Laboratory Phonology
Journal of the Association for Laboratory Phonology

Nakai, S and Scobbie, J M 2016 The VOT Category Boundary in Word-Initial Stops: Counter-Evidence Against Rate Normalization in English Spontaneous Speech. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7(1): 13, pp. 1–31, DOI: http://dx.doi.org/10.5334/labphon.49

JOURNAL ARTICLE

The VOT Category Boundary in Word-Initial Stops: Counter-Evidence Against Rate Normalization in English Spontaneous Speech

Satsuki Nakai and James M. Scobbie
Clinical Audiology, Speech and Language Research Centre, Queen Margaret University, UK
Corresponding author: Satsuki Nakai (satsuki@ovod.net)

Some languages, such as many varieties of English, use short-lag and long-lag VOT to distinguish word- and syllable-initial voiced vs. voiceless stop phonemes. According to a popular view, the optimal VOT category boundary between the two types of stops moves towards larger values as articulation rate becomes slower (and speech segments longer), and listeners accordingly shift the perceptual VOT category boundary. According to an alternative view, listeners do not shift the VOT category boundary with a change in articulation rate, because the same category boundary remains optimal across different rates of articulation in normal speech, although a shift in the optimal boundary location can be induced in the laboratory by instructing speakers to use artificially extreme articulation rates. In this study we compared the effectiveness of rate-independent VOT category boundaries applied to word-initial stop phonemes in spontaneous English speech, against the effectiveness of Miller et al.'s (1986) rate-dependent VOT category boundary applied to laboratory speech. The effectiveness of the two types of category boundaries were comparable, when spontaneous speech data were controlled for factors other than articulation rate. Our results suggest that perceptual VOT category boundaries need not shift with a change in articulation rate under normal circumstances.

Keywords: VOT; category boundary; English; spontaneous speech; rate normalization

1 Introduction

Voice onset time (the interval between stop release and onset of vocal cord vibration, hereafter VOT) is a primary acoustic cue that differentiates voiced from voiceless stop phonemes in word- and syllable- initial positions in many languages (Beckman et al., 2011; Cho & Ladefoged, 1999; Kessinger & Blumstein, 1997; Lisker & Abramson, 1964, 1967, 1970). In English, initial voiced stop phonemes are generally said to have a VOT of 15 ms or less (short-lag VOT or prevoiced), and voiceless stop phonemes some 30 ms or longer (long-lag VOT) (Lieberman & Blumstein, 1988, p. 215). Speech segment durations are affected by articulation rate, however (e.g., Gaitenby, 1965). Phonetic-phonological research has thus long been interested in how articulation rate affects VOT, and how listeners recover the correct voicing specifications of stop phonemes despite surface variation of VOT in the input.

There are two contrasting views on this issue. The widely accepted view rests on claims that the VOT category boundary location that optimally distinguishes short-lag and long-lag categories shifts with articulation rate. On this view, languages that contrast these categories such as English require rate-dependent VOT category boundaries to distinguish voiced and voiceless stop phonemes effectively, with a larger VOT value for the category boundary at a slower articulation rate (e.g., Miller et al., 1986). Based on their linguistic experience, the listeners shift the perceptual VOT category boundary, or “normalize” the boundary location, according to articulation rate to correctly identify the stop’s voicing specification from VOT.

The less accepted view states that short-lag VOT hardly changes with articulation rate and serves as a phonetic anchor in maintaining the voicing contrasts, and the same VOT category boundary location remains optimal across different rates of articulation (Kessinger & Blumstein, 1997). On this view, the listeners do not shift the VOT category boundary with a change in articulation rate in order to correctly identify the stop’s voicing specification from VOT.

Evidence from perception studies has been generally interpreted as supporting the rate normalization view. Many past studies report a shift in the perceptual VOT category boundary, with larger values for slower articulation rates, emulated by manipulating the duration of surrounding speech segments (e.g., Green et al., 1994; Green & Miller, 1985; Kidd, 1989; Miller & Dexter, 1988; Newman & Sawusch, 1996; Summerfield, 1981; Volaitis & Miller, 1992).

Even so, it has been noted that such shifts in perceptual VOT category boundary locations are often much smaller in magnitude than expected from production studies (Kessinger & Blumstein, 1998; Miller et al., 1986; Pind, 1995; Summerfield, 1975). This is most evident in Pind’s (1995) Icelandic study. In that study a mere 1.5 ms shift in the perceptual category boundary location was observed, where production data predicted a roughly 20 ms shift, although at least the observed shift was in the predicted direction and was statistically significant. The production-perception mismatch is problematic for perceptual normalization views, which assume that rate normalization processes reflect the listener’s “detailed knowledge of the temporal regularities of speech” (Nooteboom, 1979, p. 304).

From a psycho-acoustic perspective, some researchers have cast doubts on the interpretation of perceptual rate normalization studies. Diehl and Walsh (1989) found that the same nonspeech sound is perceived to be shorter before a long sound than before a short sound, and suggested that the findings of perceptual rate normalization studies may instead be attributed to general auditory contrast effects (see also Pisoni et al., 1983). Although Diehl and Walsh (1989) concerned the English /b/-/w/ contrast, if we applied the principle of auditory contrast effects to typical situations in perceptual VOT boundary experiments, the same VOT would be perceived to be shorter before a long segment (in the slow articulation condition) than before a short segment (in the fast articulation condition), which would produce a shift in the VOT category boundary in the direction reported by the perceptual rate normalization studies (see Reinisch and Sjerps [2013] for similar effects induced by temporally manipulating preceding speech contexts). In other words, the observed shifts in VOT category boundary locations in the previous perception experiments could have arisen from general auditory effects rather than speech rate normalization, which reflects listeners’ knowledge of the temporal regularities of speech.

From another perspective, Toscano and McMurray (2012) also argue against perceptual rate normalization of VOT. These authors suggest that English-speaking listeners use the duration of the vowel following a stop onset as an independent cue to the stop’s voicing specification, not as a cue to articulation rate as generally held. All else being equal, vowels following a voiced stop onset (measured from the onset of voicing) are longer than vowels following a voiceless stop onset in English (Allen & Miller, 1999). This vowel duration difference can serve as a secondary cue to the preceding stop’s voicing specification. Consequently, listeners are more biased towards the “voiced” response when the vowel following a stop onset is longer (and vice versa), which gives an appearance of rate normalization.

We suspect that the prediction of rate-dependent shift in perceptual VOT category boundary location is an artifact of rather unnatural elicitation methods used in production studies.

For example, Miller et al. (1986), Volaitis and Miller (1992), and Pind (1995) all used a "magnitude production technique", in which the participants were instructed to produce test syllables/words (e.g., /pi/) at several rates: at normal rate, twice normal rate, four times normal rate, as fast as possible, and so on. Such elicitation methods reveal what the speakers are capable of, but not necessarily what they produce in everyday communication. That is, the ranges of VOT values elicited in these studies are not ecologically grounded, and might not be relevant to central theoretical models of speech communication. (We do *not* mean that laboratory speech production studies are always or necessarily undesirable. See Xu [2010] for the advantages of well-constructed laboratory speech materials.)

We are aware that the ranges of articulation rate used in the studies employing the magnitude production technique are not entirely arbitrary. Implicitly, they are informed by Miller et al.'s (1984) study on variability in articulation rate in spontaneous speech, where articulation rate was expressed as the mean syllable duration of each pause-free stretch of speech. While we agree with Miller et al. (1984) that articulation rate may fluctuate during a conversation, the estimated variability in articulation rate in that study perhaps is inflated, because it conflates variability arising from various sources such as segments' intrinsic durations and prosodic temporal adjustments.

In Lehiste (1972), for instance, the duration of *stick* differed by a factor of 1.6 when her speakers produced the word in isolation vs. in a sentence (*the stick was discarded*) at a subjectively constant rate (see also Frank & Jaeger, 2008; Yuan et al., 2006). Unlike *stick* produced in the sentence, *stick* produced in isolation most probably underwent accentual and utterance-final lengthening, among other things, resulting in rather different durations of *stick* at similar articulation rates.[1] In our view, these additional sources of durational variability should be distinguished from general "articulation rate", manipulated in a majority of rate normalization studies by instructing participants to produce speech materials (often isolated syllables/words) at different speeds, or by resynthesizing speech materials to shorten or lengthen their overall durations, a common approach for perception experiments.

More recently, Nagao and de Jong (2007) elicited target syllables (/bi/ vs. /pi/) of a much smaller durational range than Miller et al. (1986), and reported a comparable rate-dependent shift in the VOT category boundary in production and perception, except in the fast speech rates. However, participants produced test syllables in time with a metronome, which again deviates from everyday speech production. Additionally, as the authors note, spoken syllables from the production experiment were used as stimuli in the perception experiments without controlling other acoustic cues for voicing specifications such as F_0, formant transitions, and the amplitude of aspiration noise (Haggard et al., 1981; Repp, 1979; Stevens & Klatt, 1974). It is thus unclear whether the participants identified stimuli with a long VOT as voiced in slow speech more often (and vice versa) because of perceptual rate normalization, or because of other cues for voicing compatible with the intended category *despite* atypical VOT values.

Whether or not they subscribe to rate normalization views, virtually all production studies report asymmetrical effects of articulation rate on voicing categories, with much smaller effects on short-lag than long-lag categories (Kessinger & Blumstein, 1997; Magloire & Green, 1999; Miller et al., 1986; Nagao & de Jong, 2007; Pind, 1995; Schiavetti et al., 1996; Stuart-Smith et al., 2015; Volaitis & Miller, 1992). Conceivably, for naturally occurring ranges of VOT, a rate-independent category boundary between short-lag and long-lag VOT is effective enough across different rates of articulation. Other voicing cues would

[1] This does not mean that the prosodic organization of speech and articulation rate are completely independent of each other. A change in articulation rate can affect the prosodic structure of speech (Shattuck-Hufnagel & Turk, 1996), as well as the likelihood of phonological reduction (Shockey, 1987). We touch on this issue in Section 3.3.

still be useful, as cue redundancy makes speech perception more robust and effective (Nakai & Turk, 2011; Wright, 2004).

To see how relevant the existing literature on perceptual rate normalization of category boundary locations is for naturally occurring ranges of VOT values, we examined word-initial voiced vs. voiceless English stop phonemes (the subject of many rate normalization studies) in spontaneous speech. In Miller et al. (1986) voiced vs. voiceless stop phonemes were produced at various articulation rates, and optimal category boundaries (described in Section 2.3 below) were estimated for syllables grouped by 50-ms intervals. In that study, estimating articulation rate was relatively straightforward because the speech materials were tightly controlled phonetically and prosodically: isolated /bi/ vs. /pi/.

As we pointed out, accurately quantifying the articulation rate of spontaneous speech is not easy, because word sequences and their prosodic groupings vary from utterance to utterance, adding noise to the estimated articulation rate. Therefore, in our main analysis we applied rate-independent optimal VOT category boundaries to spontaneous speech data, and compared their classification accuracy with the overall classification accuracy achieved by Miller et al.'s (1986) rate-dependent optimal category boundary. To make our spontaneous speech data roughly comparable to Miller et al.'s (1986) well-controlled speech material, in our application of rate-independent category boundaries we took into account factors known to affect VOT other than articulation rate (place of articulation, lexical stress, following vowel, word class; see, e.g., Lisker & Abramson, 1967).[2]

If such rate-independent category boundaries are as effective as Miller et al.'s (1986) rate-dependent category boundary, then we can conclude that classification accuracy is unlikely to improve by additionally taking articulation rate into account. Put differently, comparable performances of rate-independent category boundaries applied to spontaneous speech and Miller et al.'s (1986) rate-dependent category boundary applied to laboratory speech would speak against the need for perceptual rate normalization of VOT category boundaries under natural circumstances.

2 Methods

2.1. Data

The spontaneous speech sample used in this study comprised ten episodes of a BBC (the British Broadcasting Corporation) Radio 3 program "the Lebrecht Interview", broadcasted in 2011 and 2012. Each 45-minute episode featured a prominent artist or administrator in classical music, who talked to the interviewer (a music commentator) at a radio station about work and life in a conversational style. The interviewees whose speech was analyzed comprised four males and six females (age range: 37–78, $\bar{x} = 62$). They were all native speakers of English, from different parts of the world: United States ($n = 4$), United Kingdom (5), and Australia (1) (see Discussion for possible effects of dialectal differences). The episodes were streamed on iPlayer (http://www.bbc.co.uk/radio3) on a MacBook Pro and captured using Audacity, via a Soundflower input/output device at a 44.1 kHz sampling rate with 16 bit quantization, a standard used in the BBC radio studio recordings in the UK (British Broadcasting Corporation, 2010). The resulting audio recordings had a bandwidth of c. 20 kHz. No dropouts were detected.

[2] Baran et al. (1977) also examined the VOT distributions of homorganic English voiced vs. voiceless stop phonemes in spontaneous speech. They reported "an appreciable overlap" (p. 347) between the VOT distributions of voiced vs. voiceless stops, although they found no direct relation between VOT and syllable rate. The source of the overlap in Baran et al. (1977) is unclear, for they pooled together all instances of each stop phoneme without taking account of factors other than place of articulation and speaking conditions (e.g., adult-directed vs. child-directed speech).

2.2. VOT measurements

All instances of English words beginning with one of the six oral stop phonemes (/b/, /d/, /g/, /p/, /t/, and /k/) as a simplex onset were identified in the interviewees' speech for VOT measurements. Words of a foreign origin were excluded unless they were listed in the Collins online English dictionary (http://www.collinsdictionary.com/dictionary/english) with Anglicized pronunciations and judged to have been part of the English language for some time. For example, *Bach* and *Berlin* were included, but *Bayreuth* and *Dudamel* were excluded. Altogether, 10,479 words that satisfied the criteria were identified (see **Table 1**).

