L. (2016). Variability in the implementation of voicing in American English obstruents. *Journal of Phonetics*, *54*, 35–50, http://dx.doi.org/10.1016/j.wocn.2015.09.003.

De Cara, B., & Goswami, U. (2002). Similarity relations among spoken words: the special status of rimes in English. *Behavior Research Methods, Instruments, Computers*, *34*(3), 416–423, http://dx.doi.org/10.3758/BF03195470.

DiCanio, C. T., Nam, H., Amith, J. D., García, R. C., & Whalen, D. H. (2015). Vowel variability in elicited versus spontaneous speech: evidence from Mixtec. *Journal of Phonetics*, *48*, 45–59, http://dx.doi.org/10.1016/j.wocn.2014.10.003.

Disner, S. F. (1983). *Vowel quality: the relation between universal and language-specific factors (Ph.D dissertation)*. UCLA.

Dmitrieva, O., Llanos, F., Shultz, A. A., & Francis, A. L. (2015). Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English. *Journal of Phonetics*, *49*, 77–95, http://dx.doi.org/10.1016/j.wocn.2014.12.005.

Docherty, G. J. (1992). *The timing of voicing in British English obstruents*. Berlin: Walter de Gruyter.

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, *4*, 99–109.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185, http://dx.doi.org/10.2307/2289144.

Elvin, J., & Escudero, P. (2014). Comparing acoustic analyses of Australian English vowels from Sydney: Cox (2006) versus AusTalk. *In Proceedings of the international symposium on the acquisition of second language speech. Concordia working papers in applied linguistics*, 145–156.

Escudero, P., Boersma, P., Rauber, A. S., & Bion, R. A. H. (2009). A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *The Journal of the Acoustical Society of America*, *126*(3), 1379–1393, http://dx.doi.org/10.1121/1.3180321.

Evanini, K., Isard, S., & Liberman, M. (2009). Automatic formant extraction for sociolinguistic analysis of large corpora. *In Proceedings of INTERSPEECH*, 1655–1658. Brighton, UK.

Fisher-Jorgensen, E. (1954). Acoustic analysis of stop consonants. *Miscellanea Phonetica*, *2*, 42–59.

Flege, J. E., Frieda, E. M., Walley, A. C., & Randazza, L. A. (1998). Lexical factors and segmental accuracy in second language speech production. *Studies in Second Language Acquisition*, *20*(2), 155–187.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, *101*(6), 3728–3740, http://dx.doi.org/10.1121/1.418332.

Fruehwald, J. (2013). *The phonological influence on phonetic change (Ph.D dissertation)*. University of Pennsylvania.

Fruehwald, J., Gress-Wright, J., & Wallenberg, J.C. (2009). Phonological rule change: the constant rate effect. *In* Kan, S., Moore-Cantwell, C., Staubs R.(Eds.) *Proceedings of the 40th annual meeting of the north east linguistic society*, 1–12. Cambridge, MA. http://doi.org/10.1017/CBO9781107415324.004.

Furui, S. (1980). A training procedure for isolated word recognition systems. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(2), 129–136.

Garofolo, J. S., Lamel, L., Fisher, W. M., Fiscus, J. G., Pallett, D. S. D., & Dahlgren, N. L. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. Philadelphia: Linguistic Data Consortium.

Gendrot, C., & Adda-Decker, M. (2005). Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German. *In Proceedings of INTERSPEECH*, 2453–2456. Lisbon, Portugal.

Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics*, *16*(1), 78–89, http://dx.doi.org/10.1109/TAU.1968.1161953.

Godfrey, J.J., Holliman, E.C., & McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. *In Proceedings of the IEEE conference on acoustics, speech, and signal processing,* 517–520. http://doi.org/10.1109/ICASSP.1992.225858.

Gordon, M. (2006). *Syllable weight: phonetics, phonology, typology*. New York: Routledge.

Gordon, M., Barthmaier, P., & Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association*, *32*, 141–174, http://dx.doi.org/10.1017/S0025100302001020.

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-Aligner: a tool for forced alignment of laboratory speech. *Canadian Acoustics*, *39*(3), 192–193.

Haggard, M., Ambler, S., & Callow, M. (1970). Pitch as a voicing cue. *The Journal of the Acoustical Society of America*, *47*(2B), 613–617.