Of those, the VOT of the initial stop of 422 words (4%) were not measured because of overlapping speech, noise, a devoiced following vowel, unclear stop release, or the stop's realization as a glide, fricative, tap or nasal. Many (63%) of these belonged to function words, with /t/ in *to* accounting for 36% of all unmeasured tokens (though, as the most frequent /t/-initial word, 1,596 tokens still remained). Words spoken while laughing were also excluded, as we were uncertain to what extent the speaker had control over the duration of VOT. Words from disfluent sections of speech were included as an intrinsic part of spontaneous speech so long as the word was completed and identifiable, except one case of suspected substitution error (*Boint P* for *Point B*). Unfinished words were excluded, as many of them were just one syllable (e.g., *bi- Beatles*) and did not provide sufficient phonetic evidence to be absolutely sure which stop the speaker had intended.[3]

The VOT of the remaining 10,057 words were measured manually by the first author in Praat (Boersma & Weenink, 2012). VOT was defined as the interval between the first clear sign of stop release to the first clear sign of voicing that continued into the following vowel, as determined on the waveform in conjunction with spectrographic information (see **Figure 1**). This meant that no negative VOT was used; a VOT of 0 ms was assigned to prevoiced utterance-initial stops and utterance-medial stops produced with continuous voicing from before the stop release. This decision was made because the onset of voicing could not be easily determined for a majority of such cases, which were utterance-medial and had continuous voicing from segments before the stop closure (see also Lisker & Abramson, 1967; Stuart-Smith et al., 2015). As we elaborate in Section 2.3, this did not affect the locations of optimal VOT category boundaries or their classification accuracy, our main analysis tools. The portion of pseudo-regular waveform corresponding to a mixture of voicing and noisy aperiodic excitation at the release of stop closure was excluded from VOT. (VOT category boundaries estimated using this approach would be at smaller

Onset	Identified	Measured	% Excluded
/b/	2,266	2,212	2
/p/	1,035	1,013	2
/d/	1,788	1,703	5
/t/	2,705	2,529	7
/g/	824	790	4
/k/	1,861	1,810	3
Tol.	10,479	10,057	4

Table 1: Number of identified and measured simplex word-initial stops.

[3] We do not know whether BBC removed other disfluent sections of the interviews before broadcasting. However, we did not detect any sign of editing targeting disfluency; each episode contained what felt like a natural amount of fillers, hesitations, pauses, and rephrasing. In all episodes (some more so than others), utterances like the following were not uncommon: *No … no and it's I I regret … having done an academic music degree.*

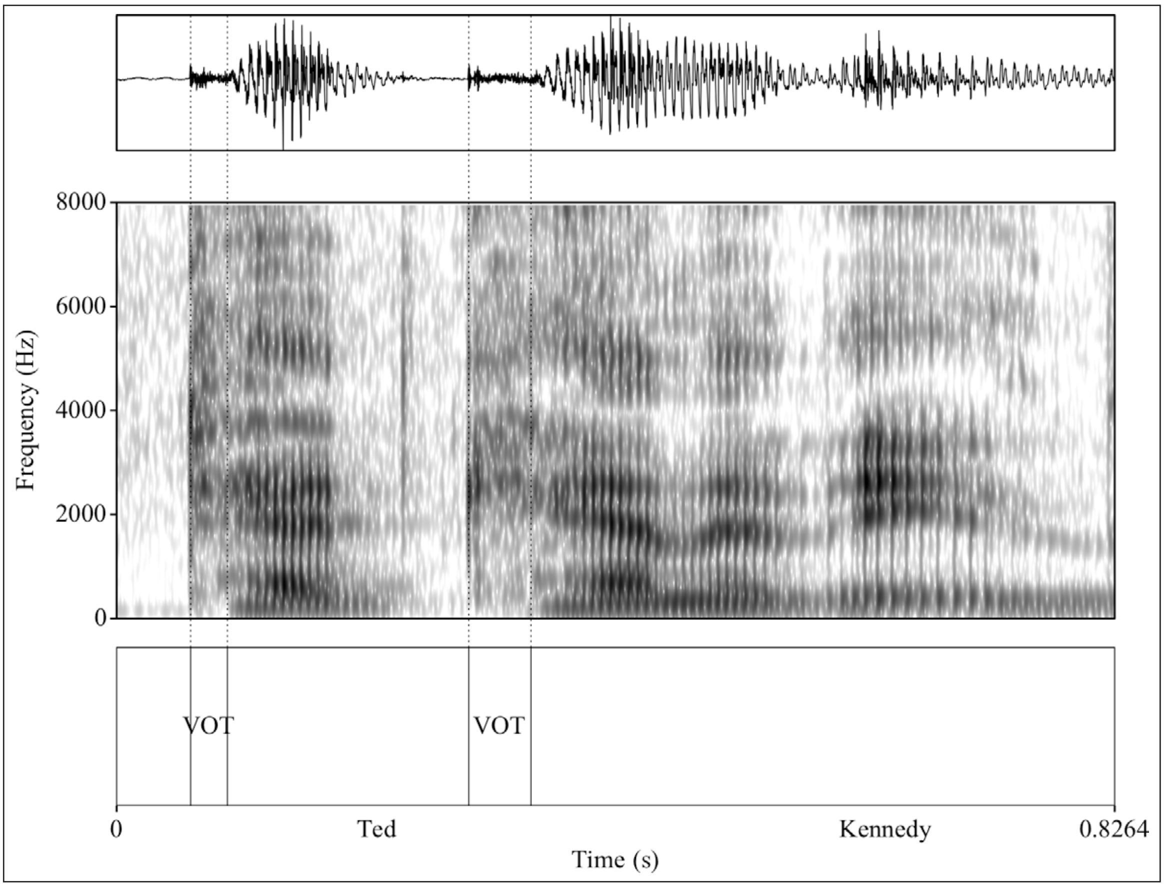

Figure 1: VOT intervals of /t/ and /k/ in *Ted Kennedy.*

values than those estimated using an approach that includes the frication portion in VOT, regardless of concurrent voicing.)

For reliability, the second author measured the VOT of roughly 5% (500 tokens) of all measured tokens, selected randomly. The Spearman's correlation coefficient between the two authors' VOT measurements for each homorganic stop pair was: r_s = 0.87 for /b/-/p/; r_s = 0.95 (/d/-/t/); r_s = 0.96 (/g/-/k/). The median difference between the repeated measurements was 1.7 ms for /b/-/p/, 2.8 ms for /d/-/t/, and 2.4 ms for /g/-/k/.

2.3. Optimal category boundary location

The optimal category boundary location between the two members of each of the three pairs of homorganic stops (/b/-/p/, /d/-/t/, and /g/-/k/) was estimated using Miller et al.'s (1986) categorization method. In this method, a candidate category boundary is placed along the VOT continuum; all items to the left of the boundary (VOT smaller than the value at the boundary) are classified as voiced, and all items to the right of the boundary are classified as voiceless. The boundary location that classifies the voicing specifications of the greatest proportion of the stop phonemes correctly (voiced and voiceless stops combined) is defined as optimal. For example, a category boundary placed at a very small VOT value (e.g., 5 ms) would classify most voiceless stop phonemes correctly but misclassify many voiced stop phonemes, resulting in a low overall classification accuracy.

In a procedural search for the optimal category boundary location, the candidate VOT boundary was moved in 1 ms steps from the smallest meaningful boundary location at 1 ms towards larger values, so that the classification accuracy improved, reached a maximum, and then started to decline. The optimal category boundary location is where the classification accuracy reaches the maximum. If maximum classification accuracy was found at multiple steps, we regarded all of them as optimal, but the midpoint of the range was used for calculations that required a single optimal VOT value.

As explained earlier, we assigned a VOT value of 0 ms to prevoiced utterance-initial stops and utterance-medial stops produced with continuous voicing from before the stop release. This did not affect the estimated optimal VOT category boundary location, as an overwhelming majority of voiceless stop phonemes and many voiced stop phonemes in our data had positive VOT values (values greater than 0 ms). Therefore, the optimal VOT category boundary, located basically at the intersection of the VOT distributions of voiced and voiceless categories, always had a positive value, as expected for a category boundary between short-lag vs. long-lag VOT (see also Miller et al. [1986], who used negative VOT values). If the optimal category boundary has a positive value, assigning 0 ms to negative VOT values makes no difference to classification accuracy, as a VOT of 0 ms would be positioned to the left of the category boundary, just like negative VOT values. Stops with a VOT of 0 ms would always be classified correctly if they are from a voiced category and wrongly if they are from a voiceless category.

2.4 Controlling spontaneous speech data

As laid out in the Introduction, our main goal is to compare the overall classification accuracy of the rate-dependent optimal category boundary applied to isolated /bi/ vs. /pi/ in Miller et al. (1986) against the accuracy of rate-independent optimal category boundaries applied to spontaneous speech data, controlled for known factors that affect VOT other than articulation rate. Rate-independent optimal category boundaries were estimated at four levels of data control: (a) all word-initial homorganic pairs of stop phonemes, (b) word-initial homorganic stop pairs in content words only,[4] (c) word-initial homorganic stop pairs in content words with word-initial (primary and non-primary) lexical stress only,[5] and (d) word-initial homorganic stop pairs in content words with word-initial lexical stress, grouped by the following vowel.

Needless to say, the controlling factors (place of articulation, word class, lexical stress, and following vowel) used here were far from exhaustive. To keep the analysis manageable in size, these factors were chosen from those reported to affect VOT durations in previous production and perception studies (e.g., Klatt, 1975; Lisker & Abramson, 1967; Yao, 2009) through inspection of items that were misclassified by the optimal category boundary at each analysis level. Among the data at the above four levels of control, the data at the final level of control (d) is the most comparable to Miller et al.'s (1986) data, which consisted of isolated /bi/ vs. /pi/ only.

3 Results

3.1 Overview of results

Table 2 provides the classification accuracy of rate-independent optimal category boundaries, along with the median VOT value of each phoneme and the semi-interquartile ranges (SIQR) of the voiceless phonemes. The SIQR was not calculated for voiced phonemes, as many of them were assigned a VOT of 0 ms, which in many cases had no numerical significance (see Section 2.2). The median VOT values of the six phonemes at the first level of analysis (all words) were comparable to the mean VOT values of corresponding

[4] Content words in our data comprised nouns, adjectives, adverbs, verbs, numbers, and interjections. Interjections seemed to behave differently from content words proper, but they were small in number and did not affect the overall results. Excluded were function words (auxiliary verbs, the copula *be*, conjunctions, prepositions, and the infinitive marker *to*) as well as function-word-like words (Selkirk, 1996), namely *going* in *going to* and *gonna* expressing future, and *got* in *have got to* expressing modality.

[5] Distinguishing primary vs. non-primary lexical stress did not affect the overall results.

All words		**/b/-/p/**	**/d/-/t/**	**/g/-/k/**
Classification accuracy		*94.8%*	*89.0%*	*91.2%*
n		3,225	4,232	2,600
Boundary location (ms)		16	24	27
M (SIQR)	Voiced	2	6	17
	Voiceless	35 (14)	45 (14)	49 (12)
Content words (all)		**/b/-/p/**	**/d/-/t/**	**/g/-/k/**
Classification accuracy		96.4%*	96.2%*	92.7%*
n		1,827	2,340	2,135
Boundary location (ms)		13	28	27
M (SIQR)	Voiced	0	6	17
	Voiceless	35 (14)	54 (14)	50 (12)
Content words (initial stress)		**/b/-/p/**	**/d/-/t/**	**/g/-/k/**
Classification accuracy		97.7%*	96.7%	94.2%
n		1,536	2,042	1,732
Boundary location (ms)		13	26	31
M (SIQR)	Voiced	0	6	17
	Voiceless	37 (14)	54 (14)	54 (13)
Content words (initial stress) grouped by following vowel		**/b/-/p/**	**/d/-/t/**	**/g/-/k/**
Classification accuracy		98.4%	98.2%*	97.8%*
n		1,536	1,996 (see Note)	1,700 (see Note)
Boundary location (ms)			see Table 4 in Section 3.5	

Table 2: Summary of the performance of rate-independent optimal category boundaries at four levels of data control. Classification accuracies in italics indicate values significantly lower than the overall classification accuracy of Miller et al.'s (1986) rate-dependent optimal category boundary. Asterisks indicate significant improvement in classification accuracy over the preceding level.

Note. /tʊ/ and /ki/ were excluded from the final level of analysis, as there were no words starting with /dʊ/ or /gi/.

phonemes in sentence context in Lisker and Abramson (1967), with bilabial stops having the shortest VOT and velar stops the longest (see also Fricke, 2013; Schiavetti et al., 1996; Stuart-Smith et al., 2015). Optimal category boundary locations for the three pairs of homorganic stops were also the shortest for bilabial stops and generally the longest for velar stops, and were roughly within the range of category boundary locations for the three places of articulation reported in Summerfield's (1975, 1981) perception studies.

Importantly, as the context other than articulation rate was progressively controlled, classification accuracy for the three pairs of homorganic voiced vs. voiceless stop contrasts gradually improved and became comparable to the classification accuracy of Miller et al.'s (1986) rate-dependent category boundary at one level or another. The results are consistent with our hypothesis that the VOT category boundary between voiced vs. voiceless stop phonemes need not be adjusted for articulation rate in spontaneous conversational speech to maintain a high degree of accurate phoneme classification. We detail below how rate-independent category boundaries fared with Miller et al.'s (1986) rate-dependent category boundary at each level of data control.

3.2 Word-initial homorganic stop pairs, unrestricted otherwise

VOT is affected by the stop's place of articulation (e.g., Lisker & Abramson, 1967), which is reflected in the perceptual VOT category boundary location between voiced vs. voiceless stops (e.g., Lisker & Abramson, 1970). At the first level of data control, we therefore grouped all word-initial stop phonemes by place of articulation. **Figure 2** plots the durational distributions of VOT for the three pairs of homorganic stops. VOT distributions for /b/-/p/ are reasonably well separated, while those for /d/-/t/ and /g/-/k/ appear to have non-negligible overlap. As given in **Table 2** above, rate-independent optimal category

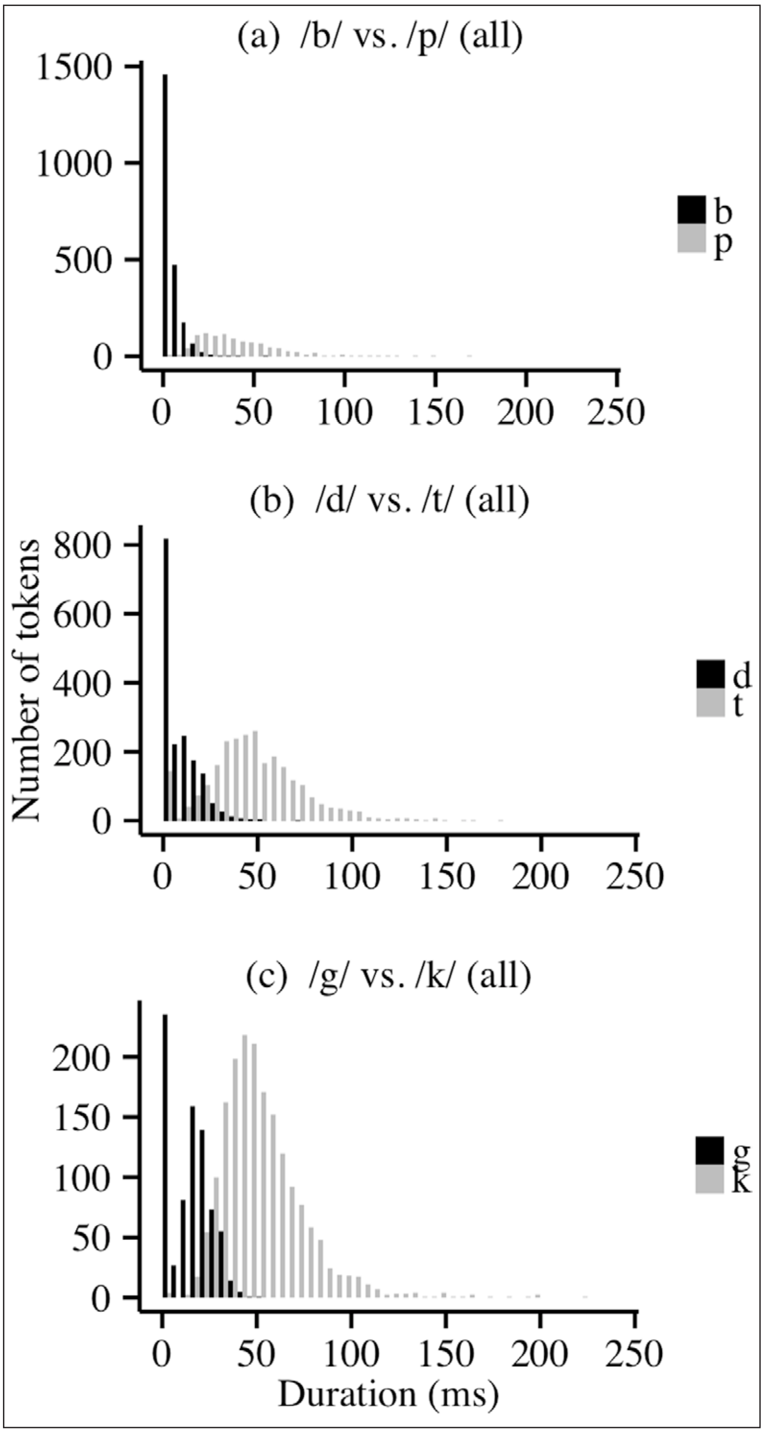

Figure 2: Durational distributions of all measured VOT of word-initial simplex stop phonemes for (a) bilabial, (b) alveolar, and (c) velar places of articulation.

boundaries correctly classified 94.8% of /b/-/p/ (at 16 ms), 89.0% of /d/-/t/ (24 ms), and 91.2% of /g/-/k/ (27 ms).

Chi-square tests[6] were used to compare the number of items correctly vs. wrongly classified by the rate-independent optimal category boundaries for the three homorganic stop contrasts against the overall classification accuracy of Miller et al.'s (1986) rate-dependent optimal category boundary for /bi/-/pi/ (97.6%, $n = 1{,}013$). (Miller et al. [1986] investigated /bi/-/pi/ only.) All three rate-independent category boundaries performed significantly worse than Miller et al.'s (1986) rate-dependent category boundary (/b/-/p/: $\chi^2(1) = 13.0$; /d/-/t/: $\chi^2(1) = 70.5$; /g/-/k/: $\chi^2(1) = 41.7$; all $ps < .001$).

3.3 Word-initial homorganic stop pairs, restricted to content words

Next, we excluded function words and examined the VOT of word-initial stops in content words only. The exclusion of function words was expected to significantly improve the accuracy of rate-independent category boundaries. Common function words are frequent in occurrence and susceptible to phonetic reduction across syllable rates (Fosler-Lussier & Morgan, 1999). Moreover, function words are more often recognized after the word's acoustic offset, that is, not immediately recognized from acoustic information alone (Bard et al., 1988), which suggests that their acoustic encoding is prone to ambiguity. As **Table 3** shows, at the previous level of data control, function words indeed contributed proportionally more to the overlap in the VOT distributions of voiced and voiceless stops for all homorganic pairs than did content words with word-initial lexical stress (but not necessarily more than content words with an unstressed word-initial syllable; more on this in Section 3.4)

Figure 3 shows the VOT distributions of each homorganic stop pair when function words were excluded. For all pairs (particularly /d/-/t/) voiced vs. voiceless stops were better separated than the previous level of data control. As given in **Table 2** above, rate-independent optimal category boundaries now correctly classified 96.4% of /b/-/p/ (at 13 ms), 96.2% of /d/-/t/ (28 ms), and 92.7% of /g/-/k/ (27 ms).

Word Type	Contrast	*n*	Misclassification
Content, Stressed initial syllable	/b/-/p/	1,536	3.3%
	/d/-/t/	2,042	4.5%
	/g/-/k/	1,732	6.4%
Content, Unstressed initial syllable	/b/-/p/	291	13.0%
	/d/-/t/	298	8.7%
	/g/-/k/	403	11.2%
Function	/b/-/p/	1,398	5.4%
	/d/-/t/	1,892	18.3%
	/g/-/k/	465	13.1%

Table 3: Misclassification rates per word type, at the first level of data control (all stops, grouped by place of articulation).

[6] All comparisons of the classification accuracy of different category boundaries in this study used Chi-square tests with Yates's continuity correction, commonly used for the analysis of 2 × 2 contingency tables. The results of Chi-square tests without Yates's correction (recommended by Field, 2005, pp. 691–692) are not reported, as the conclusions would be the same.

The improvement in the accuracy of the rate-independent optimal category boundary was statistically significant for all three stop pairs (/b/-/p/: $\chi^2(1) = 5.7$, $p = .02$; /d/-/t/: $\chi^2(1) = 100.7$, $p < .001$; /g/-/k/: $\chi^2(1) = 4.2$, $p = .04$). The accuracy of rate-independent category boundaries for /b/-/p/ and /d/-/t/ now only marginally differed from the accuracy of Miller et al.'s (1986) rate-dependent category boundary (/b/-/p/: $\chi^2(1) = 2.89$, $p = .09$; /d/-/t/: $\chi^2(1) = 3.8$, $p = .05$), although the rate-independent category boundary for /g/-/k/ still performed significantly worse than Miller et al.'s (1986) rate-dependent category boundary ($\chi^2(1) = 26.0$, $p < .001$).

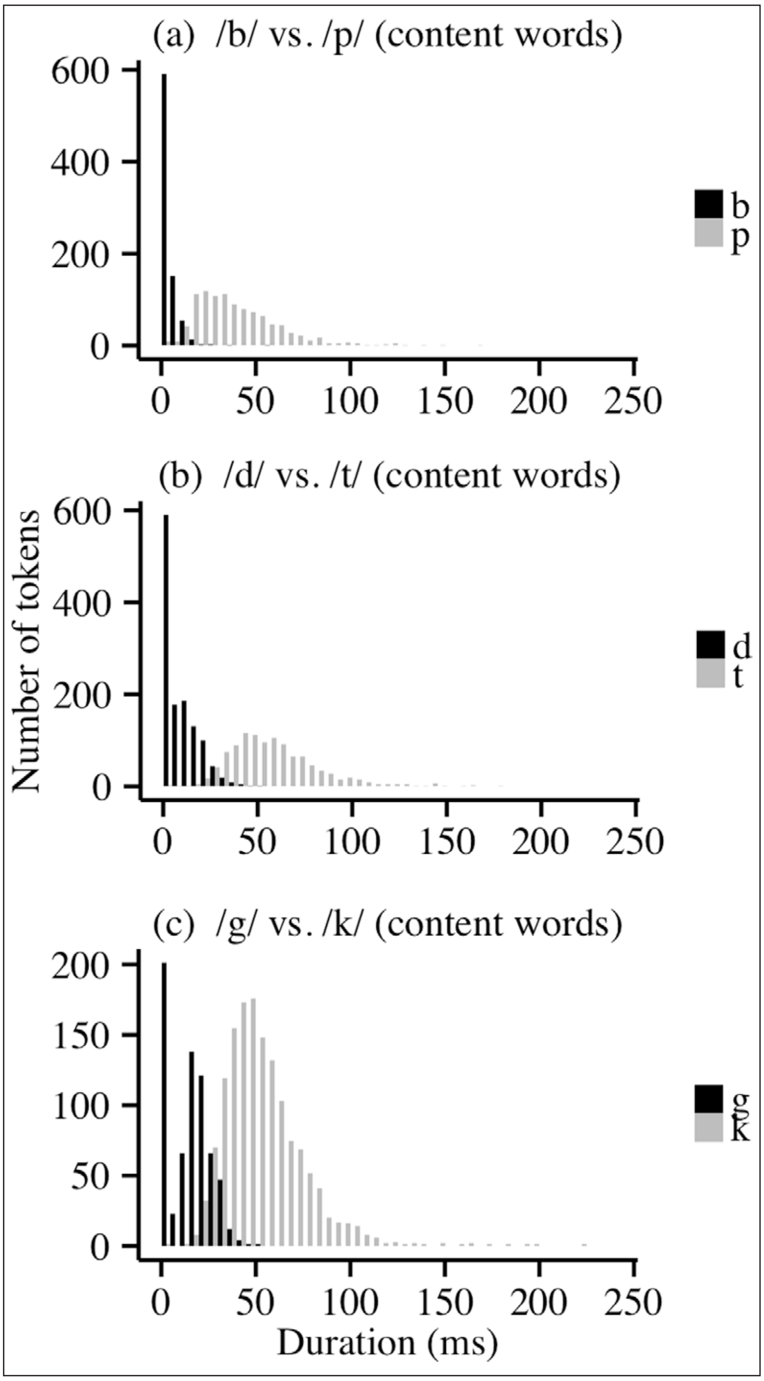

Figure 3: Durational distributions of VOT of word-initial simplex stop phonemes in content words for (a) bilabial, (b) alveolar, and (c) velar places of articulation.

As stated earlier, we excluded function words on the premise that common function words are susceptible to phonetic reduction across syllable rates and their acoustic encoding can be ambiguous. Is it possible that by excluding function words we have removed the benefits of the rate-dependent category boundary?[7]

To address this issue, we compared the effectiveness of rate-independent vs. rate-dependent VOT category boundaries for /d/ in /du/ (*do*) and /t/ in /tu/ (*to*, *too*, and *two*). We chose these words because *to* was by far the most frequent function word (n = 1,596), accounting for 43% of their occurrences, and its voiced counterpart *do* occurred reasonably often (n = 319, verb and auxiliary verb usage combined). *Too* and *two* were also frequent among content words (n = 33 and 93). All speakers produced multiple measurable tokens of *to* and *do*, and at least one measurable token of *too* or *two*.

For the estimation of articulation rate, segments in *do*, *to, too*, and *two* were not used, as their short durations (especially segments in *to* and auxiliary verb *do*) can potentially be ascribed to phonetic reduction. Instead, the mean duration of segments in the preceding word was used as a rough index of local articulation rate, assuming similar articulation rates for adjacent stretches of speech. Mean segment (rather than syllable) durations were used, as the former correlated more strongly with the duration of the target VOT: $r_s = .24$ ($p < .003$) for /du/, $r_s = .31$ ($p < .001$) for /tu/, according to Spearman's rank correlation tests.

Because of the way articulation rate was estimated, the analysis here excludes utterance-initial *do*, *to, too*, and *two*, which had no preceding word within the same utterance. Also excluded were cases where the preceding word duration could not be measured using a supralaryngeal criterion (Turk et al., 2006), for example, where the initial segment of the preceding word was a stop phoneme following a pause. This left for analysis 161 tokens of *do*, 667 *to*, 17 *too*, and 63 *two*.

Figure 4 shows the relationship between the VOT of /du/ and /tu/, and the mean segment duration of the preceding word (hereafter "articulation rate"). Several observations can be made. First, most instances of /t/ with a short VOT (< c. 25 ms) belonged to *to* produced at fast-mid articulation rates (mean duration of preceding segments < c. 100 ms). Such a short VOT was seldom found for *too* and *two*, even though these words also mainly occurred at fast-mid articulation rates. Thus, the short VOT observed for many tokens of *to* at fast-mid articulation rates seems to have arisen from phonetic reduction rather than articulation rate per se. Phonetic reduction was, unsurprisingly, unlikely to occur at slow articulation rates (see also Frank & Jaeger, 2008). Other types of reduction, for example, vowel devoicing, found for *to* but excluded from the analysis (see Section 2.2), also occurred predominantly at fast-mid articulation rates.

Second, VOT for *do* did not strictly increase with articulation rate, although there was a weak positive correlation between the two. Third, at fast-mid articulation rates, where a majority of *do* and *to* (both 80%) occurred, their VOT distributions completely overlapped at the short VOT range. As a result, rate-dependent optimal category boundaries produced little advantage over the rate-independent optimal category boundary. The rate-independent optimal category boundary for all tokens of *do*, *to, too*, and *two* yielded classification accuracy of 84.0% (at 6–7 ms). A rate-dependent category boundary yielded 84.7% classification accuracy when the optimal boundary was adjusted for each 50-ms bin of the estimated articulation rate, and 84.8% accuracy when the boundary was adjusted for each 25-ms bin. The effectiveness of the rate-dependent boundaries did not

[7] We thank one of the anonymous reviewers for suggesting this possibility.

differ significantly from that of the rate-independent boundary (50-ms bin: $\chi^2(1) = 0.10$, $p = .75$; 25-ms bin: $\chi^2(1) = 0.15$, $p = .70$).