Halle, M. & Stevens, K. N. (1971). A note on laryngeal features. *MIT RLE quarterly progress report* 10, 198-218.

Hanson, H.M., & Stevens, K.N. (2003). Models of aspirated stops in English. *In* Solé, M., Recasens, D., & Romero, J. (Eds.), *Proceedings of the 15th international congress of phonetic sciences*, 783–786. Barcelona, Spain.

Harnsberger, J. D. (2000). A cross-language study of the identification of non-native nasal consonants varying in place of articulation. *The Journal of the Acoustical Society of America*, *108*(2), 764–783, http://dx.doi.org/10.1121/1.429610.

Hasegawa-Johnson, M., Chen, K., Cole, J. S., Borys, S., Kim, S. S., Cohen, A., Zhang, T., Choi, J. Y., Kim, H., Yoon, T., & Chavarria, S. (2005). Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus. *Speech Communication, 46*(3–4), 418–439, http://dx.doi.org/10.1016/j.specom.2005.01.009.

Hayes, B. (1999). Phonetically-driven phonology: the role of Optimality Theory and inductive grounding. In M. Darnell, E. Moravcsik, M. Noonan, F. Newmeyer, & K. Wheatley (Eds.), *Functionalism and formalism in linguistics (Vol. 1): general papers* (pp. 243–285). Amsterdam: John Benjamins.

Hoit, J. D., Solomon, N. P., & Hixon, T. J. (1993). Effect of lung volume on voice onset time (VOT). *Journal of Speech and Hearing Research*, *36*, 516–521, http://dx.doi.org/10.1044/jshr.3603.516.

Hoole, P. (1997). *Techniques for investigating laryngeal articulation and the voice-source*, 35. Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM).

Hoole, P., & Pouplier, M. (2015). Interarticulatory coordination. In M. Redford (Ed.), *The handbook of speech production* (pp. 131–157). Somerset, MA: Wiley.

Hunnicutt, L., & Morris, P. (2016). Pre-voicing and aspiration in Southern American English. *In University of Pennsylvania working papers in linguistics (Vol. 22)*, 215–224. http://doi.org/10.1017/CBO9781107415324.004.

Jacewicz, E., Fox, R. A., & Salmons, J. (2007). Vowel duration in three American English dialects. *American Speech*, *82*(4), 367–385, http://dx.doi.org/10.1215/00031283-2007-024.

Jacewicz, E., Fox, R. A., & Lyle, S. (2009). Variation in stop consonant voicing in two regional varieties of American English. *Journal of the International Phonetic Association*, *39*(3), 313–334, http://dx.doi.org/10.1017/S0025100309990156.

Joos, M. (1948). Acoustic phonetics. *Language*, *24*(2), 5–136.

Keating, P. A. (1985). Universal phonetics and the organization of grammars. In Victoria Fromkin (Ed.), *Phonetic linguistics: essays in honor of Peter Ladefoged* (pp. 115–132). Orlando: Academic Press.

Keating, P. A., Byrd, D., Flemming, E. S., & Todaka, Y. (1994). Phonetic analyses of word and segment variation using the TIMIT corpus of American English. *Speech Communication*, *14*, 131–142.

Keating, P.A. (2003). Phonetic and other influences on voicing contrasts. *In* Solé, M., Recasens, D., & Romero, J. (Eds.), *Proceedings of the 15th international congress of phonetic sciences*, 20–23. Barcelona, Spain.

Kendall, T. (2009). *Speech rate, pause and linguistic variation: an examination through the sociolinguistic archive and analysis project (Ph.D dissertation)*. Duke University.

Kessinger, R. H., & Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics*, *25*, 143–168, http://dx.doi.org/10.1006/jpho.1996.0039.

Kessinger, R. H., & Blumstein, S. E. (1998). Effects of speaking rate on voice-onset time and vowel production: some implications for perception studies. *Journal of Phonetics*, *26*, 117–128, http://dx.doi.org/10.1006/jpho.1997.0069.

Kingston, J. (1992). The phonetics and phonology of perceptually motivated articulatory covariation. *Language and Speech*, *35*(1, 2), 99–113.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, *70*(3), 419–454, http://dx.doi.org/10.1353/lan.1994.0023.