As we argued in the Introduction, spontaneous speech is not readily amenable to articulation rate measurement because of numerous confounding factors that cannot be controlled easily. However, to the extent that the mean segment durations of the preceding word reflected articulation rate, we found no clear advantage of rate-dependent over rate-independent VOT category boundaries.

The failure of rate-dependent VOT category boundaries to improve the classification accuracy of /du/-/tu/ does not mean that /tu/ with a short VOT cannot be acoustically distinguished from /du/. A further inspection of the data reveals that /u/ (measured from the onset of voicing) was shorter in a majority of instances of /tu/ than /du/, especially where VOT does not distinguish the two (see **Figure 5**). If we classify all instances with a short /u/ (< 80 ms) as /tu/ regardless of VOT, and apply a rate-independent VOT category boundary to the rest, we obtain a classification accuracy of 95.2% (at 23 ms), a significant improvement to the 84.0% accuracy of the rate-independent category boundary (at 6–7 ms) that ignores the following vowel duration ($\chi^2(1) = 59.7$, $p < .001$). Dividing the following vowel durations into further groups did not significantly improve the classification accuracy. (Classification accuracy achieved here was still poorer than the 97.6% of Miller et al.'s [1986] rate-dependent category boundary [$\chi^2(1) = 7.5$, $p < .007$]. We return to this issue in the discussion.)

Importantly, the short duration of /u/ found in many instances of /tu/ (particularly *to*) does not seem to have arisen primarily from fast articulation rates. As can be seen in **Figure 6**, across articulation rates we find /tu/ whose voiced portion is shorter than 80 ms, used in the earlier analysis to distinguish /du/ from /tu/, where VOT was neutralized. In contrast, only a handful of instances of /du/ had such a short /u/ even at fast articulation rates.

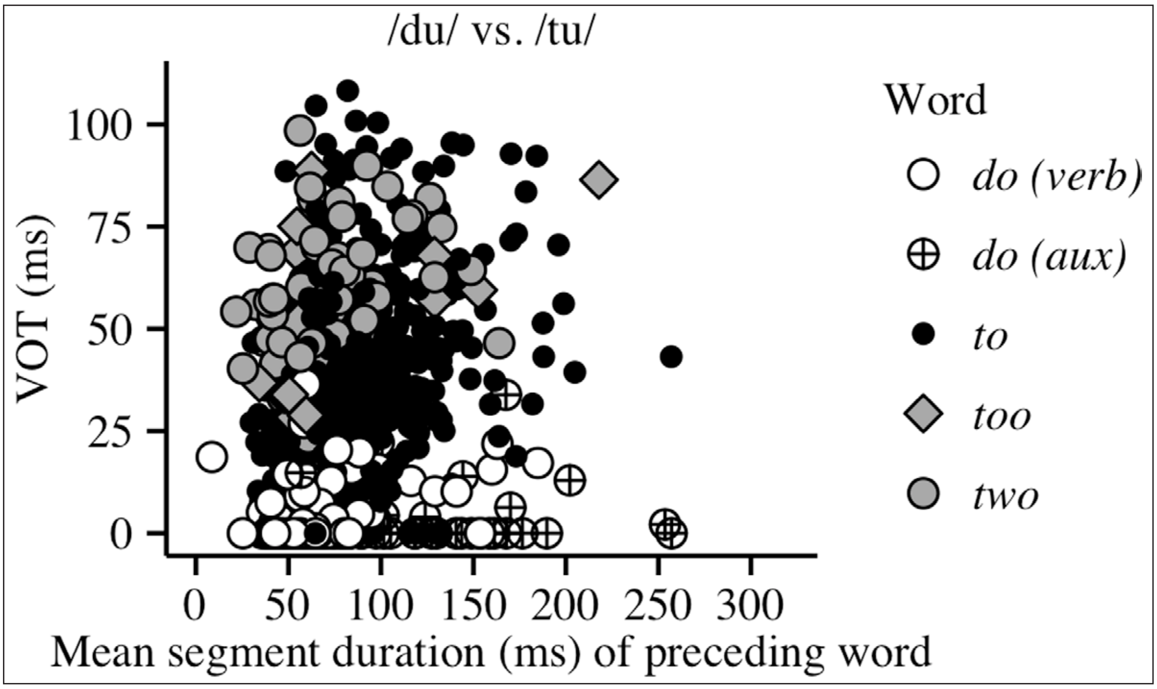

Figure 4: Relationship between mean segment duration of the preceding word and VOT of the initial stop in *do*, *to*, *too*, and *two*.

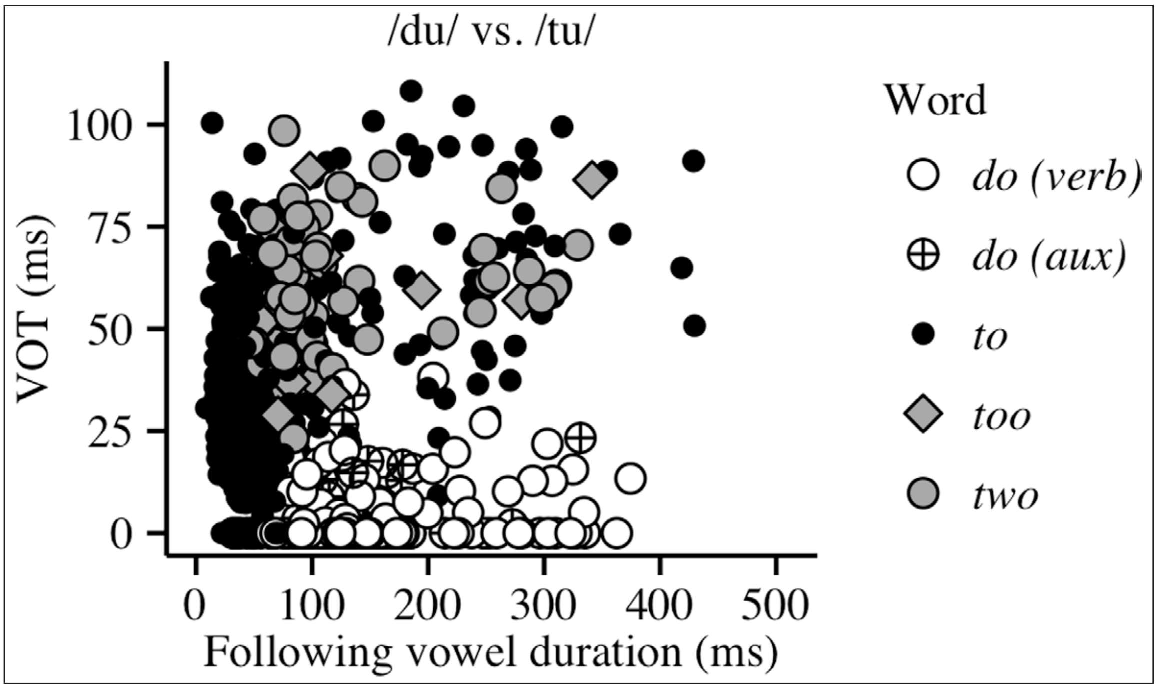

Figure 5: Relationship between VOT and duration of /u/ (measured from the onset of voicing) in *do*, *to*, *too*, and *two.*

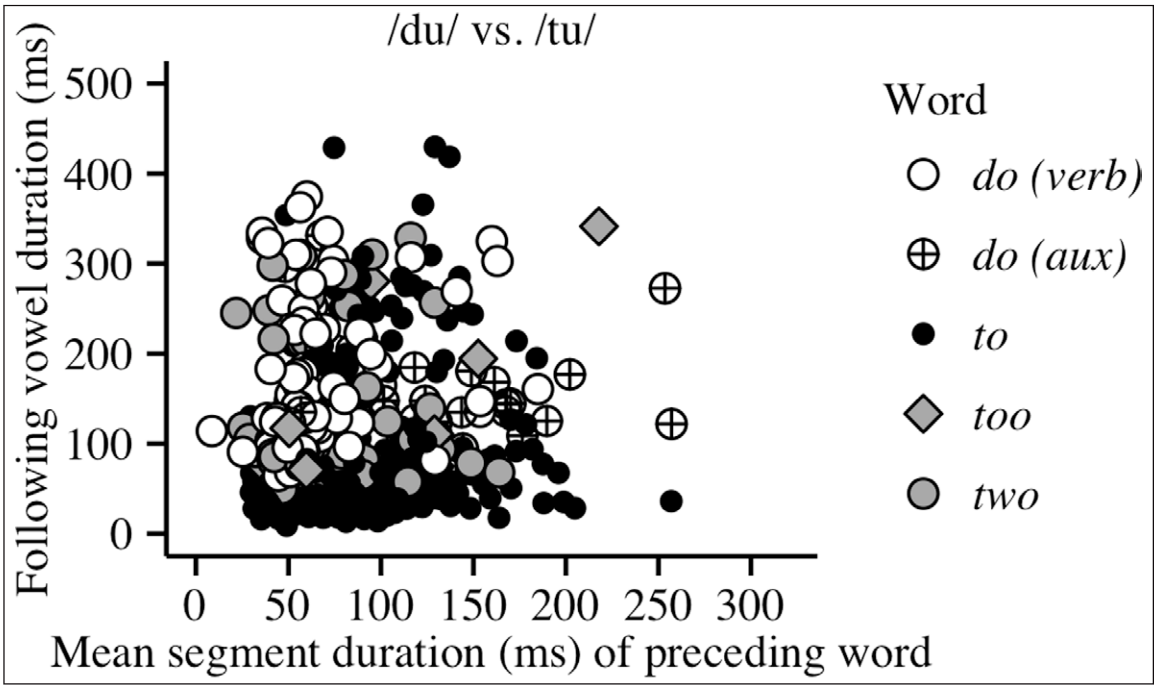

Figure 6: Relationship between duration of /u/ (measured from the onset of voicing) in *do*, *to*, *too*, and *two*, and mean segment duration of the preceding word.

In line with the above observations, regression models fitted to the data indicated that only 2% of variance in /u/ duration was explained by articulation rate alone, while 17% of variance was explained when the preceding stop's voicing specification (/d/ vs. /t/) was added to the model (a significant increase in explanatory power at $p < .001$). As **Figure 7** shows, /u/ is generally shorter in /tu/ than in /du/, and a very short /u/ suggests that the word is *to*. These results are consistent with Toscano and McMurray's (2012) finding that

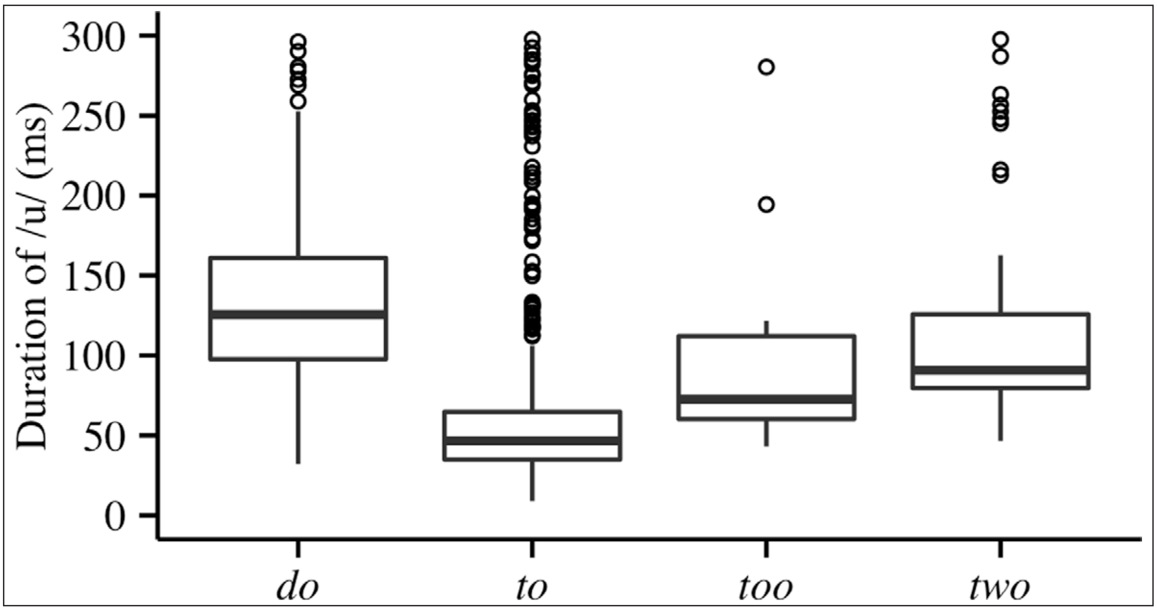

Figure 7: Duration of /u/ (measured from the onset of voicing) in *do*, *to*, *too*, and *two*. Each box shows the 25th–75th percentile of the durational distribution of /u/, and horizontal lines inside the boxes median values. Whiskers show the entire distribution, excluding outliers (shown as circles).

English-speaking listeners interpret the following vowel duration as a cue to the voicing specification of the preceding stop onset rather than articulation rate.

3.4 Word-initial homorganic stop pairs in lexically stressed syllables of content words

We saw that the stops were more likely to be misclassified in lexically unstressed than in stressed syllables of content words at the initial level of data control, where the optimal category boundary was estimated for all measured VOT for each homorganic pair of word-initial stop phonemes (see **Table 3** above). This observation is consistent with Lisker and Abramson's (1967) report that VOT values for English voiced vs. voiceless stops were less distinct in lexically unstressed syllables. As can be seen in **Figure 8**, in our data too the VOT distributions for /b/-/p/ and /d/-/t/ had a greater overlap in lexically unstressed than stressed syllables. As a result, fewer voiced vs. voiceless stops in lexically unstressed syllables were correctly classified than were stops in stressed syllables, even when optimal VOT category boundaries were separately estimated for the two types of syllables (/b/-/p/: 97.7% vs. 92.1%, $\chi^2(1) = 22.5, p < 0.001$; /d/-/t/: 96.7% vs. 94.0%, $\chi^2(1) = 4.6, p = 0.03$). As for /g/-/k/, their VOT distributions completely overlapped for unstressed syllables, though this may be ascribed to the paucity of /g/ in word-initial unstressed syllables.

As one would expect, further excluding content words without word-initial lexical stress shifted classification accuracy in the right direction (see **Table 2** above): 97.7% for /b/-/p/ (at 13 ms), 96.7% for /d/-/t/ (26 ms), and 94.2% for /g/-/k/ (31 ms). The classification accuracy of rate-independent category boundaries for /b/-/p/ and /d/-/t/ no longer differed significantly from the overall accuracy of Miller et al.'s (1986) rate-dependent category boundary ($\chi^2(1) = 0, p = 1$; $\chi^2(1) = 1.8, p = .18$), though the accuracy of the rate-independent category boundary was still poorer for /g/-/k/ ($\chi^2(1) = 16.4, p < .001$).

Except for /b/-/p/, however, the effectiveness of rate-independent category boundaries did not differ significantly from the previous level, where the data consisted of word-initial stops in all content words (/b/-/p/: $\chi^2(1) = 4.1, p = .04$; /d/-/t/: $\chi^2(1) = 0.47, p = .49$; /g/-/k/: $\chi^2(1) = 1.6, p = .21$). The lack of significant improvement compared to the previous level for all stop pairs can be ascribed to the relatively small number of content

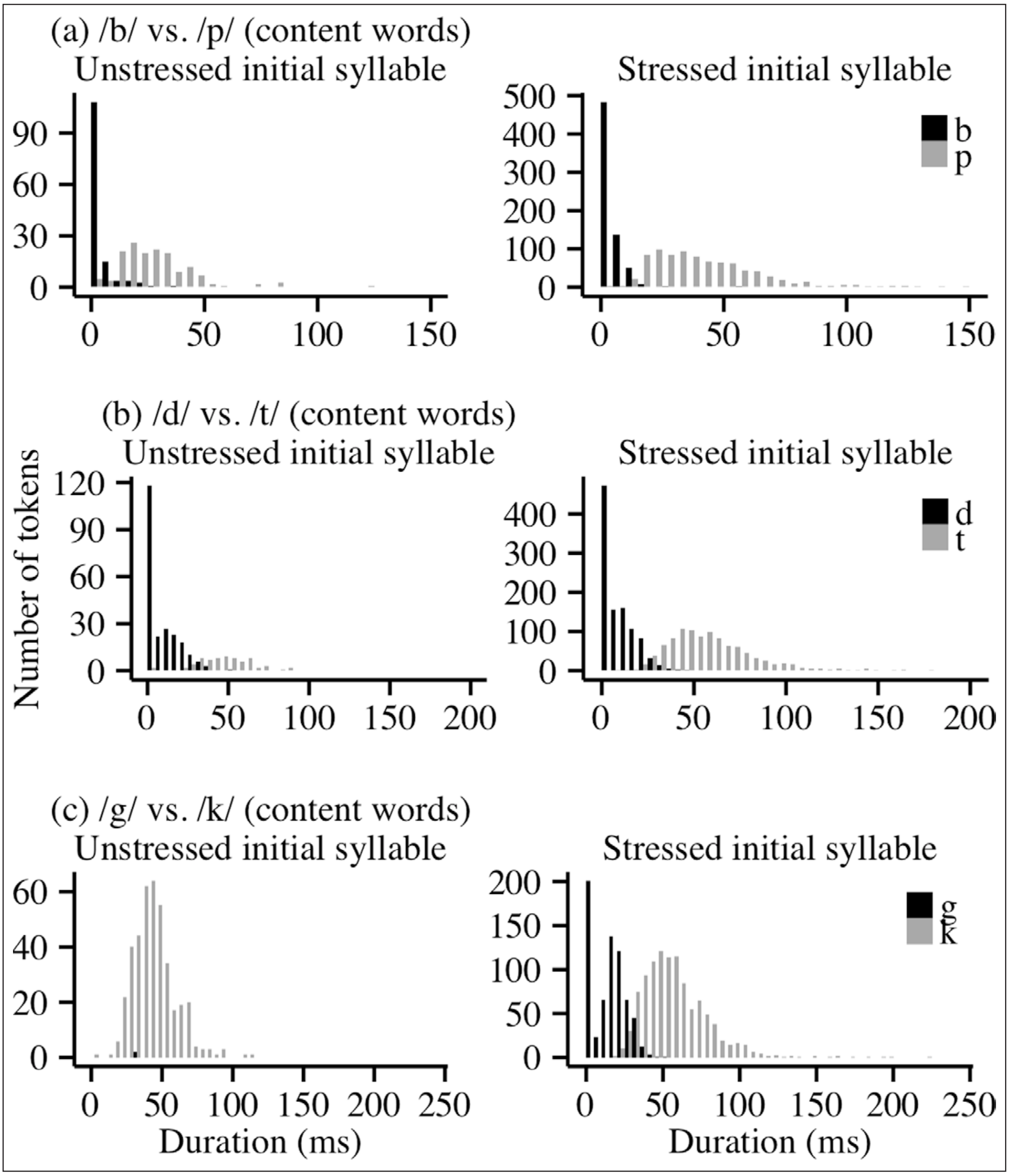

Figure 8: VOT distributions of voiced and voiceless simplex stop phonemes in word-initial syllable of content words without lexical stress (left panels) vs. with lexical stress (right panels) for (a) bilabial, (b) alveolar, and (c) velar places of articulation.

words with non-initial stress. As shown in **Table 3** above, content words with non-initial stress were not many, accounting for only 10% of measured tokens, consistent with Cutler and Carter's (1987) report.

Interestingly, the voicing specifications of stops in word-initial unstressed syllables were largely predictable from the following vowel; 93% of /b/, 97% of /d/, and both of the two tokens of /g/ were followed by /ɪ/, while 99% of /p/, and all instances of /t/ and /k/ were followed by /ə/. When each stop pair was analyzed separately depending on the following vowel, rate-independent category boundaries classified voiced vs. voiceless stops with high accuracy: 99% for /b/-/p/ and /d/-/t/, and 100% for /g/-/k/. These classification accuracies were higher than the overall accuracy of Miller et al.'s (1986) rate-dependent category boundary, though the difference was significant for /g/-/k/ only (/b/-/p/: $\chi^2(1) = 0.8$, $p = 0.38$; /d/-/t/; $\chi^2(1) = 2.6$, $p = .11$; /g/-/k/; $\chi^2(1) = 8.3$, $p = .004$).

3.5 Word-initial homorganic stop pairs in lexically stressed syllables of content words, grouped by the following vowel

The vowel following a stop onset has been reported to affect the VOT of the stop onset, and the locations of perceptual VOT category boundaries between voiced vs. voiceless stop onsets (Higgins et al., 1998; Klatt, 1975; Nearey & Rochet, 1994; Summerfield, 1975, 1981). Though there are some discrepancies in the details, the general finding is that stops tend to be accompanied by a longer VOT when they precede phonologically high vowels than non-high vowels.

At the final and most allophonically-rich level of data control, word-initial homorganic stop phonemes in lexically stressed syllables of content words were grouped by the following vowel phoneme, and separate rate-independent optimal category boundaries were estimated for each group. None of the speakers had a strong regional accent beyond that of their country of origin (England, USA, or Australia). The vowel groups used here therefore reflected broad dialectal differences reported for the three varieties, for example, /ɒ/ for the vowel in *pot* in Anglo English, /ɑ/ for American English, and /ɔ/ for Australian English (Harrington et al., 1997; Wells, 1996). Because only one Australian speaker was represented in our data, vowel phonemes only reported for Australian English were placed with vowels of the same phonological height in other varieties: /ɐ/ and /ɐː/ were grouped with /æ/, and /ʉ/ was grouped with /u/. As there were only several instances of them, Anglo English /əʊ/ and Australian /əʉ/ were grouped with /ɜ/. Diphthongs were grouped based on their initial element; for example, /aɪ/ and /aʊ/ were grouped together. **Table 4** gives the resulting optimal VOT category boundary locations.

These category boundaries produced an overall classification accuracy of 98.4% for /b/-/p/, 98.2% for /d/-/t/, and 97.8% for /g/-/k/ (see **Table 2** above), excluding /tʊ/ and /ki/, whose inclusion would have led to higher classification accuracies, as their voiced counterparts (/dʊ/ and /gi/) did not occur in our data. For all three pairs of stops the classification accuracy achieved here was slightly better than, though not significantly different from, the 97.6% accuracy achieved by Miller et al.'s (1986) rate-dependent category boundary for /bi/-/pi/ (/b/-/p/: $\chi^2(1) = 1.41$, $p = .24$; /d/-/t/: $\chi^2(1) = .83$, $p = .36$; /g/-/k/: $\chi^2(1) = .01$, $p = .93$). Compared to the previous level of data control, the classification accuracy improved significantly for /d/-/t/ and /g/-/k/ ($\chi^2(1) = 8.28$, $p = .004$; $\chi^2(1) = 27.6$, $p < .001$) but not for /b/-/p/ ($\chi^2(1) = 1.67$, $p = .20$). We do not know why the following vowels affected the boundary location for /d/-/t/ and /g/-/k/ more than for /b/-/p/, but Nearey and Rochet (1994) report similar findings in perception.

Based on previous studies on perceptual VOT category boundary locations (Nearey & Rochet, 1994; Summerfield, 1975, 1981), we had expected larger VOT values at the category boundary in high vowel contexts and smaller values in low vowel contexts, particularly for alveolar and velar stops. This appeared to be true of our data to some extent, but the differences in VOT boundary location between vowel contexts seemed more directly linked to the difference in the relative frequency of occurrences of voiced vs. voiceless stop phonemes between vowel contexts than to the phonological height of the following vowel.

Figure 9 shows the relationship between the estimated optimal VOT category boundary location in various vowel contexts, and the difference in logarithmic token frequency between voiced vs. voiceless members of each homorganic stop pair in each context. As we saw in **Table 4**, the range of estimated boundary locations was large in some cases. To ensure some degree of reliability of the estimated boundary location, we only used

Following vowel	Phonological height	/b/-/p/	/d/-/t/	/g/-/k/
/i/	High	6–9 (*n* = 327)	28–30 (*n* = 117)	(see Note)
/ɪ/		16–18 (*n* = 200)	32–36 (*n* = 246)	42 (*n* = 98)
/u/		13–24; no overlap (*n* = 10)	36–40 (*n* = 451)	26–78; no overlap (*n* = 6)
/ʊ/		12–21 (*n* = 76)	(see Note)	39–51; no overlap (*n* = 91)
/e/	Mid	13–15 (*n* = 137)	26–27 (*n* = 296)	31 (*n* = 189)
/ɛ/		15–16 (*n* = 63)	19–21 (*n* = 95)	37–39 (*n* = 101)
/ɜ/		12 (*n* = 115)	17–26; no overlap (*n* = 77)	36–39 (*n* = 144)
/o/		9–27; no overlap (*n* = 14)	1–34; no overlap (*n* = 12)	41–43 (*n* = 91)
/ɔ/		14–21 (*n* = 89)	21–23; no overlap (*n* = 69)	25–28; no overlap (*n* = 96)
/ʌ/		11–14; no overlap (*n* = 35)	33–36; no overlap (*n* = 93)	19–21 (*n* = 238)
/æ/	Low	14–15 (*n* = 257)	26–33 (*n* = 81)	18 (*n* = 122)
/ɒ/		12–14; no overlap (*n* = 48)	18–31; no overlap (*n* = 19)	24–29 (*n* = 156)
/ɑ/		14–16 (*n* = 148)	19 (*n* = 116)	24 (*n* = 254)
/a/		9–23; no overlap (*n* = 17)	21 (*n* = 324)	19–22 (*n* = 114)

Table 4: Rate-independent optimal category boundary locations for homorganic stops in stressed word-initial syllables, grouped by the following vowel. A range of values represents maximum classification accuracy found at multiple steps. It corresponds to a gap in distribution, where voiced vs. voiceless VOT did not overlap.

Note. /tʊ/ and /ki/ were excluded from analysis, as no words in the data started with /dʊ/ or /gi/.

(1) boundary locations that could be estimated within 1 ms and (2) the midpoint of the estimated range when the boundary location could be defined within 5 ms or less. **Figure 9** suggests that the more frequently a voiceless stop onset occurred relative to its voiced counterpart before a given vowel (the larger the value on the x-axis), the smaller the estimated VOT boundary was. According to Pearson correlation tests, this correlation was significant for all three stop pairs (/b/-/p/: $r = -.81$; /d/-/t/: $r = -.75$; /g/-/k/: $r = -.94$; all $ps < .02$). At the same time, the results exhibited a tendency consistent with the observation of boundaries at a large VOT value for high vowel contexts and a small VOT value for low vowel contexts for alveolar and velar stops.