Kirby, J.P., & Ladd, D.R. (2015). Stop voicing and f0 perturbations: evidence from French and Italian. *In* The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th international congress of phonetic sciences*. Paper number 0740. Glasgow, UK.

Kirov, C., & Wilson, C. (2012). The specificity of online variation in speech production. *In* Miyake, N., Peebles, D., & Cooper, R. P. (Eds.), *Proceedings of the 34th annual conference of the cognitive science society*, 587–592. Sapporo, Japan.

Klatt, D. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research*, *18*(4), 686–706.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognizing the familiar, generalizing to the similar, and adapting to the novel. *Psychological Review*, *122*(2), 148–203.

Kleinschmidt, D.F., & Jaeger, T.F.(submitted). Inferring listeners' beliefs about unfamiliar talkers. Unpublished manuscript.

Kleinschmidt, D.F., Weatherholtz, K., & Jaeger, T.F.(submitted). Sociolinguistic perception as inference under uncertainty. Unpublished manuscript.

Koenig, L. L. (2000). Laryngeal factors in voiceless consonant production in men, women, and 5-year-olds. *Journal of Speech, Language and Hearing Research*, *43*(5), 1211–1228.

Kong, E. J. (2009). *The development of phonation-type contrasts in plosives: cross-linguistic perspectives (Ph.D dissertation)*. The Ohio State University.

Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: cue weighting and executive function. *Journal of Phonetics*, *59*, 40–57 http://dx.doi.org/10.1016/j.wocn.2016.08.006.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin Review*, *13*(2), 262–268, http://dx.doi.org/10.3758/BF03193841.

Kuhl, P. K. (1981). Discrimination of speech by nonhuman animals: basic auditory sensitivities conducive to the perception of speech-sound categories. *The Journal of the Acoustical Society of America*, *70*(2), 340–349, http://dx.doi.org/10.1121/1.386782.

Kurath, H., & McDavid, R. I. (1961). *The pronunciation of English in the Atlantic States: Based upon the collections of the linguistic atlas of the Eastern United States*. Ann Arbor: University of Michigan Press.

Labov, W., Rosenfelder, I., & Fruehwald, J. (2013). One hundred years of sound change in Philadelphia: linear incrementation, reversal, and reanalysis. *Language*, *89*(1), 30–65.

Labov, W., Yaeger, M., & Steiner, R. (1972). *Quantitative study of sound change in progress*. Report on NSF project No. 65-3287.

Lasry, M. J., & Stern, R. M. (1984). A posteriori estimation of correlated jointly Gaussian mean vectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *4* (PAMI-6), 530–535.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language*, *48*(4), 839–862.

Lindau, M., & Wood, P. (1977). Vowel features. *UCLA Working Papers in Phonetics*, *38*, 41–48.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J. J. Ohala, & J. Jaeger (Eds.), *Experimental phonology* (pp. 13–44). Orlando: Academic Press.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W. J. Hardcastle, & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). The Netherlands: Springer Netherlands.

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

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, *49*(2–2), 606–608.

Löfqvist, A. (1980). Interarticulator programming in stop production. *Journal of Phonetics*, *8*, 475–490.

Löfqvist, A., & Yoshioka, H. (1984). Intrasegmental timing: laryngeal-oral coordination in voiceless consonant production. *Speech Communication*, *3*, 279–289.

Maddieson, I. (1997). Phonetic universals. In J. Laver, & W. J. Hardcastle (Eds.), *Handbook of phonetic sciences* (pp. 619–639). Oxford: Blackwells Publishers.

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PloS One*, *7*(8), e43230, http://dx.doi.org/10.1371/journal.pone.0043230.

McDonough, J. & Ladefoged, P. (1993). Navajo stops. *UCLA Working Papers in Phonetics, 84,* 151–164.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219–246, http://dx.doi.org/10.1037/a0022325.

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: a look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, *43*, 106–115.

Möbius, B. (2004). Corpus-based investigations on the phonetics of consonant voicing. *Folia Linguistica*, *38*(1–2), 5–26.

Morris, R. J., & Brown, W. S. (1994). Age-related differences in speech variability among women. *Journal of Communication Disorders*, *27*(1), 49–64, http://dx.doi.org/10.1016/0021-9924(94)90010-8.