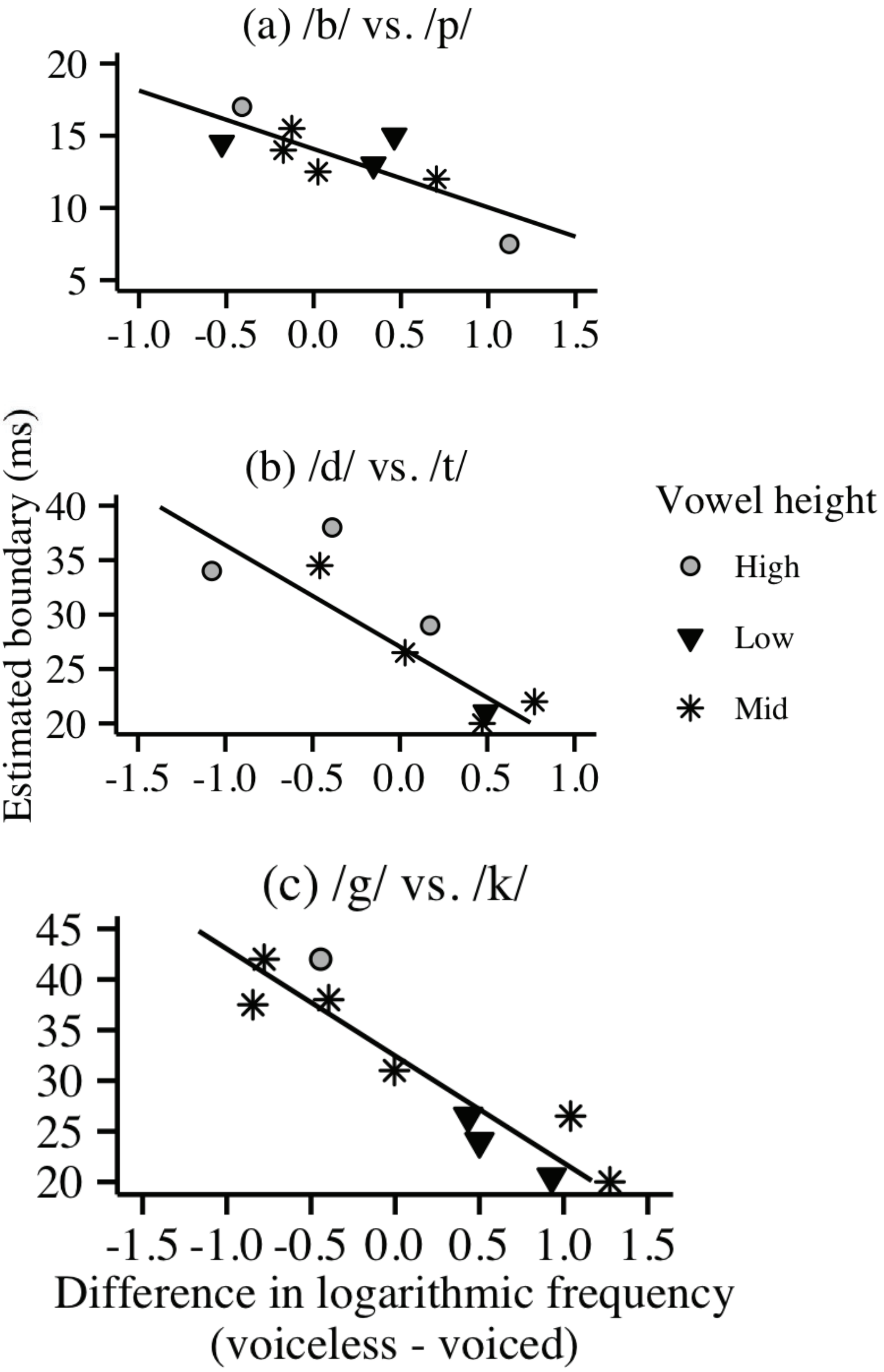

Figure 9: Relationship between estimated VOT category boundary between voiced vs. voiceless stops in various vowel contexts, and difference in their logarithmic token frequency in each context.

Recall that the optimal category boundary location was determined on the basis of maximum classification accuracy for voiced and voiceless stops combined (see Section 2.3). All else being equal, the more frequent a phoneme is within the region of distributional overlap, the greater the phoneme's contribution to the calculation of overall classification accuracy; this pushes the optimal category boundary away from that phoneme. The results above suggest that for alveolar and velar places of articulation, voiceless stops tend to occur less frequently than their voiced counterparts in high vowel contexts and more frequently in low vowel contexts in word-initial position in English, and the category boundary is pushed in different directions depending on the vowel context, towards the less frequent voicing category. Notice that this produces an effect similar to the frequency effects on the perceptual category boundary location between phonemes reported by Kataoka and Johnson (2007).

It is worth noting, in addition, that the total range of VOT values for voiced vs. voiceless stops differed between vowel contexts in our data, in a manner consistent with the observed difference in boundary location between vowel contexts.

First, the more frequently a voiced stop occurred before a given vowel, the longer its maximum VOT tended to be, likely contributing to a larger VOT value at the optimal category boundary. For /d/ and /g/, Pearson correlation tests indicated a significant positive correlation between each stop's logarithmic token frequencies and maximum VOT values in different vowel contexts (/d/: $r = .85$; /g/: $r = .83$; both $ps < .001$). For /b/, which had a large outlier, the correlation was not significant in a Pearson test ($r = .34$, $p = .24$) but significant in a Spearman test, which is robust to the presence of outliers ($r_s = .64$, $p = .01$).

Conversely, the more frequently a voiceless stop occurred before a given vowel, the shorter its minimum VOT was, likely contributing to a smaller VOT value at the optimal category boundary. Pearson tests indicated a significant negative correlation between each voiceless stop's logarithmic token frequencies and minimum VOT values in different vowel contexts (/p/: $r = -.73$, $p = .003$; /t/: $r = -.63$, $p = .02$; /k/: $r = -.74$, $p = .002$). The picture was similar for voiced vs. voiceless stops in lexically unstressed word-initial syllables discussed in Section 3.4, although the observations in infrequent vowel contexts were very small in number.

The above observation itself does not necessarily imply different underlying VOT distributions for a given stop phoneme in more vs. less frequent vowel contexts, as the likelihood of obtaining extreme values from the same underlying distribution increases with sample size.[8] However, the results of Fricke's (2013) recent study of voiceless stop onsets in English spontaneous speech point to the possibility that underlying VOT distributions themselves may in fact differ between more vs. less frequent contexts in which the stop occurs. At any rate, our observation does suggest that in real-life conversation we are more likely to encounter extreme VOT values for a stop in a more frequent vowel context. This can also push the perceptual category boundary location towards the less frequent phoneme.

4 Discussion

In this study we examined the effectiveness of the rate-independent VOT category boundary for word-initial English voiced vs. voiceless stop phonemes in unscripted conversational speech. Articulation rate varied in our data in, we assume, a natural way; variation in articulation rate was certainly observable across and within speakers, at a qualitative level. Yet, our data suggested that there is no compelling need for listeners to normalize

[8] We thank Holger Mitterer for pointing this out.

perceptual VOT category boundary locations for word-initial voiced vs. voiceless stops in accordance with articulation rate, supporting Kessinger and Blumstein's (1997) proposal.

Rate-independent optimal VOT category boundaries classified all three pairs of homorganic, word-initial voiced vs. voiceless categories in content words at accuracy comparable to (or better than) Miller et al.'s (1986) rate-dependent category boundary, when the stops were analyzed separately depending on the presence of lexical stress and the following vowel phoneme. The inclusion of function words led to lower classification accuracy, but in our analysis of /du/ vs. /tu/ (*do* vs. *to*, *too*, and *two*), classification accuracy did not much improve by adopting rate-dependent category boundaries (using the mean segment duration of the preceding word as an index of local articulation rate). Classification accuracy improved significantly, however, by postulating the short duration of /u/ (measured from the onset of voicing) in /tu/ as an additional cue to the /du/-/tu/ opposition. Crucially, the short duration of /u/ in /tu/ relative to /du/ was found across our measure of articulation rate and could not be ascribed to fast articulation rates of /tu/.

Thus, the lack of large shifts in perceptual VOT category boundary locations for word-initial stops in previous rate normalization studies can be seen to reflect the listeners' experience of temporal regularities of speech they normally encounter. The small but consistent shifts in VOT category boundary locations found in these perception studies are perhaps better interpreted as arising from cue integration (Toscano & McMurray, 2012) or general auditory (proximal or distal) contrast effects (Diehl & Walsh, 1989; Holland & Lockhead, 1968; Pisoni et al., 1983).

The point we wish to make here is simple: If rate normalization reflects the temporal regularities of the ambient language, then we have little reason to expect such a process where the language does not require it. For example, the durational distributions of singleton vs. geminate sonorants in Cypriot Greek are reported to be well separated across different rates of articulation (Arvaniti, 1999). We therefore do *not* expect Cypriot Greek speakers to shift the perceptual category boundary for the contrasts with a change in articulation rate. On the other hand, our data suggest that English function words are less likely to be reduced under slow articulation rates. We therefore expect English-speaking listeners to less often report reduced function words in ambiguous speech stimuli when surrounding speech is slow (Baese-Berk et al., 2014; Dilley & Pitt, 2010).

In the absence of relevant information, we have little to say about rate-dependent shifts in perceptual category boundaries reported for other contrasts, for example, the /b/-/w/ distinction in English (e.g., Miller & Liberman, 1979) or consonant and vowel quantity in other languages (Icelandic: Pind, 1995; Japanese: Fujisaki et al., 1975; see also Hirata & Lambacher, 2004).

That said, we think that a lack of need for rate normalization may be found for more contrasts, as durational variation arising from different articulation rates is presumably more malleable than other aspects of speech that are thought to necessitate perceptual normalization such as formant frequencies associated with vocal tract length (but see Johnson, 1997). Even though listeners can apparently cope with such situations (e.g., Ladefoged & Broadbent, 1957; Syrdal & Gopal, 1979), perceptual normalization is not cost-free (Mullennix et al., 1989, 2002; Nakai, 2013; Sommers et al., 1994).

While the rate-dependent VOT category boundary did not seem to significantly improve the classification accuracy of word-initial voiced vs. voiceless stop phonemes, vowel-dependent VOT category boundaries did. The vowel-dependent category boundary produced an optimal category boundary with a relatively small VOT value in a vowel context where voiceless stops were more frequent than voiced stops, and a large VOT value where voiced stops were more frequent than voiceless stops. This resulted in a VOT boundary location at a generally larger VOT value for phonologically high vowel contexts and a smaller VOT value for low vowel contexts for alveolar and velar stops, as previously reported in perception studies (Nearey & Rochet, 1994; Summerfield, 1975, 1981).

If shifting a VOT category boundary depending on articulation rate were costly to the perception mechanism, would vowel-dependent category boundaries not be costly too? With a caveat that we did not conduct any perception experiments, it is plausible that the listeners use categories other than phonemes as basic units in their analysis of incoming speech where it makes sense to do so, as proposed by Reinisch et al. (2014). Rather than normalizing a phoneme-based category boundary depending on the following vowel, the listeners may look for units larger than a phoneme (e.g., CV) and use category boundaries specific to these units.

Notice that this scenario is compatible with the observation that *do* and *to* can be largely distinguished on the basis of the duration of /u/, where VOT is neutralized. The scenario also sits well with previous findings that the listeners interpret acoustic cues to the place of articulation of stop onsets (e.g., burst, formant transition) differently depending on the following vowel (Cooper et al., 1952; Dorman et al., 1977) and that the listeners can largely identify the following vowel from the brief period immediately after the stop release (Blumstein & Stevens, 1980). Arguably, structural or phonological contexts such as the following vowel in the same word are different in kind from contexts such as articulation rate. Vowels, being discrete units that constitute a part of a word, can more readily serve as a part of the basic unit of analysis in speech perception, unlike articulation rate, which forms continuity and is presumably unspecified in the lexicon.

Another point we wish to stress is that the greater overlap in the VOT distributions of voiced vs. voiceless stops in function words compared to content words with initial lexical stress was more directly linked to the difference in their overall frequencies, rather than their different lexical statuses. In our data, function words overall had a higher token frequency than did content words with initial lexical stress (calculated for lemmas), as shown in **Figure 10**. And, as shown in **Figure 11**, at the initial level of analysis, the more frequent content and function words were, the more likely they were to have a VOT that fell on the opposite side of the optimal category boundary, producing an overlap between voiced vs. voiceless VOT distributions (content words: $r_s = .32, p < .001$; function words: $r_s = .78, p < .001$). For example, word-initial stops in frequent content words like light verbs (e.g., *do*, *get*, *give*) were more likely to have a VOT that fell on the opposite side of the optimal category boundary than were stops in infrequent function words like *per*. Thus, the relatively high misclassification rate for /du/-/tu/ found in Section 3.3 may be at least partially ascribed to the high frequencies of /du/ and /tu/, especially *do* and *to*.

The foregoing observations of the relationships between frequency and VOT values, in relation to vowel contexts and lexical status, may be conceived in terms of predictability, which is highly correlated to frequency (e.g., Bell et al., 2009). A growing body of studies report phonetic and phonological reduction of frequent and/or predictable words and segments, which cannot simply be attributed to fast articulation rates (e.g., Aylett & Turk, 2006; Baese-Berk & Goldrick, 2009; Baran et al., 1977; Bybee, 2000; Ernestus, 2000; Fosler-Lussier & Morgan, 1999; Frank & Jaeger, 2008; Fricke, 2013; Gahl et al., 2012; Jurafsky et al., 2001; Lieberman, 1963; Munson, 2007). Where it is predictable, voicing specifications may not need to be as clearly signaled by VOT for successful communication, considering the facilitative effects of listener expectations on word recognition (e.g., Rubenstein & Pollack, 1963) and listener tolerance for acoustic mismatches in reduced speech (Brouwer et al., 2012). (Of course, other cues to the stop's voicing specification may also be present, as was the case for *to* with a very short VOT.)

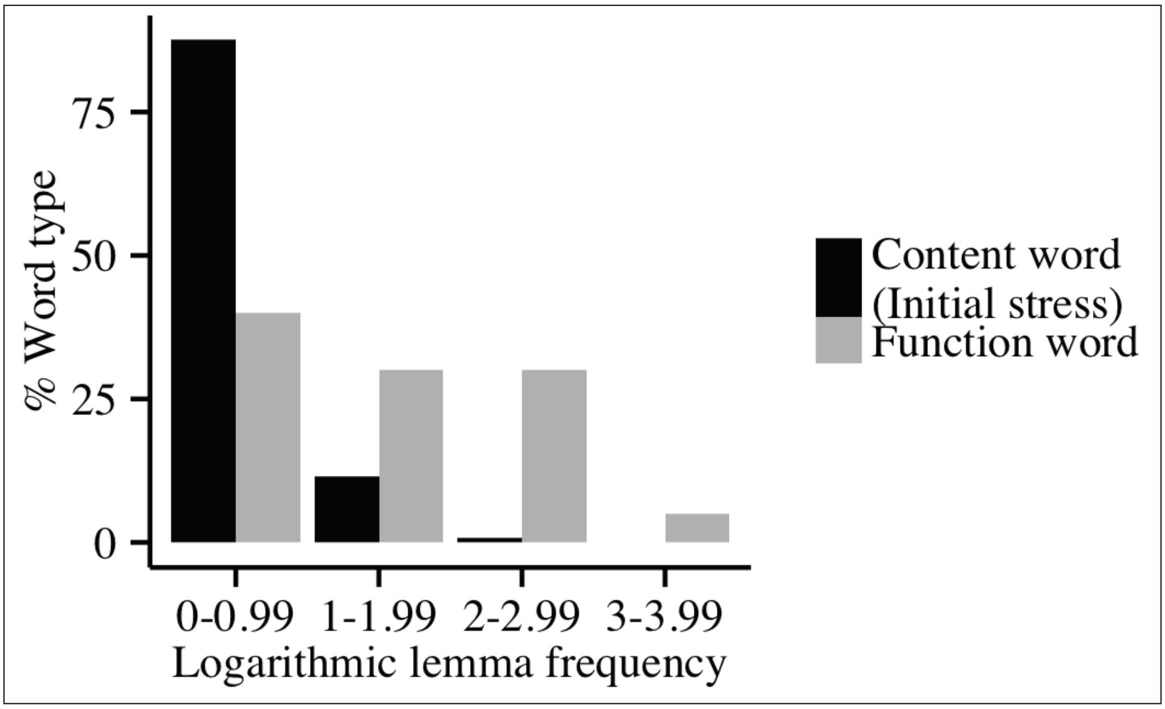

Figure 10: Lemma frequency of content words (with word-initial lexical stress) vs. function words.

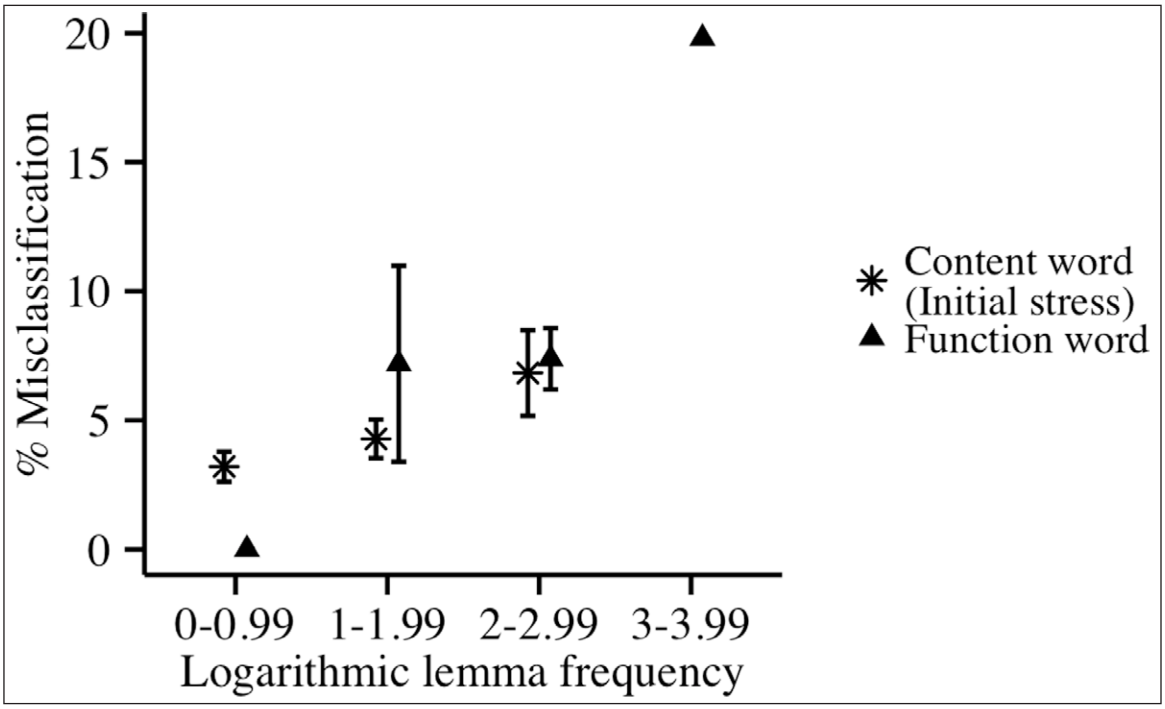

Figure 11: Relationship between lemma frequency and misclassification rates for content words (with initial lexical stress) and function words, at the initial level of data control. Error bars indicate the standard error of the mean.

This is not to suggest that speakers consciously produce unpredictable speech segments more clearly and predictable speech segments less clearly. Clear enunciation of words and enhanced segmental contrasts (including those signaled by VOT) can result from listener-oriented considerations given by the speaker (Bradlow, 2002), but this is not always true

(Baese-Berk & Goldrick, 2009; Bard et al., 1988; Gahl et al., 2012). That is, ambiguous renditions of predictable segments and words are not necessarily a product of speakers' conscious production strategy.

A relatively short VOT for voiceless stops in frequent words and vowel contexts can arise from ease of lexical access on the speaker's part as well as ease of articulation (Balota et al., 1989; Bard et al., 2000; Bell et al., 2009; Fricke, 2013; Gahl et al., 2012; Munson, 2007). This account, however, would not predict a relatively long VOT for voiced stops in frequent words and vowel contexts, for the ease of production is associated with reduced duration.

Another possibility, though not mutually exclusive from the above, is that clarity of enunciation of some segments/words reflects their phonetic representations. Since Norris et al.'s (2003) influential work, several studies have shown that perceptual category boundary locations for segmental contrasts are affected by ambiguous sounds when the ambiguous sounds are recognized by the listener as a part of a legitimate word (Clarke & Luce, 2005; Eisner et al., 2013; Kraljic & Samuel, 2005; Maye et al., 2008).

Conceivably, phonetic representations of less frequent segments and words are more likely to be shaped by their clear enunciations, as unpredictable segments/words produced with ambiguous pronunciations are less likely to lead to immediate recognition (see Pierrehumbert [2002] for a similar proposal in relation to lexical neighborhood density, and Wedel [2006] in relation to diachronic maintenance of phonemic contrasts). If so, then position and/or context-sensitive representations of phonemes (Dahan & Mead, 2010; Eisner et al., 2013; Mitterer et al., 2013) would predict more distinct phonetic representations of contrasts in positions and contexts where segments are less predictable, and word-specific phonetic representations (Bybee, 2000; Johnson, 2004; Klatt, 1979; Pierrehumbert, 2002; Wedel, 2006) would predict more distinct phonetic representations of less predictable words.

As a final note, the range of VOT values used to signal voiced vs. voiceless stop phonemes can differ between speakers of the same language, depending on factors such as gender and geographical origin (Docherty et al., 2011; Oh, 2011; Scobbie, 2006). It is currently unclear, however, to what extent such factors affect category boundary locations for voicing contrasts along acoustic cues like VOT, as past production studies focused on phonetic targets rather than category boundaries. In perception studies listener sensitivity to social-indexical acoustic variation has been shown, most notably for English vowels (e.g., Hay et al., 2006; Niedzielski, 1999). Listener sensitivity to social-indexical variation in VOT has also been shown, but shifts in perceptual VOT boundary locations for word-initial stops induced through manipulation of speaker gender (Toscano, 2011) or speaker adaptation training (Clarke & Luce, 2005; VanDam, 2007) appear very small in magnitude.[9] We welcome further studies on the role of inter-personal and social-indexical factors in the production and perception of speech segments from various angles.

[9] In our data optimal VOT category boundary locations for content words differed, for example, between male and female speakers only by 1 ms for /b/-/p/ and 2 ms for /g/-/k/ in the expected direction (larger VOT values at the boundary for female speakers). The boundary location differed between male and female speakers by 1 ms for /d/-/t/ in the opposite direction. These should be interpreted with caution, because our male vs. female speakers were not homogenous with regard to other social-indexical aspects (e.g., geographical origin), which may have affected the estimations, in addition to other factors we did not control. We note, however, that the above boundary shifts for /b/-/p/ and /g/-/k/ are comparable, in magnitude as well as direction, with the gender-related VOT category boundary shift observed in Toscano's (2011) perception study.

Acknowledgements

We thank the British Broadcasting Corporation for permission to use the audio recordings of their program for this study. We are grateful to Holger Mitterer, Mirjam Ernestus, Jennifer Cole, and two anonymous reviewers for their help in improving this paper.

Competing Interests

The authors declare that they have no competing interests.

References

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

Arvaniti, A. 1999. Effects of speaking rate on the timing of single and geminate sonorants. *Proceedings of the XIVth International Congress of Phonetic Sciences*, 599–602.

Aylett, M., & Turk, A. E. 2006. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America*, *119*(5), 3048–3058. DOI: http://dx.doi.org/10.1121/1.2188331

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

Baese-Berk, M. M., Heffner, C. C., Dilley, L. C., Pitt, M. A., Morrill, T. H., & McAuley, J. D. 2014. Long-term temporal tracking of speech rate affects spoken-word recognition. *Psychological Science*, *25*(8), 1546–1553. DOI: http://dx.doi.org/10.1177/0956797614533705

Balota, D. A., Boland, J. E., & Shields, L. W. 1989. Priming in pronunciation: Beyond pattern recognition and onset latency. *Journal of Memory and Language*, *28*(1), 14–36. DOI: http://dx.doi.org/10.1016/0749-596X(89)90026-0

Baran, J. A., Laufer, M. Z., & Daniloff, R. 1977. Phonological contrastivity in conversation: A comparative study of voice onset time. *Journal of Phonetics*, *5*, 339–350.

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. 2000. Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, *42*(1), 1–22. DOI: http://dx.doi.org/10.1006/jmla.1999.2667

Bard, E. G., Shillcock, R. C., & Altmann, G. T. 1988. The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception and Psychophysics*, *44*(5), 395–408. DOI: http://dx.doi.org/10.3758/BF03210424

Beckman, J., Helgason, P., McMurray, B., & Ringen, C. 2011. Rate effects on Swedish VOT: Evidence for phonological overspecification. *Journal of Phonetics*, *39*(1), 39–49. DOI: http://dx.doi.org/10.1016/j.wocn.2010.11.001

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111. DOI: http://dx.doi.org/10.1016/j.jml.2008.06.003

Blumstein, S. E., & Stevens, K. N. 1980. Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America*, *67*(2), 648–662. DOI: http://dx.doi.org/10.1121/1.383890

Boersma, P., & Weenink, D. 2012. Praat: Doing phonetics by computer (Version 5.3.23). Retrieved from http://www.fon.hum.uva.nl/praat/

Bradlow, A. R. 2002. Confluent talker-and listener-oriented forces in clear speech production. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology 7* (pp. 241–273). Berlin: Mouton de Gruyter. DOI: http://dx.doi.org/10.1515/9783110197105.241

British Broadcasting Corporation. 2010. *Radio technical standards: BWAV specification*. Retrieved from http://www.bbc.co.uk/guidelines/dq/pdf/radio/delivery_requirements_ mar_10.pdf.

Brouwer, S., Mitterer, H., & Huettig, F. 2012. Speech reductions change the dynamics of competition during spoken word recognition. *Language and Cognitive Processes*, *27*(4), 539–571. DOI: http://dx.doi.org/10.1080/01690965.2011.555268

Bybee, J. 2000. The phonology of the lexicon: Evidence from lexical diffusion. In M. Barlow & S. Kemmer (Eds.), *Usage-based models of language* (pp. 65–85). Stanford, CA: Center for the Study of Language and Information.

Cho, T., & Ladefoged, P. 1999. Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics*, *27*(2), 207–229. DOI: http://dx.doi.org/10.1006/jpho.1999.0094

Clarke, C., & Luce, P. 2005. Perceptual adaptation to speaker characteristics: VOT boundaries in stop voicing categorization. *Proceedings of ISCA Workshop on Plasticity in Speech Perception*, 23–26.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. 1952. Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America*, *24*(6), 597–606. DOI: http://dx.doi.org/10.1121/1.1906940

Cutler, A., & Carter, D. M. 1987. The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*(3–4), 133–142. DOI: http://dx.doi.org/10.1016/0885-2308(87)90004-0

Dahan, D., & Mead, R. L. 2010. Context-conditioned generalization in adaptation to distorted speech. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(3), 704–728. DOI: http://dx.doi.org/10.1037/a0017449

Diehl, R. L., & Walsh, M. A. 1989. An auditory basis for the stimulus-length effect in the perception of stops and glides. *Journal of the Acoustical Society of America*, *85*(5), 2154–2164. DOI: http://dx.doi.org/10.1121/1.397864

Dilley, L. C., & Pitt, M. A. 2010. Altering context speech rate can cause words to appear or disappear. *Psychological Science*, *21*(11), 1664–1670. DOI: http://dx.doi.org/10.1177/0956797610384743

Docherty, G. J., Watt, D., Llamas, C., Hall, D., & Nycz, J. 2011. Variation in voice onset time along the Scottish-English border. *Proceedings of the XVIIth International Congress of Phonetic Sciences*, 591–594.

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. 1977. Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception and Psychophysics*, *22*(2), 109–122. DOI: http://dx.doi.org/10.3758/BF03198744

Eisner, F., Melinger, A., & Weber, A. 2013. Constraints on the transfer of perceptual learning in accented speech. *Frontiers in Psychology*, *4*, 1–9. DOI: http://dx.doi.org/10.3389/fpsyg.2013.00148

Ernestus, M. 2000. *Voice assimilation and segment reduction in casual Dutch: A corpus-based study of the phonology-phonetics interface*. Utrecht, Netherlands: Landelijke Onderzoekschool Taalwetenschap.

Field, A. P. 2005. *Discovering statistics using SPSS* (2nd ed.). London: Sage Publications.

Fosler-Lussier, E., & Morgan, N. 1999. Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication*, *29*, 137–158. DOI: http://dx.doi.org/10.1016/S0167-6393(99)00035-7

Frank, A., & Jaeger, T. F. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, 933–938.

Fricke, M. 2013. *Phonological encoding and phonetic duration*. Doctoral dissertation. Available from ProQuest Dissertations & Theses database (UMI No. 3616451).

Fujisaki, H., Nakamura, K., & Imoto, T. 1975. Auditory perception of duration of speech and non-speech stimuli. In G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 197–219). London: Academic Press. DOI: http://dx.doi.org/10.1016/B978-0-12-248550-3.50017-9

Gahl, S., Yao, Y., & Johnson, K. 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language, 66*(4), 789–806. DOI: http://dx.doi.org/10.1016/j.jml.2011.11.006

Gaitenby, J. 1965. The elastic word. *Haskins Laboratories Status Report on Speech Research, 2*, 3–1.