Munson, B., McDonald, E. C., DeBoe, N. L., & White, A. R. (2006). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics*, *34*, 202–240, http://dx.doi.org/10.1016/j.wocn.2005.05.003.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. Cambridge: MIT press.

Nartey, J. N. A. (1982). *On fricative phones and phonemes: measuring the phonetic differences within and between languages (Ph.D dissertation)*. UCLA.

Nearey, T. M. (1978). *Phonetic feature systems for vowels (Ph.D dissertation)*. Indiana University.
Nearey, T. M., & Assmann, P. F. (2007). Probabilistic 'sliding template' models for indirect vowel normalization. In M.-J. Solé, P. S. Beddor, & M. Ohala (Eds.), *Experimental approaches to phonology* (pp. 246–270). New York: Oxford University Press.
Nearey, T. M., & Rochet, B. L. (1994). Effects of place of articulation and vowel context on VOT production and perception for French and English stops. *Journal of the International Phonetic Association*, *24*(1), 1–18, http://dx.doi.org/10.1017/S0025100300004965.
Newman, R. S. (2003). Using links between speech perception and speech production to evaluate different acoustic metrics: a preliminary report. *The Journal of the Acoustical Society of America*, *113*(5), 2850–2860, http://dx.doi.org/10.1121/1.1567280.
Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3), 1181–1196, http://dx.doi.org/10.1121/1.1348009.
Nielsen, K.Y. (2007). Implicit phonetic imitation is constrained by phonemic contrast. *In Proceedings of the 16th international congress of phonetic sciences*, 1961–1964. Saarbrücken, Germany.
Nielsen, K. Y. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*, 132–142, http://dx.doi.org/10.1016/j.wocn.2010.12.007.
Nielsen, K. Y., & Wilson, C. (2008). A hierarchical Bayesian model of multi-level phonetic imitation. In N. Abner, & J. Bishop (Eds.), *Proceedings of the 27th west coast conference on formal linguistics* (pp. 335–343). Los Angeles: Cascadilla Proceedings Project.
Ohde, R. N. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *The Journal of Acoustical Society of America*, *75*(1), 224–230.
Ohde, R. N., & Stevens, K. N. (1983). Effect of burst amplitude on the perception of stop consonant place of articulation. *The Journal of the Acoustical Society of America*, *74* (3), 706–714, http://dx.doi.org/10.1121/1.389856.
Ostendorf, M. (2001). A prosodically labeled database of spontaneous speech. *In ISCA tutorial and research workshop (ITRW) on prosody in speech recognition and understanding*, 5–7. Red Bank, New Jersey.
Pajak, B., Bicknell, K., & Levy, R. (2013). A model of generalization in distributional learning of phonetic categories. *In* Demberg, V.,Levy, R. (Eds.), *Proceedings of the 4th workshop on cognitive modeling and computational linguistics*, 11–20. Sofia, Bulgaria.
Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). LibriSpeech: an ASR corpus based on public domain audio books. *In Proceedings of the IEEE conference on acoustics, speech and signal processing (ICASSP)*, 5206–5210.
Paul, D.B., & Baker, J.M. (1992). The design for the Wall Street Journal-based CSR Corpus. *In Proceedings of the DARPA speech and natural language workshop*, 357–362.
Peirce, J. W. (2007). PsychoPy – Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13, http://dx.doi.org/10.1016/j.jneumeth.2006.11.017.
Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.
Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of Acoustical Society of America*, *32*(6), 693–703.
Pierrehumbert, J. B., & Talkin, D. (1992). Lenition of /h/ and glottal stop. In G. Docherty, & D. R. Ladd (Eds.), *Papers in laboratory phonology II: gesture, segment, prosody* (pp. 90–117). Cambridge: Cambridge University Press.
Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95, http://dx.doi.org/10.1016/j.specom.2004.09.001.
Port, R. F., & Rotunno, R. (1979). Relation between voice-onset time and vowel duration. *The Journal of the Acoustical Society of America*, *66*(3), 654–682, http://dx.doi.org/10.1121/1.383692.
Repp, B. H. (1979). Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language and Speech*, *22*(2), 173–189.
Rosenfelder, I., Fruehwald, J., Evanini, K., & Yuan, J. (2011). *FAVE (Forced Alignment and Vowel Extraction) Program Suite*. ⟨http://fave.ling.upenn.edu⟩.
Schmidt, R. A., Zelaznik, H., Hawkins, B., Frank, J. S., & Quinn, J. T. (1979). Motor-output variability: a theory for the accuracy of rapid motor acts. *Psychological Review*, *86*(5), 415–451, http://dx.doi.org/10.1037/0033-295X.86.5.415.
Schöner, G. (2002). Timing, clocks, and dynamical systems. *Brain and Cognition*, *48*(1), 31–51, http://dx.doi.org/10.1006/brcg.2001.1302.
Schuppler, B., Ernestus, M., Scharenborg, O., & Boves, L. (2011). Acoustic reduction in conversational Dutch: a quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics*, *39*, 96–109, http://dx.doi.org/10.1016/j.wocn.2010.11.006.
Scobbie, J.M. (2005). Interspeaker variation as the long term outcome of dialectally varied input: speech production evidence for fine-grained plasticity. *In* Hazan, V., & Iverson, P. (Eds.), *ISCA workshop on plasticity in speech perception*, 56–59. London.
Scobbie, J. M. (2006). Flexibility in the face of incompatible English VOT systems. In L. Goldstein, D. H. Whalen, & C. T. Best (Eds.), *Laboratory phonology 8: varieties of phonological competence (phonology and phonetics, Vol. 4)*. New Haven, Conn.
Shaw, J. A., Gafos, A. I., Hoole, P., & Zeroual, C. (2009). Syllabification in Moroccan Arabic: evidence from patterns of temporal stability. *Phonology*, *26*(1), 187–215, http://dx.doi.org/10.1017/S0952675709001754.
Shultz, A. A., Francis, A. L., & Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *The Journal of the Acoustical Society of America*, *132*(2), EL95, http://dx.doi.org/10.1121/1.4736711.
Smiljanić, R., & Bradlow, A. R. (2008). Stability of temporal contrasts across speaking styles in English and Croatian. *Journal of Phonetics*, *36*, 91–113, http://dx.doi.org/10.1016/j.wocn.2007.02.002.
Smith, B. L. (1978). Effects of place of articulation and vowel environment on "voiced" stop consonant production. *Glossa*, *12*(2), 163–173.
Solé, M.-J. (2007). Controlled and mechanical properties in speech. In M.-J. Solé, P. S. Beddor, & M. Ohala (Eds.), *Experimental approaches to phonology* (pp. 302–321). Oxford: Oxford University Press.
Solé, M.-J., & Estebas, E. (2000). Phonetic and phonological phenomena: V.O.T. A cross-language comparison. *In Proceedings of the XVII AEDEAN* conference, 437–444. Vigo, Spain.
Sonderegger, M. (2015). Trajectories of voice onset time in spontaneous speech on reality TV. *In* The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th international congress of phonetic sciences*. Paper number 0903. Glasgow, UK.
Sonderegger, M., & Keshet, J. (2010). Automatic discriminative measurement of voice onset time. *In* Kobayashi, T., Hirose, K., & Nakamura, S. (Eds.), *Proceedings of INTERSPEECH*, 2242–2245. Makuhari, Japan.
Sonderegger, M., & Keshet, J. (2012). Automatic discriminative measurement of voice onset time. *The Journal of the Acoustical Society of America*, *132*(6), 3965–3979, http://dx.doi.org/10.1121/1.4763995.
Stuart-Smith, J., Rathcke, T., Sonderegger, M., & Macdonald, R. (2015). A real-time study of plosives in Glaswegian using an automatic measurement algorithm. *In Language variation-European perspectives V: selected papers from the seventh international conference on language variation in Europe (ICLaVE 7)*, *17*, 225–237.
Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology Human Perception and Performance*, *7*(5), 1074–1095, http://dx.doi.org/10.1037/0096-1523.7.5.1074.
Suomi, K. (1980). *Voicing in English and Finnish stops: a typological comparison with an inter language study of the two languages in contact (Ph.D Dissertation)*. University of Turku.
Swartz, B. L. (1992). Gender difference in voice onset time. *Perceptual and Motor Skills*, *75*(3), 983–992.
Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *The Journal of the Acoustical Society of America*, *128*(4), 2090–2099, http://dx.doi.org/10.1121/1.3467771.
Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: contextual influences. *The Journal of the Acoustical Society of America*, *125*(6), 3974–3982, http://dx.doi.org/10.1121/1.3106131.
Torre, P., & Barlow, J. A. (2009). Age-related changes in acoustic characteristics of adult speech. *Journal of Communication Disorders*, *42*(5), 324–333, http://dx.doi.org/10.1016/j.jcomdis.2009.03.001.
Torreira, F., & Ernestus, M. (2012). Weakening of intervocalic /s/ in the Nijmegen Corpus of Casual Spanish. *Phonetica*, *69*(3), 124–148, http://dx.doi.org/10.1159/000.
Turk, A., & Shattuck-Hufnagel, S. (2014). Timing in talking: what is it used for, and how is it controlled? *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, *369*(1658), 20130395, http://dx.doi.org/10.1098/rstb.2013.0395.
Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: influence of place of articulation and speaking rate on the internal structure of voicing categories. *The Journal of the Acoustical Society of America*, *92*(2), 723–735, http://dx.doi.org/10.1121/1.403997.
Weismer, G. (1979). Sensitivity of voice-onset time (VOT) measures to certain segmental features in speech production. *Journal of Phonetics*, *7*, 197–204.
Weismer, G. (1980). Control of the voicing distinction for intervocalic stops and fricatives: some data and theoretical considerations. *Journal of Phonetics*, *8*, 427–438.
Whalen, D. H., Abramson, A. S., Lisker, L., & Mody, M. (1990). Gradient effects of fundamental frequency on stop consonant voicing judgments. *Phonetica*, *47*(1–2), 36–49.
Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic f0 of vowels. *Journal of Phonetics*, *23*, 349–366.
Whiteside, S. P., & Irving, C. J. (1998). Speakers' sex differences in voice onset time: a study of isolated word production. *Perceptual and Motor Skills*, *86*(2), 651–654.
Wightman, C. W., & Ostendorf, M. (1994). Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, *2*(4), 469–481, http://dx.doi.org/10.1109/89.326607.
Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, *72*(1), 37–48, http://dx.doi.org/10.1016/j.jml.2013.12.003.
Yao, Y. (2007). Closure duration and VOT of word-initial voiceless plosives in English in spontaneous connected speech. *UC Berkeley phonology lab annual reports*, 183–225.
Yao, Y. (2009). Understanding VOT variation in spontaneous speech. In M. Pak (Ed.), *Current numbers in unity and diversity of languages* (pp. 1122–1137). Seoul: Linguistic Society of Korea.
Yao, Y., Tilsen, S., Sprouse, R. L., & Johnson, K. (2010). Automated measurement of vowel formants in the Buckeye corpus. *UC Berkeley phonology lab annual reports*, 80–94.
Yoon, T., & Kang, Y. (2013). *The Korean Phonetic Aligner Program Suite*. ⟨http://korean.utsc.utoronto.ca/kpa/⟩.
Yu, A.C.L., Abrego-Collier, C., Phillips, J., Pillion, B., & Chen, D. (2015). Investigating variation in English vowel-to-vowel coarticulation in a longitudinal phonetic corpus. *In* The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th international congress of phonetic sciences*. Paper number 0519. Glasgow, UK.
Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *In Proceedings of Acoustics '08*, 5687–5790. http://doi.org/10.1121/1.2935783.
Yuan, J., & Liberman, M.Y. (2011). Automatic measurement and comparison of vowel nasalization across languages. *In* Lee, W.-S., & Zee, E. (Eds.), *Proceedings of the 17th international congress of phonetic sciences*, 2244–2247. Hong Kong.
Zavaliagkos, G., Schwartz, R., & Makhoul, J. (1995). Batch, incremental and instantaneous adaptation techniques for speech recognition. *In Proceedings of the international conference on acoustics, speech, and signal processing* , 676–679. Detroit, MI.
Zlatin, M. A. (1974). Voicing contrast: perceptual and productive voice onset time characteristics of adults. *The Journal of the Acoustical Society of America*, *56*(3), 981–994, http://dx.doi.org/10.1121/1.1903359.
Zue, V. W. (1976). *Acoustic characteristics of stop consonants: a controlled study (Ph.D dissertation)*. Cambridge: Massachusetts Institute of Technology.