Green, K. P., & Miller, J. L. 1985. On the role of visual rate information in phonetic perception. *Perception and Psychophysics, 38*(3), 269–276. DOI: http://dx.doi.org/10.3758/BF03207154

Green, K. P., Stevens, E. B., & Kuhl, P. K. 1994. Talker continuity and the use of rate information during phonetic perception. *Perception and Psychophysics*, *55*(3), 249–260. DOI: http://dx.doi.org/10.3758/BF03207596

Haggard, M., Summerfield, Q., & Roberts, M. 1981. Psychoacoustical and cultural determinants of phoneme boundaries: Evidence from trading F_0 cues in the voiced–voiceless distinction. *Journal of Phonetics*, *9*, 49–62.

Harrington, J., Cox, F., & Evans, Z. 1997. An acoustic phonetic study of broad, general, and cultivated Australian English vowels. *Australian Journal of Linguistics, 17*(2), 155–184. DOI: http://dx.doi.org/10.1080/07268609708599550

Hay, J., Nolan, A., & Drager, K. 2006. From fush to feesh: Exemplar priming in speech perception. *Linguistic Review 23*(3), 351–379. DOI: http://dx.doi.org/10.1515/TLR.2006.014

Higgins, M. B., Netsell, R., & Schulte, L. 1998. Vowel-related differences in laryngeal articulatory and phonatory function. *Journal of Speech, Language, and Hearing Research, 41*(4), 712–724. DOI: http://dx.doi.org/10.1044/jslhr.4104.712

Hirata, Y., & Lambacher, S. G. 2004. Role of word-external contexts in native speakers' identification of vowel length in Japanese. *Phonetica, 61*, 177–200. DOI: http://dx.doi.org/10.1159/000084157

Holland, M. K., & Lockhead, G. R. 1968. Sequential effects in absolute judgments of loudness. *Perception and Psychophysics, 3*(6), 409–414. DOI: http://dx.doi.org/10.3758/BF03205747

Johnson, K. 1997. Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego, CA: Academic Press.

Johnson, K. 2004. Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (Eds.), *Spontaneous speech: Data and analysis.* (pp. 29–54). Tokyo: National International Institute for Japanese Language.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. 2001. Evidence from reduction in lexical production. In J. Bybee (Ed.), *Frequency and the emergence of linguistic structure* (pp. 229–254). Amsterdam: John Benjamins. DOI: http://dx.doi.org/10.1075/tsl.45.13jur

Kataoka, R., & Johnson, K. 2007. *Frequency effects in cross-linguistic stop place perception: A case of /t/ - /k/ in Japanese and English.* Retrieved from University of California, Berkley, Phonology Lab Annual Report Web site: http://linguistics.berkeley.edu/phonlab/annual_report/documents/2007/Kataoka_Johnson.pdf

Kessinger, R. H., & Blumstein, S. E. 1997. Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics*, *25*(2), 143–168. DOI: 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*(2), 117–128. DOI: http://dx.doi.org/10.1006/jpho.1997.0069

Kidd, G. R. 1989. Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(4), 736–748. DOI: http://dx.doi.org/10.1037/0096-1523.15.4.736

Klatt, D. H. 1975. Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research, 18*(4), 686–706. DOI: http://dx.doi.org/10.1044/jshr.1804.686

Klatt, D. H. 1979. Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, *7*, 279–312.

Kraljic, T., & Samuel, A. G. 2005. Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178. DOI: http://dx.doi.org/10.1016/j.cogpsych.2005.05.001

Ladefoged, P., & Broadbent, D. E. 1957. Information conveyed by vowels. *Journal of the Acoustical Society of America*, *29*(1), 98–104. DOI: http://dx.doi.org/10.1121/1.1908694

Lehiste, I. 1972. The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America*, *51*(6B), 2018–2024. DOI: http://dx.doi.org/10.1121/1.1913062

Lieberman, P. 1963. Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, *6*, 172–187.

Lieberman, P., & Blumstein, S. E. 1988. *Speech physiology, speech perception, and acoustic phonetics*. Cambridge, England: Cambridge University Press. DOI: http://dx.doi.org/10.1017/CBO9781139165952

Lisker, L., & Abramson, A. S. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422. DOI: http://dx.doi.org/10.1080/00437956.1964.11659830

Lisker, L., & Abramson, A. S. 1967. Some effects of context on voice onset time in English stops. *Language and Speech*, *10*(1), 1–28.

Lisker, L., & Abramson, A. S. 1970. The voicing dimension: Some experiments in comparative phonetics. *Proceedings of the VIth International Congress of Phonetic Sciences*, 563–567.

Magloire, J., & Green, K. P. 1999. A cross-language comparison of speaking rate effects on the production of voice onset time in English and Spanish. *Phonetica*, *56*(3–4), 158–185. DOI: http://dx.doi.org/10.1159/000028449

Maye, J., Aslin, R. N., & Tanenhaus, M. K. 2008. The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32*(3), 543–562. DOI: http://dx.doi.org/10.1080/03640210802035357

Miller, J. L., & Dexter, E. R. 1988. Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 369–378. DOI: http://dx.doi.org/10.1037/0096-1523.14.3.369

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*(1–3), 106–115. DOI: http://dx.doi.org/10.1159/000261764

Miller, J. L., Grosjean, F., & Lomanto, C. 1984. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, *41*(4), 215–225. DOI: http://dx.doi.org/10.1159/000261728

Miller, J. L., & Liberman, A. M. 1979. Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception and Psychophysics*, *25*(6), 457–465. DOI: http://dx.doi.org/10.3758/BF03213823

Mitterer, H., Scharenborg, O., & McQueen, J. M. 2013. Phonological abstraction without phonemes in speech perception. *Cognition*, *129*(2), 356–361. DOI: http://dx.doi.org/10.1016/j.cognition.2013.07.011

Mullennix, J. W., Bihon, T., Bricklemyer, J., Gaston, J., & Keener, J. M. 2002. Effects of variation in emotional tone of voice on speech perception. *Language and Speech*, *45*(3), 255–283. DOI: http://dx.doi.org/10.1177/00238309020450030301

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. 1989. Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, *85*(1), 365–378. DOI: http://dx.doi.org/10.1121/1.397688

Munson, B. 2007. Lexical access, lexical representation, and vowel production. In J. Cole & J. I. Hualde (Eds.), *Laboratory phonology 9* (pp. 201–228). Berlin: Walter de Gruyter.

Nagao, K., & de Jong, K. 2007. Perceptual rate normalization in naturally produced rate-varied speech. *Journal of the Acoustical Society of America*, *121*(5), 2882–2898. DOI: http://dx.doi.org/10.1121/1.2713680

Nakai, S. 2013. An explanation for phonological word-final vowel shortening: Evidence from Tokyo Japanese. *Laboratory Phonology*, *4*(2), 513–553. DOI: http://dx.doi.org/10.1515/lp-2013-0016

Nakai, S., & Turk, A. E. 2011. Separability of prosodic phrase boundary and phonemic information. *Journal of the Acoustical Society of America*, *129*(2), 966–976. DOI: http://dx.doi.org/10.1121/1.3514419

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. DOI: http://dx.doi.org/10.1017/S0025100300004965

Newman, R. S., & Sawusch, J. R. 1996. Perceptual normalization for speaking rate: Effects of temporal distance. *Perception and Psychophysics*, *58*(4), 540–560. DOI: http://dx.doi.org/10.3758/BF03213089

Niedzielski, N. 1999. The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, *18*(1), 62–85. DOI: http://dx.doi.org/10.1177/0261927X99018001005

Nooteboom, S. G. 1979. Complex control of simple decisions in the perception of vowel length. *Proceedings of the IXth International Congress of Phonetic Sciences*, *2*, 298–304.

Norris, D., McQueen, J. M., & Cutler, A. 2003. Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. DOI: http://dx.doi.org/10.1016/S0010-0285(03)00006-9

Oh, E. 2011. Effects of speaker gender on voice onset time in Korean stops. *Journal of Phonetics*, *39*(1), 59–67. DOI: http://dx.doi.org/10.1016/j.wocn.2010.11.002

Pierrehumbert, J. 2002. Word-specific phonetics. In C. Gussenhoven, T. Rietveld, & N. Warner (Eds.), *Laboratory phonology 7* (pp. 101–139). Berlin: Mouton de Gruyter. DOI: http://dx.doi.org/10.1515/9783110197105.101

Pind, J. 1995. Speaking rate, voice-onset time, and quantity: The search for higher-order invariants for two Icelandic speech cues. *Perception and Psychophysics*, *57*(3), 291–304. DOI: http://dx.doi.org/10.3758/BF03213055

Pisoni, D. B., Carrell, T. D., & Gans, S. J. 1983. Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception and Psychophysics*, *34*(4), 314–322. DOI: http://dx.doi.org/10.3758/BF03203043

Reinisch, E., & Sjerps, M. J. 2013. The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics*, *41*(2), 101–116. DOI: http://dx.doi.org/10.1016/j.wocn.2013.01.002

Reinisch, E., Wozny, D. R., Mitterer, H., & Holt, L. L. 2014. Phonetic category recalibration: What are the categories? *Journal of Phonetics*, *45*, 91–105. DOI: http://dx.doi.org/10.1016/j.wocn.2014.04.002

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. doi:10.1177/002383097902200207

Rubenstein, H., & Pollack, I. 1963. Word predictability and intelligibility. *Journal of Verbal Learning and Verbal Behavior*, *2*(2), 147–158. DOI: http://dx.doi.org/10.1016/S0022-5371(63)80079-1

Schiavetti, N., Whitehead, R. L., Metz, D. E., Whitehead, B., & Mignerey, M. 1996. Voice onset time in speech produced during simultaneous communication. *Journal of Speech, Language, and Hearing Research*, *39*(3), 565–572. DOI: http://dx.doi.org/10.1044/jshr.3903.565

Scobbie, J. M. 2006. Flexibility in the face of incompatible English VOT systems. In L. Goldstein, D. H. Whalen, & C. T. Best (Eds.), *Papers in laboratory phonology 8: Varieties of phonological competence* (pp. 367–392). Berlin: Mouton de Gruyter.

Selkirk, E. 1996. The prosodic structure of function words. In K. Demuth & J. L. Morgan (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 187–213). Mahwah, NJ: Lawrence Erlbaum.

Shattuck-Hufnagel, S., & Turk, A. E. 1996. A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, *25*(2), 193–247. DOI: http://dx.doi.org/10.1007/BF01708572

Shockey, L. 1987. Rate and reduction: Some preliminary evidence. In R. Channon & L. Shockey (Eds.), *In honor of Ilse Lehiste* (pp. 217–225). Dordrecht: Foris. DOI: http://dx.doi.org/10.1515/9783110886078.217

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. 1994. Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America*, *96*(3), 1314–1324. DOI: http://dx.doi.org/10.1121/1.411453

Stevens, K. N., & Klatt, D. H. 1974. Role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America*, *55*(3), 653–659. DOI: http://dx.doi.org/10.1121/1.1914578

Stuart-Smith, J., Sonderegger, M., Rathcke, T., & Macdonald, R. 2015. The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology*, *6*(3–4), 505–549. DOI: http://dx.doi.org/10.1515/lp-2015-0015

Summerfield, Q. 1975. How a full account of segmental perception depends on prosody and vice versa. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 51–68). Berlin: Springer-Verlag. DOI: http://dx.doi.org/10.1007/978-3-642-81000-8_4

Summerfield, Q. 1981. Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5), 1074–1095. DOI: http://dx.doi.org/10.1037/0096-1523.7.5.1074

Syrdal, A. K., & Gopal, H. S. 1979. A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, *79*(4), 1086–1100. DOI: http://dx.doi.org/10.1121/1.393381

Toscano, J. C. 2011. *Perceiving speech in context: Compensation for contextual variability during acoustic cue encoding and categorization*. Doctoral dissertation. Available from ProQuest Dissertations & Theses database (UMI No.3473251).

Toscano, J. C., & McMurray, B. 2012. Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, and Psychophysics, 74*(6), 1284–1301. DOI: http://dx.doi.org/10.3758/s13414-012-0306-z

Turk, A. E., Nakai, S., & Sugahara, M. 2006. Acoustic segment durations in prosodic research: A practical guide. In S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, … J. Schlieβer (Eds.), *Methods in empirical prosody research* (pp. 1–27). Berlin: Walter de Gruyter. DOI: http://dx.doi.org/10.1515/9783110914641.1

VanDam, M. 2007. *Plasticity of phonological categories*. Doctoral dissertation. Available from ProQuest Dissertations & Theses database (UMI No.3277973).

Volaitis, L. E., & Miller, J. L. 1992. Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America*, *92*(2), 723–735. DOI: http://dx.doi.org/10.1121/1.403997

Wedel, A. B. 2006. Exemplar models, evolution and language change. *Linguistic Review*, *23*(3), 247–274. DOI: http://dx.doi.org/10.1515/TLR.2006.010

Wells, J. C. 1996. *Accents of English 2: The British Isles*. Cambridge, England: Cambridge University Press.

Wright, R. 2004. A review of perceptual cues and cue robustness. In B. Hayes, R. M. Kirchner, & D. Steriade (Eds.), *Phonetically based phonology* (pp. 34–57). Cambridge, England: Cambridge University Press. DOI: http://dx.doi.org/10.1017/CBO9780511486401.002

Xu, Y. 2010. In defense of lab speech. *Journal of Phonetics*, *38*(3), 329–336. DOI: http://dx.doi.org/10.1016/j.wocn.2010.04.003

Yao, Y. 2009. Understanding VOT variation in spontaneous speech. *Proceedings of the 18th International Congress of Linguists*.

Yuan, J., Liberman, M., & Cieri, C. 2006. Towards an integrated understanding of speaking rate in conversation. *Proceedings of 9th International Conference on Spoken Language Processing*, 541–544.

How to cite this article: Nakai, S and Scobbie, J M 2016 The VOT Category Boundary in Word-Initial Stops: Counter-Evidence Against Rate Normalization in English Spontaneous Speech. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7(1): 13, pp. 1–31, DOI: http://dx.doi.org/10.5334/labphon.49

Published: 07 October 2016

]u[*Laboratory Phonology: Journal of the Association for Laboratory Phonology* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS