Individual talker and token covariation in production of multiple cues to stop voicing.

Meghan Clayards

Department of Linguistics, McGill University, Montreal, QC

School of Communication Sciences and Disorders, McGill University, Montreal, QC

2001 McGill College, 8th Floor, Montreal, QC, Canada

H3A 1G1

meghan.clayards@mcgill.ca

Tel: + 514 398 4235

Running Head: Variability in Multiple Cues

Abstract

Background/Aims: Previous research found that individual talkers have consistent differences in production of segments impacting perception of their speech by others. Speakers also produce multiple acoustic-phonetic cues to phonological contrasts. Less is known about how multiple cues co-vary within a phonetic category and across talkers. We examined differences in individual talkers across cues and whether token-by-token variability is a result of intrinsic factors or speaking style by examining within-category correlations. Methods: We examined correlations for three cues (VOT, talker-relative onset f0 and talker-relative following vowel duration) to word-initial labial stop voicing in English. Results: VOT for /b/ and /p/ productions and onset f0 for /b/ productions varied significantly by talker. Token-by-token within-category variation was largely limited to speaking-rate effects. VOT and f0 were negatively correlated within category for /b/ productions after controlling for speaking rate and talker mean f0, but in the opposite direction expected for an intrinsic effect. Within-category talker means were correlated across VOT and vowel duration for /p/ productions. Some talkers produced more prototypical values than others, indicating systematic talker differences. Conclusion: Relationships between cues are mediated more by categories and talkers than by intrinsic physiological relationships. Talker differences reflect systematic speaking style differences.

Keywords

Multiple cues; VOT; stop consonant voicing; production; individual differences

1.0 Introduction

1.1 Multiple cues

Every phonological contrast has a multitude of phonetic correlates, many of which have been demonstrated to affect perception (Raphael, 2005). Throughout this paper we will refer to these phonetic correlates as 'cues' when they have been demonstrated to have an effect on perception. As an example, Lisker (1986) pointed out that there are many potential cues to word-medial voicing in English (e.g. "rapid"-"rabid") such as duration of the preceding vowel, duration of the closure, voice onset time (VOT) presence of vocal fold vibration during closure, burst amplitude, fundamental frequency going into and out of the closure (offset and onset f0) and first formant frequency (F1) going into and out of the closure. Such multiple correlates to phonological contrasts or categories are in fact ubiquitous and perhaps unsurprising given the complex acoustic consequences of a dynamically changing and complex sound generation system such as the vocal tract. A classic literature in speech perception using synthetic speech stimuli has also demonstrated that many acoustic variables contribute to our percepts of stop place of articulation (e.g. Bailey & Summerfield, 1980), stop voicing (e.g. Port & Dalby, 1982) fricative place of articulation (e.g. Mann & Repp, 1980) and many other contrasts. These studies found that the perceptual effect of changes in the value of one cue (e.g. lengthening closure duration in a word-medial stop) can be offset by changes in the value of another cue (e.g. lengthening the preceding vowel), a phenomena termed trading relations (see Repp, 1982 for a review and Toscano & McMurray, 2010 for further discussion). Thus, many acoustic dimensions "make up" phonological categories. While the literature on perceptual trading relations has explored how multiple cues work together in perception, we know much less about how acoustic cues pattern together in the speech signal itself especially within phonological categories and

across talkers. This paper explores one fundamental aspect: to what extent are cues in the speech signal independent of each other and to what extent are they intrinsically linked at the level of the category and at the level of the talker.

We know that cues co-vary across categories. For example, in English, voiceless stops before stressed vowels have long VOTs and relatively high f0 in the onset of the following vowel while voiced stops have short VOTs and relatively low f0 in the following vowel (House & Fairbanks, 1953; Lehiste & Peterson, 1961; Lisker & Abrahmson, 1964; Umeda, 1981; Ohde, 1984). That is, VOT and f0 are positively correlated across voicing categories. However it is a separate question whether within categories (e.g. within voiceless stops) longer VOT is associated with higher f0. In other words, are they still correlated conditional on the category? Conversely, an analog of perceptual trading relations could exist in production. In other words, when a speaker produces an ambiguous value of one cue (e.g. a VOT intermediate between what is expected for voiced and voiceless stops) they could compensate by producing a more extreme value of another cue (e.g. a very high f0). The first goal of this paper is to examine these possible relationships between cues on a token-by-token basis within-category, discussed further in Section 1.2 below. Either of the patterns described above would have implications for our understanding of speech production as well as speech perception. Another possibility is that there are relationships between cues not on a token-by-token basis, but on a talker-by-talker basis as discussed in Section 1.3 below. The second goal of this paper is to examine these individual talker differences.

1.2 Token-by-token variation of multiple cues within categories

There are good reasons to think that some cues could co-vary on a token-by-token basis – that is when we examine individual productions of a single category we might find correlations between cues. If, for example, two cues share a single articulatory or aerodynamic source, then one would expect that they would be inextricably linked in production – intrinsically linked according to the terminology of Wang and Fillmore (1961). For example, if lowering the velum has the effect of lowering the amplitude of the signal as well as increasing the bandwidth of the formants, than one might expect the degree of amplitude reduction and the amount of increase in bandwidth to be correlated. Similarly it has been argued that lowering the larynx to facilitate voicing by increasing supra-glottal volume also decreases the stiffness of vocal folds, thus lowering onset f0 (Hombert, Ohala & Ewan, 1979; Hoole, et al. 2004; Hoole & Honda, 2011). In this situation we would expect these cues to exhibit a correlation on a token-by-token basis within category as well as between categories, and we would expect these correlations within and between categories to be in the same direction as in Panel A of Figure 1.

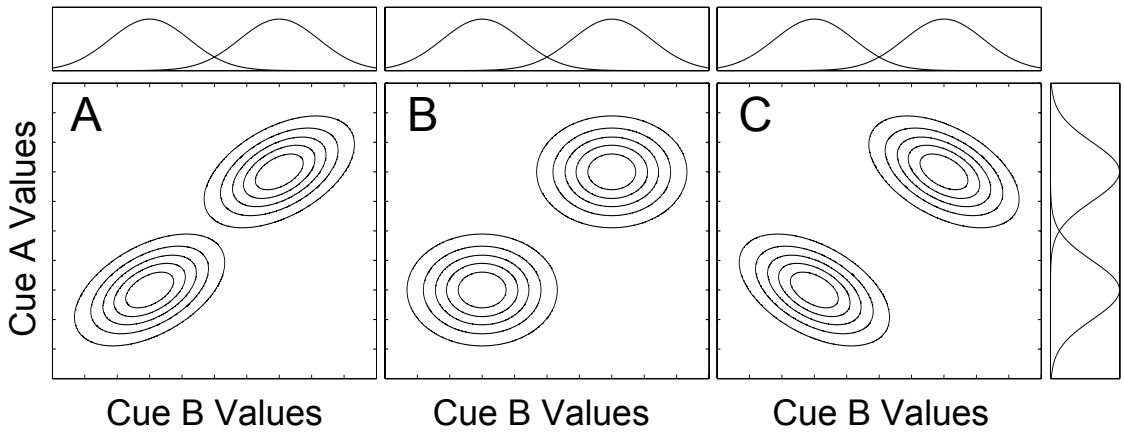

Figure 1: Hypothetical distributions of two cues to two categories showing three possible relationships between the cues [A] token-by-token within-category correlations with a possible intrinsic source, [B] uncorrelated within-category, [C] within-category correlations with "cue trading". Ellipses represent equal probability lines of multivariate distributions. Distributions for each individual dimension are represented as continuous histograms above and to the right of the three panels.

On the other hand, not all of the multiple cues associated with a particular contrast need be produced by a single articulatory movement. For the production of intervocalic Dutch stops, Slis and Cohen (1969) argue on the basis of articulatory data that the movements and musculature responsible for the timing of the oral closure (supra-glottal gestures) – which govern such cues as the duration of the closure interval and the duration of the preceding vowel – are independent from the movements and musculature responsible for the timing of voicing (glottal gestures) – and thus onset f0, amount of voicing during closure and time to voice onset (But see Ohde, 1984 for counter arguments). Thus some of the cues involved in voicing (at least in Dutch) may be independent of one another on a token-by-token basis, while others may not be. It has also been noted that the category correlation between voicing and onset f0 holds across languages despite very different implementations of the voicing contrast (e.g. Dimitrieva, Llanos, Schultz & Francis, 2015).

Based on cross-linguistic patterns and articulatory evidence, Kingston and Diehl have argued that many cues are not automatic consequences of primary articulatory gestures but are under the control of the speaker, although not necessarily conscious control (Kingston & Diehl, 1994). The listener may then learn to associate these cues and thus

produce them together (Nearey, 1997; Holt, Lotto, & Kluender, 2001). Or, alternatively, speakers may produce these cues because they cohere or integrate for auditory reasons as argued by Kingston and Diehl and colleagues. In their view, these cues also serve to enhance the perception of the contrast by contributing to *intermediate perceptual properties* (IPP) (Kingston & Diehl, 1994; Kingston, Diehl, Kirk, & Castleman, 2008). For example, in voicing contrasts, they argue that lowered f0 and voicing during closure both contribute to an IPP of low frequency energy continuity (Kingston, et al., 2008). Thus the lower f0 at the stop closure enhances the percept of voicing beyond what the vocal fold vibration contributed. The possibility that speakers of a language (unconsciously) choose to co-vary cues in order to enhance perception (i.e. low f0 for voiced stops and higher for voiceless) would imply a relationship between them that is not necessarily intrinsic to the articulatory system. That is, we expect the between category correlations that we know exist between cues, but there is no reason to expect within-category correlations as illustrated panel A of Figure 1. Instead we might expect a pattern like Panel B. In this pattern the cues are correlated across categories but independent within the categories (conditionally independent).

Alternatively, if cues are produced together for perceptual reasons (either due to IPP processes or simply to provide redundancy in the signal), speakers might use one cue to compensate for ambiguities created by another cue. For example, if a speaker intended to produce a voiceless stop but produced a VOT that is shorter than is prototypical for such a stop, they might also increase f0 in order to make sure that they conveyed the intended message to the listener. This strategy would essentially take advantage of trading relations in perception where a less ambiguous cue in one dimension can compensate for a more ambiguous cue in another dimension. Note that such a

production strategy does not depend on the perceptual interactions described by Kingston and Diehl and colleagues discussed above. Even if cues do not perceptually cohere, and even if cues are not correlated in any way within categories in production, multiple cues will produce trading relations in perception, as long as listeners attend to the cues to make judgements about categories (see Clayards, 2008; Toscano & McMurray, 2010). However, if speakers adopted a compensatory strategy in production, one would expect that ambiguous values of one cue would tend to co-occur with more extreme values of another creating within-category correlations as illustrated in panel C of Figure 1 in which the within-category correlations are in the opposite direction of the between-category correlation. It should also be noted that even if speakers are not able to modulate cues on a token-by-token basis in production, they might do so on an individual basis. In other words, a talker who tends to use VOT more distinctively may use f0 less or vice versa. This would also tend to create the pattern in panel C of Figure 1, however it would be true of talker means, not of token-by-token variability. Patterns due to talker means are discussed further in Section 1.3.

In summary, there are reasons to expect any of the three within-category patterns. Cues could be linked intrinsically in the articulatory system creating within-category patterns in the same direction as the between-category patterns; cues could be independent within categories despite being correlated between categories; cues could be produced in a compensatory way on a token-by-token basis creating within-category correlations in the opposite direction from the between-category correlation. Observing any of these patterns would tell us more about the mechanisms that determine the internal structure of categories. However another important source of variability in speech is not just

categories but also talkers and the relationship between cues may be closely linked to the individuals who produced them.

1.3 Individual differences in production

There is a growing literature on individual differences in production and perception of speech. Here we focus on production differences and their consequences for speech perception by other listeners. Research has shown that individual talkers are distinct in their production of phonetic cues. Newman, Clouse and Burnham (2001) showed that individuals can have very different productions of sibilants, and that in some cases the distribution of spectral centre of gravity for /s/ of one talker overlaps entirely with /ʃ/ for another talker. They also found that talkers differed in how much the two categories overlapped, and that the degree of overlap had consequences for how well listeners were able to recognize the sounds. VOTs in English are known to be systematically different between talkers, including amount of pre-voicing and amount of voicing lag (e.g. Allen, Miller & DeSteno, 2003; Scobbie, 2008) even when controlling for speaking rate, while Theodore, Miller and DeSteno (2009) found that speaking rate effects on VOT were also talker-specific. McMurray and Jongman (2011) investigated talker effects in fricative production and found significant effects of talker on almost every cue investigated. Individual differences have also been demonstrated for articulatory data (e.g. Johnson, Ladefoged & Lindau, 1993) which can result in different acoustics (Noiray, Iskarous & Whalen, 2014). Thus variability in production is well documented and has consequences for listeners.

While individual talker differences have been shown to impact on how easily some talkers are understood (Newman et al., 2001; Ferguson, 2004; Hazan & Markham,

2004), there is also an advantage for listening to talkers that are familiar (Nygaard & Pisoni, 1998). Some of these advantages are tied to language-specific knowledge and so likely reflect knowledge of how variability is structured in the language (Goggin et al, 1991). This suggests that one way that listeners deal with individual talker variability is by adapting to the patterns of talkers they are exposed to. There have to date been many demonstrations that listeners do in fact adapt to the properties of the speech that they are exposed to (e.g. Bertelson, Vroomen, & de Gelder, 2003; Norris, McQueen and Cutler, 2003) and this may be tied to individual talkers (Eisner & McQueen, 2005). Theodore and colleagues (Theodore & Miller, 2010; Theodore, Myers & Lomibao, 2015) showed that listeners are sensitive to talker differences in VOT and can generalize expectations about talker VOT to novel tokens. Because speech is highly multidimensional – even for a single contrast, and there are many phonological contrasts – tracking and adapting to each of these dimensions for each talker seems like a daunting task for any listener. One possibility is that these dimensions don't vary randomly between talkers but instead bear a systematic relationship. For example talkers with relatively long VOTs in voiceless stops could have relatively high f0 in those stops as well (analogous to the pattern in panel A of Figure 1, but for talker means instead of tokens). Conversely, talkers who signal voicing relatively more consistently with VOT may use f0 relatively less consistently (analogous to the pattern in panel C of Figure 1, but for talker means instead of tokens). This would greatly simplify the task for the listener as it would require them to identify some aspect of the variation in order to predict many other aspects. Thus the second goal of this paper is to examine how cues vary across talkers within a single phonetic category and across phonetic categories within a voicing contrast.

1.4 Present study

The present study examined three acoustic cues to voicing in voiced and voiceless word-initial mono-syllabic labial stops (e.g. "beach" and "peach") – VOT, f0 at vowel onset and length of the following vowel. We included following vowel duration as a cue to voicing although it has been argued to be cue to speaking rate that should be interpreted relative to VOT (e.g. Kohler, 1979; Boucher, 2002) as it seems to play an independent role in perception of voicing (McMurray, Clayards, Tanenhaus & Aslin, 2008; Toscano & McMurray, 2012). Vowel duration is also used as a measure of speaking rate as a control variable in evaluating the other two cues as it is known to effect VOT especially in /p/ productions (Miller, Green & Reeves, 1986).

Very few studies have reported on the correlations of cues within categories. One well-known example is locus equations, which demonstrate very high correlations for F2 at vowel onset and F2 at vowel midpoint within a single stop category but across vowels (Sussman, Fruchter, Hilbert & Sirosh, 1998). However, it is not clear whether this constitutes a particular kind of contextual variation (different from the within-category variation of interest here) or a single integrated cue for the listener. Most relevant to the current paper, a recent study by Dmitrieva, Llanos, Shultz and Francis (2015) examined correlations between VOT and onset f0 for Spanish and English. They found a weak negative trend between VOT and onset f0 for voiceless stops in English (longer VOTs with lower f0) and a weak positive correlation for voiced stops in English (shorter VOTs with lower f0). A second recent study examined correlations between VOT and onset f0 in French and Italian stops. Kirby & Ladd (2015) found the same pattern in voiceless Italian stops as Dmitrieva et al. found for English, i.e. longer VOTs with lower f0, but no correlations for French. In both Italian and French, negative VOTs

occur with voiced stops and the researchers found that longer voicing lead was correlated with lower f0. Together these two studies have generally found a negative relationship between f0 and VOT for voiceless stops and a positive relationship for voiced stops, though for some languages there was no relationship. However, there is an important limitation to these results in that differences between individuals were only partially considered – f0 measurements were normalized to talkers' overall means, however differences in individual talkers' VOTs were not considered and are crucial to understanding the patterns.

If the two cues are intrinsically linked by articulatory mechanisms then we would expect the pattern in Panel A of Figure 1 for each individual talker. However differences in mean values of the cue across talkers could obscure the relationship when pooling across talkers. For example, for any given talker, a relationship like panel A of Figure 1 could hold, but if talkers have different average VOTs (as has been shown, see above), the group data could look like panel B. Similarly, an apparent relationship between the cues across talkers could be due to differences in mean values between talkers. For example, some talkers could produce large differences in cue values for the two categories for both cues while others produce smaller differences for both cues. If we then pooled across talkers, a pattern like the one in Panel A of Figure 1 would emerge with some talkers' productions falling towards the middle and others towards the extremes. On the other hand, talkers could instead differ in the extent to which they rely on a particular cue versus another cue to signal a contrast. This could be due to physiological differences between talkers or simply due to individual preferences. Shultz, Francis and Llanos (2012) used the same data as in Dmitrieva et al. (2015) to examine the relationship between production cue weights for f0 and VOT for individual

talkers. They used linear discriminant analysis (LDA) to compute coefficients for VOT and f0 for each talker. They found a significant negative correlation between the two such that the more that a given talker signalled the contrast with VOT, the less they signalled it with f0. This trade-off relationship predicts a pattern similar to panel C of Figure 1 when the results are pooled across talkers, and this is what was reported for the voiceless category of Dmitrieva et al. (2015). Kirby and Ladd (2015) also pooled data across talkers. It is therefore possible that the patterns reported in both papers are due to differences in individual talkers' average VOT and average f0 rather than an intrinsic relationship between the cues. It is therefore critical to examine the behaviour of individual talkers as well as to separate talker-by-talker variation from token-by-token variation in multiple cues. Examining how individual talkers use cues to signal the voicing contrast (i.e. across categories) will also allow us to better understand differences between talkers.

2.0 Methods

2.1 Participants

Nine talkers (7 female) at the University of Rochester, in Rochester, NY initially produced a set of recordings. All were native monolingual English speakers from a variety of North American dialect regions (age 23-30). An additional 11 talkers (3 female) were recruited from McGill University, Montreal Canada with the same characteristics (all were native monolingual speakers of either American or Canadian English from a variety of dialect regions in the same age range) who completed the same task. The total number of talkers was 20 (10 female).

2.2 Procedure

Talkers read the words from index cards, repeating each word three times in succession. The order of cards was randomly shuffled for each talker. The cards were then re-shuffled and each word was read 3 times again in a different random order for a total of 6 recordings of each word (2 talkers read the list a third time and some words had to be discarded due to speech errors or background noise resulting in slightly different numbers of tokens per talker). No carrier phrase was used. The first set of talkers (Rochester) were recorded in a sound attenuated booth using a Marantz portable digital recorder (PMD 670) and a Technica lapel microphone worn approximately six inches from the mouth. Sound files were digitized at a sampling rate of 32,000 Hz. The second set (Montreal) were recorded in a sound attenuated booth directly onto a computer at a sampling rate of 44,100 Hz using Praat (Boersma and Weenink, 2011) and a Logitech H390 USB headset microphone. Four of these talkers (2 female) read from a longer set of words (which included 10 disyllabic words with word-medial stops not analyzed here). These talkers produced all of the words in a randomized order, repeating each word three times when they read it as before. They repeated this procedure 10 times with a different randomized order each time. This created a total of 30 recordings for each word (some productions were again discarded due to errors, noise or recording problems). In the analyses reported here, all the data is grouped together and mixed effects regression is used which accounts for differences in amount of data between talkers.

2.3 Word List

Three minimal pairs were chosen: beach, peach, bees, peas, beak and peak. Words were originally chosen to be part of a perceptual experiment and were chosen to be minimal pairs with the same vowel quality and to be picturable.

2.4 Acoustic Measurements

All acoustic analyses were performed using the Praat software package (Boersma & Weenink, 2011). Temporal intervals (e.g. length of the vowel) were marked by hand according to characteristic markers in the waveforms.

Both positive and negative Voice Onset Time (VOT) was measured. Positive VOT was defined as the time from the beginning of the stop burst to the onset of voicing of the vowel. The onset of voicing was defined as the point where the (quasi) periodic portion of the waveform first crossed zero in the positive direction. Some speakers produced a period of voicing before the release burst. In this case, negative VOT was measured from the onset of periodic voicing until the stop burst and the period between the burst and the onset of voicing in the subsequent vowel was not analyzed.

The onset of the vowel was always the onset of voicing. The offset of the vowel was defined differently depending on the following consonant. Before stops and affricates the end of the vowel was the last pitch cycle before a significant drop in amplitude. Before fricatives it was the last pitch cycle before significant frication noise. Words ending in voiced consonants ("bees" and "peas") had longer vowels than those ending in voiceless consonants ("beach", "beak", "peach" and "peak", see Table 1) as expected (House & Fairbanks, 1953, Mack, 1982). We therefore subtracted the mean vowel duration (averaged across talkers) for words with voiced codas from each of the words with a voiced coda and the mean vowel duration (averaged across talkers) for voiceless codas for each of the words with a voiceless coda (hereafter final-voicing-relative vowel duration). This measure was used as a measure of speaking rate in the individual

difference analyses below. A second measure took account of individual differences in speaking rate. Each talker's mean vowel duration was subtracted from the final-voicing-relative vowel duration (hereafter relative vowel duration for simplicity).

Onset f0 measurements were obtained by hand-measuring the duration of the first three pitch cycles of the following vowel as in Cole, Kim, Choi, and Hasegawa-Johnson (2007). Very low onset f0 values due to creak were removed (below 70 Hz). Because men and women have different average f0 (see Table 1), onset f0 was transformed to semitones relative to each talker's mean onset f0 using the formula 12 ln(x/individual mean onset f0)/ln 2 (as in Schultz et al. 2012 and Dmitrieva et al. 2015). Resulting values are on a logarithmic scale accounting for higher dispersion for higher f0 values. Positive values are above the talker's mean onset f0 and negative values are below the talker's mean onset f0.

3.0 Results

3.1 Distribution of Cues

A total of 1,371 tokens were analysed. Raw means for each category (untransformed) are summarized in Table 1. Figure 2 shows the smoothed density plots of each cue for each category. Vowel duration and onset f0 are normalized as described above (talker and voicing relative vowel duration, talker-relative semitones for onset f0). There is little overlap between the categories for VOT and considerable overlap for all other cues consistent with previous observations that VOT is the dominant cue to this contrast in word-initial position (Lisker & Abrahmson, 1964). Onset f0 is higher and following vowel shorter for /p/ productions as expected given the previous literature (House & Fairbanks, 1953; Lehiste & Peterson, 1961; Umeda, 1981; Ohde, 1984).

Table 1: Means for each of the cues by voicing category and talker gender. Mean for vowel duration also grouped by word-final voicing.

Gender		VOT (ms)	f0 (Hz)	Vowel (ms)	Final Voicing		Vowel (ms)
/b/	f	2.4	225.4	206.1	/b/	Voiced	296.1
	m	7.5	116.2	182.4		Voiceless	144.0
/p/	f	76.1	245.9	172.4	/p/	Voiced	266.9
	m	68.0	125.3	162.5		Voiceless	119.8

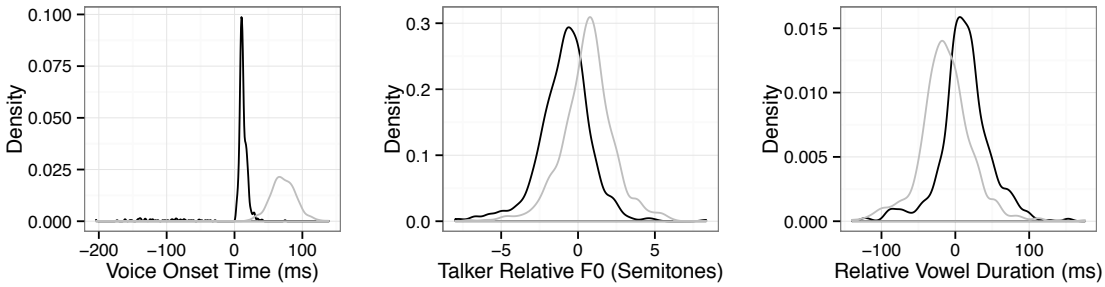

Figure 2: Density plots smoothed with a Gaussian kernel showing the distribution of values for each cue for each category pooled across all speakers. Black lines are /b/ productions, grey lines are /p/ productions. Onset f0 is normalized relative to talker mean onset f0. Vowel duration is normalized for voicing of the following consonant and talker mean duration.

3.2. Correlations between cues

We now turn to the raw correlations between cues. The left panel of Figure 3 shows measurements of VOT versus onset f0 for each token for the voiced and voiceless categories. The middle panel shows VOT versus relative vowel duration measurements. The right panel shows relative vowel duration measurements versus onset f0. To examine the relationship between the three cues we began with simple correlational analysis for each token within-category as was done in Dmitrieva et al. (2015) and Kirby and Ladd (2015) for VOT and onset f0. Because the pre-voiced VOTs were not contiguous with the short lag VOTs we examined correlations for these separately (see

Dmitireva et al., 2015 and Kirby & Ladd, 2015 for discussion of reasons why pre-voiced and short-lag stops may pattern differently). To see if these correlations were in the same direction as the between-category correlation (as in panel A of Figure 1) or in the opposite direction (panel C of Figure 1) the correlation between categories was also calculated. Note that in the case of the between-category correlations it is the sign of the correlation that is of interest and not the magnitude. The magnitude of the between-category correlation will be proportional to how well the two cues signal the contrast independently of each other. Because voiced stops have on average shorter VOT, lower onset f0 and longer vowel duration, we expect a positive between-category correlation between VOT and onset f0, a negative between-category correlation between VOT and vowel duration, and negative between-category correlation between onset f0 and vowel duration, and this is what we found (Table 2). For the within-category correlations the sign, magnitude and significance are all relevant.

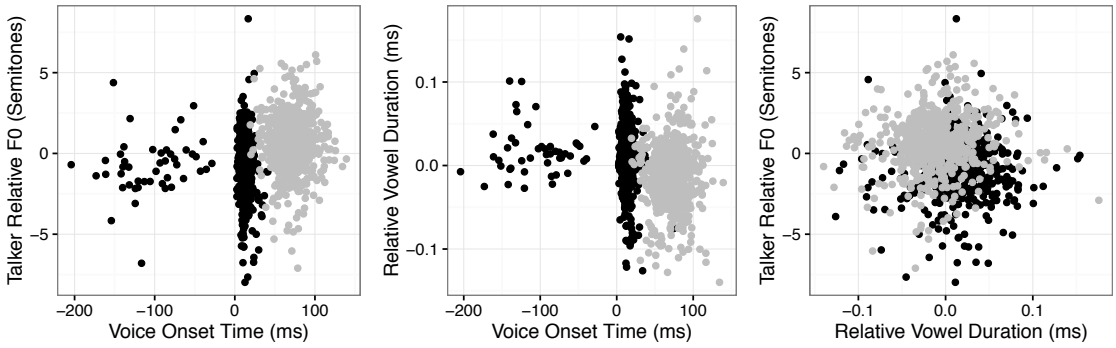

Figure 3. Scatter plots of individual tokens for VOT, talker-relative f0 and final-voicing and talker-relative vowel duration. Black circles are /b/ tokens and grey circles are /p/ tokens.

Table 2: Pearson R correlations for VOT, onset f0 and final voicing and talker-corrected vowel duration for each voicing category as well as both categories together. VOT values for /b/ are split into positive (+ve) and negative (-ve) VOT values.

	Data	Pearson r	*p*	N
f0/VOT	All	0.309	0.000	1271
	/p/	0.000	0.995	652
	/b/+ve	-0.111	0.008	573
	/b/-ve	-0.208	0.160	47
Vowel duration /VOT	All	-0.280	0.000	1366
	/p/	0.054	0.158	686
	/b/+ve	-0.155	0.000	635
	/b/-ve	-0.022	0.886	46
f0/ Vowel duration	All	-0.079	0.005	1269
	/p/	0.116	0.003	652
	/b/	0.019	0.630	617

Within categories, we found no correlations between VOT and f0 for the /p/ productions ($r(650)=0.00$, $p=0.995$) unlike the weak negative correlation found by Dmitrieva et al. (2015) for English voiceless stops. We found a negative correlation in the /b/ productions between positive VOTs and onset f0 ($r(571)=-0.111$, $p=0.008$) and a non-significant negative correlation for the negative VOTs ($r(45)=-0.208$, $p=0.160$) which should be treated with caution given the low N. These are in the opposite direction as the results of Dmitrieva et al. (2015) for English. For vowel duration and VOT there was a non-significant positive correlation within the /p/ productions ($r(684)=0.054$, $p=0.158$). Within the /b/ productions there was a negative correlation for positive VOT values (longer vowels go with shorter VOTs, ($r(632)=-0.155$, $p<0.001$) and a non-significant negative correlation for negative VOT values (longer vowels go with more negative VOTs, ($r(44)=-0.022$, $p=0.886$) but again the results for negative VOTs are based on a very small sample. The positive relationship within the /p/ productions is consistent with a speaking rate effect, where longer vowels and longer VOT are both part of a slower speaking rate (Kessinger & Blumstein, 1997; 1998); however, this

effect was small and non-significant. Within the /b/ productions, the pattern with the negative VOT values could also be explained this way as a slower speaking rate leads to longer pre-voicing (Kessinger & Blumstein, 1997). However, the pattern for positive VOT values within /b/ productions is not consistent with overall slowing as we found longer vowels occurred with shorter VOT. Finally, onset f0 and relative vowel duration were positively correlated within /p/ productions ($r(650)=0.116$, $p=0.003$) and showed a non-significant negative correlation within /b/ productions ($r(615)=0.019$, $p=0.630$).

So far, none of these trends is clearly consistent with any of the panels in Figure 1 (intrinsic, compensatory or simply uncorrelated) because the within-category correlations are not consistent across the two voicing categories. The patterns are also inconsistent with some previous findings. However, these simple correlations don't allow us to examine whether talker differences in mean values for each category observed in the previous section are contributing to or obscuring a relationship as discussed in the introduction. The goal of the next section is to explore these relationships more fully as well as to look at individual differences in use of cues.

3.3 Individual differences in individual cues

Before examining the question of whether cues are correlated across talkers, we first determined if there were significant talker differences between cues for each voicing category using the same modelling strategy as Allen et al. (2003). We built mixed effects regression models for each variable tested (VOT, onset f0 and relative vowel duration) using the lmer package in R (Bates, et al. 2014). The models for VOT and onset f0 also included final-voicing relative vowel duration as a control measure for speaking rate. Because talkers repeated each word three times in a row, they may have adopted a prosodic contour that repeated with each set of three. Inspection of VOT, f0

and vowel duration values according to repetition suggested that VOT increased, f0 decreased and vowel duration increased over the three repetitions. To control for this we added repetition as a continuous (centered) variable to the each of the models. Then, as in Allen et al. (2003), we added random intercepts for talkers to determine if individuals reliably differ from each other in how they produced each cue. Allen et al. (2003) examined VOT for /p/ productions, but here we extend it to /b/ productions and to two other cues. As a stricter test of individual differences in /b/ productions, only productions with positive VOTs were included. Thus any differences found are not due to some individuals using negative VOT and others not. Note that this may underestimate individual differences. A random intercept by word was also included to model potential differences in VOT according to word-level factors like neighbourhood density (Goldinger & Summers, 1989; Fox, Reilly & Blumstein, 2015). We examined the improvement to model fit of including the random talker intercepts using a Likelihood Ratio Test ($\chi 2$) comparing models with just the random word intercept and models including both intercepts (see also Kliegl et al. 2011 for discussion of using random effects to examine individual differences). Finally, to test whether individuals varied in how their productions were affected by speaking rate and repetition we included random slopes by participants for each. Significance of these slopes was tested by likelihood ratio tests by comparing intercept-only models to models containing each of the random slopes in turn. Correlations between random effects were not fit in any of the models. Significance of the fixed effects was determined with a Wald test (fixed effects are reported for models which include all the random intercepts and slopes). Results of the model comparisons are given in Table 3.

Table 3. Results from model comparison tests of talker differences for each cue and category for /p/ productions (Left) and /b/ productions (Right). Models were lmer(X ~ {vowel_length} + rep + (1+{vowel_length} +rep ||talker) + (1|word)). /b/ productions include only positive VOTs (no pre-voicing)

/p/		*χ2 (df)*	*p*	/b/		*χ2 (df)*	*p*
VOT	Intercept	210.1 (1)	0.00	VOT	Intercept	66.6 (1)	0.00
	Vwl len.	1.4 (1)	0.23		Vwl len.	0 (1)	1.00
	Rep.	0 (1)	1.00		Rep	0.2 (1)	0.68
f0	Intercept	0 (1)	1.00	f0	Intercept	4.3 (1)	0.04
	Vwl len.	8.8 (1)	0.00		Vwl len.	4.5 (1)	0.03
	Rep.	121.6 (1)	0.00		Rep	25.8 (1)	0.00
Vwl	Intercept	0 (1)	1.00	Vwl	Intercept	0 (1)	1.00
	Rep.	53.3 (1)	0.00		Rep	57.3 (1)	0.00

Table 3 shows that when controlling for repetition and where possible speaking rate, there were significant talker differences in: VOT for both /p/ productions ($\chi 2(1)=210.1$, $p<0.005$) and /b/ productions ($\chi 2(1)=66.6$, $p<0.005$); onset f0 for /b/ productions ($\chi 2(1)=4.3$, $p=0.04$); but not vowel duration. Note that our measures of onset f0 and vowel duration have already factored out most of the between-talker differences by making the measures relative to individual talkers' means across categories. What is being tested here is whether talkers differ in the degree to which voicing category perturbs their onset f0 or vowel duration. Thus talker differences in mean f0 and speaking rate have been eliminated. These results confirm previous findings that English-speaking adults differ in their use of VOT to signal a voicing contrast and extend it to positive VOTs for /b/ productions as well as to onset f0 for /b/ productions.

The models provided evidence that talkers differed in the effect of repetition on their productions for onset f0 for /p/ productions ($\chi 2(1)=121.6$, $p<0.005$) and /b/ productions

($\chi 2(1)=25.8$, $p<0.005$) and vowel duration for /p/ productions ($\chi 2(1)=53.3$, $p<0.005$) and /b/ productions ($\chi 2(1)=57.3$, $p<0.005$). In the other cases the model estimates indicated that there was essentially no variation by talkers. Talkers also differed in the effect of speaking rate (vowel duration) only on onset f0 for /p/ productions ($\chi 2(1)=8.8$, $p<0.005$) and /b/ productions ($\chi 2(1)=4.5$, $p=0.03$).

Analysis of fixed effects for the models found that at slower speaking rates (longer vowels) VOTs were longer for /p/ productions ($\beta=2.24$, SE=0.85, t=2.64, p=0.01) while repetition reduced f0 for /p/ productions ($\beta=-0.59$, SE=0.20, t=2.87, $p<0.005$) and /b/ productions ($\beta=-0.33$, SE=016, t=2.02, p=0.04).

Given the significant individual differences found in these analyses, we now turn to the question of how these cues pattern together both within categories and across talkers.

3.4 Talker means and token-by-token deviations

To test whether there was a relationship between the cues within talkers we computed the mean VOT, onset f0 and relative vowel duration for each talker and each voicing category. To test whether there were relationships that could not be explained by talker differences we computed the deviances from these means for each production. This allowed us to separate out variation in each cue that was due to differences in voicing and talker from variation that was not. If we observe relationships between cues that is not due to voicing or talker, we could more safely characterize this variation as arising due to intrinsic factors or conversely to token-by-token compensation.

Figure 4 shows the mean onset f0, mean relative vowel duration and mean VOT for each talker by voicing category.

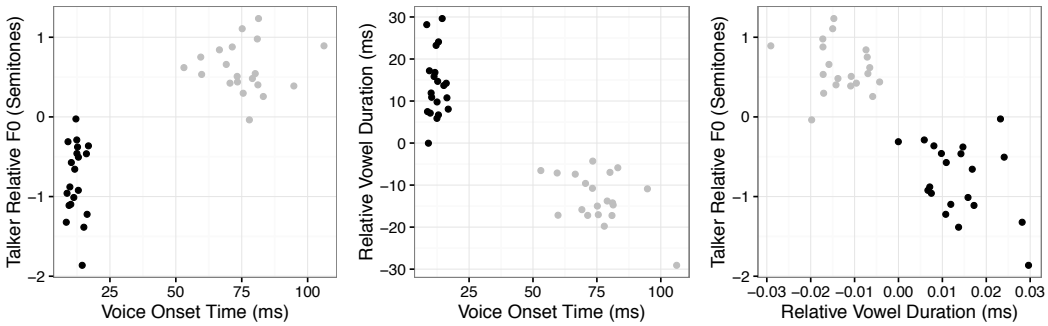

Figure 4: Talker means for each voicing category for VOT, final-voicing-and-talker-relative vowel duration and talker-relative f0. Black circles are /b/ productions and grey circles are /p/ productions.

To test these relationships statistically, we built on the individual differences models constructed in Section 3.3 above. To each model we added fixed effects of the talker means and deviations for the other two cues. For example, the model predicting VOT for /p/ productions had a fixed effect of mean onset f0 for each talker's /p/ productions as well as a fixed effect for the deviations from those means and a fixed effect of mean relative vowel duration for each talker's /p/ productions as well as a fixed effect for the deviations from those means. As before, because the negative VOT productions could be different from positive VOT productions and because there were relatively few negative VOTs, we included only productions with positive VOTs in the analysis. This means that for talkers with some negative VOT values, the mean VOT of /b/ productions does not reflect those negative VOTs. Therefore this is again a conservative estimate of the relationship between the cues. The random effects structure was

expanded to include a random slope by participant for each of the deviation measures. The random slope for repetition by participant was removed from the models for VOT because the previous models of individual difference found the variance by talkers was near zero and thus increased model complexity without improving fit (see Barr et al. 2013 and Matuschek et al. 2016 for discussion of when to include random effects). Significance of fixed effects was tested using a Wald test on the *t* statistic. Collinearity between all fixed effects was tested using Kappa (the condition number) using the collin.fnc function in the languageR package (Baayen, 2009) and found to be very low in both the /b/ productions ($k = 1.79$) and the /p/ productions ($k = 1.97$).

Table 4: Model output from lmer(VOT ~ vowel duration mean + vowel duration deviation + onset f0 mean + onset f0 deviation + repetition + (1 + vowel duration deviation + onset f0 deviation||talker) + (1|word)) for data from each voicing category.

		Estimate	Std.Error	t.value	p
/p/	Intercept	74.94	3.08	24.36	0.00
	Vowel duration mean	-5.15	2.17	-2.38	0.02
	Vowel duration deviation	1.51	0.58	2.60	0.01
	Onset f0 mean	-1.20	2.50	-0.48	0.63
	Onset f0 deviation	-1.05	1.11	-0.94	0.35
	Repetition	-0.48	0.73	-0.65	0.51
/b/	Intercept	12.23	0.66	18.40	0.00
	Vowel duration mean	-0.18	0.51	-0.35	0.72
	Vowel duration deviation	-0.57	0.22	-2.61	0.01
	Onset f0 mean	0.08	0.56	0.15	0.88
	Onset f0 deviation	-0.53	0.32	-1.66	0.10
	Repetition	0.05	0.27	0.19	0.85

Results of the models for VOT are found in Table 4. The model for /p/ productions found a negative relationship between VOT and vowel duration mean (β=,-5.15 *SE*=2.17, t=2.38, p=0.02). Since vowels following voiceless stops are on average

shorter than those following voiced stops, this means that talkers with on average more prototypical vowel duration (after taking into account the talker's mean vowel duration) for /p/ also produce more prototypical VOTs. Additionally, vowel duration deviation and VOT had a positive relationship (β=1.51, *SE*=0.58, *t*=2.6, *p*=0.01). In other words, productions that were longer than the talkers' category means also had longer VOTs. The model for /b/ productions found no relationship between talker mean vowel duration and VOT but a negative relationship between vowel duration deviation and VOT (β=-0.57, *SE*=0.22, *t*=2.61, *p*=0.01) indicating that productions with vowels longer than the talker's mean had shorter VOTs. Together these results indicate that talkers with on average more prototypical vowel duration for /p/ productions produce more prototypical VOTs and that decreased speaking rate results in more prototypical productions of VOT for both /p/ and /b/ productions.

The models for both /p/ and /b/ productions found no relationships between VOT and talker's mean onset f0 or the deviations from those means. Repetition did not influence VOT.

Two parallel models were constructed to examine onset f0. Results are in Table 5. The only reliable effect on onset f0 was VOT deviation for /b/ productions wherein longer VOTs occur with lower onset f0 (β=-0.02, *SE*=0.09, *t*=2.20, *p*=0.03). This is in line with the raw correlations observed in Section 3.2 and is confirmed here after controlling for talker means and repetition. It is in the opposite direction from the effect reported in Dmitrieva et al. (2015) for English and in the opposite direction from an intrinsic effect. It is in the same direction as Kirby and Ladd (2015) report for Italian short lag stops.

Repetition decreased onset f0 significantly in /p/ productions (β=-0.59, *SE*=0.21, *t*=-2.88, $p<0.005$) and marginally in /b/ productions (β=-0.33, *SE*=0.18, *t*=1.86, *p*=0.06).

Table 5: Model output from lmer(onset f0 ~ vowel duration mean + vowel duration deviation + VOT mean + VOT deviation + rep + (1 + vowel duration deviation+VOT deviation+rep||talker) + (1|word)) for data from each voicing category (positive VOTs only).

		Estimate	Std.Error	t.value	p
/p/	Intercept	0.52	0.09	5.56	0.00
	VOT mean	-0.08	0.06	-1.31	0.19
	VOT deviation	-0.07	0.10	-0.66	0.51
	Vowel duration mean	-0.04	0.06	-0.69	0.49
	Vowel duration deviation	0.05	0.06	0.86	0.39
	Repetition	-0.59	0.21	-2.88	0.00
/b/	Intercept	-0.86	0.12	-7.26	0.00
	VOT mean	-0.06	0.09	-0.59	0.56
	VOT deviation	-0.20	0.09	-2.20	0.03
	Vowel duration mean	-0.12	0.09	-1.26	0.21
	Vowel duration deviation	-0.05	0.09	-0.55	0.58
	Repetition	-0.33	0.18	-1.86	0.06

The third set of models examined vowel duration. As before, this measure was final voicing and talker mean corrected. Results are summarized in Table 6. For /p/ productions vowel duration was significantly affected by VOT deviation in the same way that we saw in the VOT model. In other words, longer than average VOTs were associated with longer than average vowel durations (β=5.10, *SE*=2.49, *t*=2.05, *p*=0.04), likely due to speaking rate. Repetition did not affect vowel duration.

Table 6: Model output from lmer(relative vowel duration ~ f0 mean+ f0 deviation + VOT mean + VOT deviation + repetition + (1 + f0 deviation + VOT deviation ||talker) + (1|word)) for data from each voicing category (positive VOTs only).

		Estimate	Std.Error	t.value	p
/p/	Intercept	-12.10	1.42	-8.55	0.00
	VOT mean	-1.82	1.41	-1.29	0.20
	VOT deviation	5.10	2.49	2.05	0.04
	Onset f0 mean	0.50	1.47	0.34	0.73
	Onset f0 deviation	1.62	1.55	1.05	0.30
	Repetition	3.50	3.08	1.14	0.26
/b/	Intercept	13.30	2.37	5.61	0.00
	VOT mean	-0.12	1.43	-0.08	0.93
	VOT deviation	-3.86	3.47	-1.11	0.27
	Onset f0 mean	-2.35	1.53	-1.53	0.13
	Onset f0 deviation	-0.01	1.79	0.00	1.00
	Repetition	2.47	2.20	1.12	0.26

In summary, we found that talker means revealed that some talkers spoke with more prototypical values: talkers with shorter average vowels for /p/ productions had longer VOTs. For /b/ productions, there were no relationships between talker means for onset f0 or vowel duration and VOT.

Analysis of deviations from talker means found that on a token-by-token basis, a slower speaking rate (longer vowels) led to longer /p/ VOTs and shorter /b/ VOTs. We also found that for /b/ productions, lower f0 was associated with longer VOTs, which is inconsistent with an intrinsic effect. Thus we found evidence that, other than speaking rate effects, cues seem to be correlated more at the level of talkers than within categories on a token-by-token basis.

3.4 Individual differences in contrast signalling

In the previous sections we found significant differences in how talkers produced each category and these differences were consistent with some talkers speaking more carefully or hyperarticulating. However, we only examined behaviour of talkers within a single category and not how they signalled the contrast between the two categories. In this section we test whether there is a relationship between the degree to which individual talkers rely on one cue versus another cue to signal the contrast. Talkers who produce one cue more prototypically may produce another cue more ambiguously (as in Shultz et al. 2012). Conversely if some talkers are speaking more carefully, they may produce all cues more prototypically than other speakers. To determine how consistently each cue was produced by each speaker, we quantified the amount of overlap between the categories. The amount of overlap between categories can be used as a metric of how precisely that cue conveys the contrast (Nearey & Hogan, 1986; Clayards, 2008; Noiray et al.,, 2014) and has been shown to influence perception (Newman et al., 2001; Clayards, Tanenhaus, Aslin & Jacobs, 2008). In the extreme, a cue for which the distributions of two categories are entirely overlapping does not convey any information about the contrast. On the other hand, a cue for which there is no overlap between the categories distinguishes the contrast perfectly. The amount of overlap was quantified using Cohen's D (Green & Swets, 1966) which was calculated by the equation in 1.

(1) $$D' = \frac{\mu_{"b"} - \mu_{"p"}}{\left(\sigma_{"b"} + \sigma_{"p"}\right)/2}$$

Where $\mu_{\text{“b”}}$ and $\mu_{\text{“p”}}$ refer to the means of the /b/ and /p/ categories respectively and $\sigma_{\text{“b”}}$ and $\sigma_{\text{“p”}}$ refer to the standard deviations of the /b/ and /p/ categories respectively.

Cohen's D values were calculated for each individual and each cue separately and are shown in Figure 5. Values for VOT were calculated on positive VOTs only to be more comparable to Shultz et al. (2012).

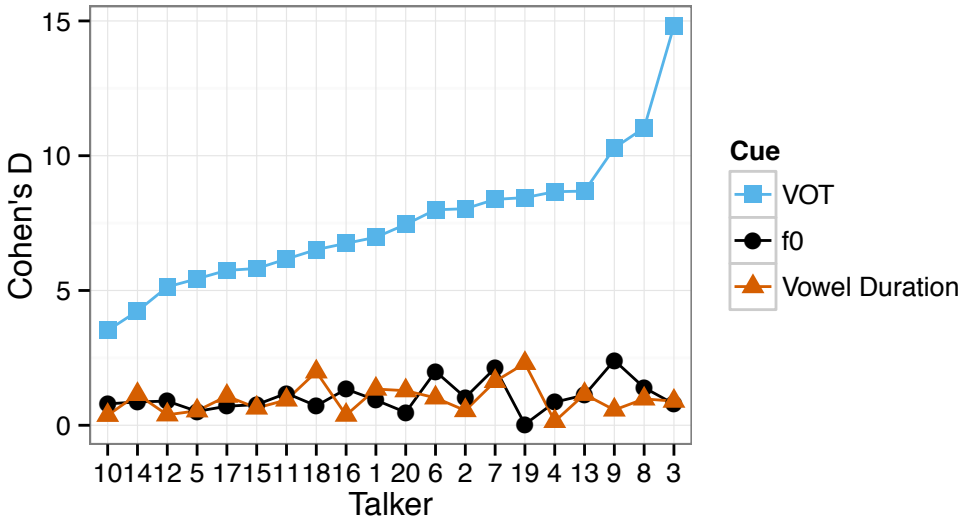

Figure 5: Cohen's D values of each cue for each talker.

For each talker, the Cohen's D value for VOT is much greater than any of the other cues. There is also considerable variability among speakers in the values for VOT but much less in the values of other cues. Correlations between Cohen's D values of each cue found no significant relationships (Table 7). To more directly compare our results to Shultz et al. (2012) we also performed a linear discriminant analysis (LDA) using the lda() function of the MASS package in R. For each talker we computed the LDA coefficient for each cue. We then compared the correlations between the coefficients for each cue pair. Unlike Schultz et al. who found a significant negative correlation between onset f0 and VOT, we found a positive correlation that was marginally significant (R=0.408, p = 0.075). We also found a positive correlation between vowel duration and onset f0 coefficients (R=0.510, p = 0.021).

Table 7: Correlations between Cohen's D values or LDA Coefficients calculated for each talker and each cue.

Cohen's D	Pearson *R*	*p*	*N*
VOT/f0	0.270	0.249	20
VOT/Vowel duration	0.109	0.646	20
Vowel duration/f0	-0.192	0.417	20
Linear Discriminant Analysis	Pearson *R*	*p*	*N*
VOT/f0	0.408	0.075	20
VOT/Vowel duration	0.352	0.128	20
Vowel duration/f0	0.510	0.021	20

4.0 Discussion

We investigated three cues to voicing in word-initial labial stops. As expected, voiceless stops had longer VOT, higher f0 and shorter following vowels than voiced stops. Of interest, however, were the relationships within each voicing category and within and between talkers. In other words, we were interested in whether the variation was systematic on a token-by-token basis and could be attributed to factors intrinsic to the production system, or whether it was systematic on a talker-by-talker basis and could be attributed to individual differences. Other than speaking rate effects, our evidence was more in favour of the latter.

We found no evidence for intrinsic relationships between the cues that were likely to be traced back to articulatory interactions (within-category correlations in the same direction as between-category correlations as in Panel A of Figure 1). Even VOT and f0 which have been argued to be related for physiological reasons at the level of the musculature of the larynx (Hombert et al.,1979; Ohde, 1984), only displayed within-category token-by-token correlations for the /b/ productions which were in the opposite direction from the overall pattern, in other words they were not in the direction

predicted by the physiological mechanism. Of course, lack of an effect should be treated with caution, especially given other results in the literature. However, some previous results should also be treated with caution, as they did not account for individual differences in mean values.

We also found little evidence that talkers were compensating for ambiguous cues in one dimension with prototypical cues in the other dimension (i.e. within-category correlations in the opposite direction as between-category correlations as in Panel C of Figure 1, however note that very few VOT values could be considered ambiguous, weakening the possibility of testing this hypothesis). The one exception is the VOT/onset f0 relationship for /b/ productions discussed above. Instead, the token-by-token relationships we observed can be best understood as arising from more hyper-articulation with slower speaking rate. This was observed in the vowel duration effects on VOT. With slower speaking rate (longer vowels), VOT was longer for /p/ productions and shorter for /b/ productions. This is consistent with previous studies that reported longer VOTs for (long-lag) voiceless stops with slower speaking rate (Kessinger & Blumstein, 1997; 1998; Miller et al., 1986; Port, 1981); however, it is in contrast to previous findings that reported no changes in VOT for (short-lag) voiced stops (Kessinger & Blumstein, 1997) or an increase in VOT for (short-lag) voiced stops (Miller et al., 1986; Boucher, 2002) with slower speaking rate. Previous results had also reported that the ratio of VOT to syllable duration was constant across syllable durations (i.e. ratio~syllable duration had a slope of zero, Boucher, 2002). A simple regression on our data found this not to be the case. There were significant negative slopes for both /b/ and /p/ models ($p<0.05$). The shorter VOTs in voiced stops with slower speaking rates and the fact that the VOTs could not be fully explained by vowel

duration suggest that our talkers were deliberately hyper-articulating the VOT cue. Recent studies have found that voiceless stops are produced with longer VOTs when a minimal pair is salient (Baese-Berke & Goldrick, 2009) even after controlling for speaking rate (Buz, Jaeger & Tanenhaus, 2014). Like many classic studies, our words were also produced in the presence of salient minimal pairs which might have triggered this targeted hyper-articulation.

Previous studies had found significant individual talker differences in the values of individual cues (e.g. Allen et al., 2003; Newman et al., 2001). Within VOT it has been shown that some of these differences may be systematic across stop categories (Chodroff, Godfrey, Khudanpur, & Wilson, 2015; Scobbie, 2006). We extended this to multiple cues to the same contrast. The analyses reported in Table 2 also showed that talkers differed from each other significantly in VOT for both /b/ and /p/ (controlling for speaking rate), and f0 for /b/ (controlling for speaking rate, and talker mean f0). Furthermore, these differences were systematic. Some talkers used more prototypical values (longer VOT, shorter vowels for /p/, shorter VOT, longer vowels for /b/) than others. This suggests that some of the variation between talkers, at least in cues to voicing, is due to speaking style or degree of hyper-articulation. Studies of clear versus conversational style speech have generally focused on global variables (see Smiljanic and Bradlow, 2009 for a review). Krause and Braida (2004) examined specific phonetic cues and found mixed results for the effect of clear speech on VOT, however, this study only included two talkers. Thus more would be required before classifying some speakers as using "clear speech". What is also unclear from this data is whether the differences between talkers seen here are global properties of these talkers, or simply reflect how they decided to act in this particular task. Recent work on day-to-day

fluctuations in talker VOT extracted from a large corpus suggests that individual talkers' VOTs do shift substantially from day to day while often remaining stable over longer time scales (Sonderegger, 2015). The individuals in this study were also not from a homogenous dialect region. It is possible that the differences we observed are due to dialect differences (see Scobbie 2006 for examples of dialect differences in Scottish VOT). Nonetheless, studies with homogenous dialect groups have also found individual differences (e.g. Hazan & Baker, 2011). Furthermore, the fact that the differences observed here line up with patterns of hyperarticulation makes the dialect explanation less plausible.

As expected, our analysis of Cohen's D (degree of overlap between the categories) found that VOT was by far the strongest cue to this contrast. Interestingly we also found that it varied much more between talkers in degree of overlap than the other two cues. This suggests that while all cues are affected by speaking style, the strongest cues may be affected most while secondary cues are more robust. In line with this observation, Lisker and Abramson (1967) found that differences between stop voicing categories decreased in carrier phrases relative to citation form speech, suggesting more hypo-articulation in carrier phrases. Ohde (1984) also found that VOT was more affected than f0. An intriguing implication of this pattern is a possible role in sound change. One theory about the origin of tonogenesis is that it arises from differences in f0 due to voicing taking over the contrastive role from VOT (Hombert et al., 1979) in a process often called phonologization. Hypo-articulation of high frequency words has long been hypothesized to be a driver of sound change (Lindblom, Guion, Hura, Moon, & Willerman, 1995) along with enhancement of secondary cues (Kirby, 2013). Kirby (2010; 2013) argues through computational simulations that both loss of precision in the

primary cue (decrease in Cohen's D) and enhancement of a secondary cue (increase in Cohen's D) is required to account for phonologization. He proposes a probabilistic enhancement hypothesis in which cues are enhanced over time in direct proportion to their precision as measured by Cohen's D. Thus, if the differential effect of hyper-articulation on precision between VOT and onset f0 observed here is also in effect in high versus low frequency words (in English as well as in languages undergoing tonogenesis like Seoul Korean), then high frequency words provide exactly the conditions for Kirby's probabilistic enhancement of f0. Bang, Sonderegger, Kang, Clayards, and Yoon (2015) tested this prediction by examining word frequency effects in Seoul Korean and found evidence that enhancement of f0 proceeds in parallel with weakening of VOT in high frequency words over time, just as predicted. Future work should test whether onset f0 as a cue to voicing and secondary cues in general are more robust to hypo-articulation than primary cues and in particular if word frequency has the effect of reducing differences in precision between cues.

5.0 Conclusions

We investigated multiple cues to the word-initial /p/-/b/ contrast in English across tokens and across talkers. We found that within-category, token-by-token variability was best characterized as more hyper-articulated productions for slower speaking rates. After accounting for differences in talker means the only correlations between cues on a token-by-token basis was a negative correlation between onset f0 and VOT for /b/ productions. We also found significant inter-talker variation. This variation was systematic across cues such that some talkers produced more prototypical/hyper-articulated speech in multiple dimensions but VOT varied most between talkers in

degree of hyper-articulation. Future research should determine whether these talker differences extend to other sound contrasts within the same individuals.

Acknowledgements

A portion of this data was part of the author's PhD thesis and was presented at the LSA meeting in San Francisco in January 2009. Data collection was supported by NIH research Grant DC-005071 to Michael K. Tanenhaus and Richard N. Aslin and by FQRSC grant 145433 to the author.

References

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America*. *113* (1), 544-522.

Baayen, Harald R. (2009). LanguageR. R package. http://CRAN.R-project.org/package=language R

Baese-Berk, M., & Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and cognitive processes*, *24*(4), 527-554.

Bang, H.-Y., Sonderegger, M. Kang. Y., Clayards, M., Yoon, T.-J. (2015). The effect of word frequency on the time-course of tonogenesis in Seoul Korean. *Proceedings of the 18th International Congress of Phonetic Sciences,* Glasgow, UK.

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68, 255-278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7, http://CRAN.R-project.org/package=lme4.

Bailey, P. J., & Summerfield, Q. (1980). Information in speech: observations on the perception of [s]-stop clusters. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 536-563.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*, 592–597.

Boersma, P., & Weenink, D. (2011). *Praat: doing phonetics by computer* (Version 4.6.09)

Boucher, V. J. (2002). Timing relations in speech and the identification of voice-onset times: A stable perceptual boundary for voicing catgories across speaking rates. *Perception and Psychophysics, 64*(1), 121-130.

Buzz, E., Jaeger, F., & Tanenhaus, M. K. (2014). Contextual confusability leads to targeted hyperarticulation. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

Chodroff, E., Godfrey, J., Khudanpur, S., & Wilson, C. (2015) Structured variability in acoustic realization: a corpus study of voice onset time in American English stops. In *Proceedings of the 18th International Congress of Phonetic Sciences.*

Clayards, M. (2008) *The ideal listener: Making optimal use of acoustic cues for speech perception*. PhD Thesis, University of Rochester

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804-809.

Cole, J., Kim, H., Choi, H., & Hasegawa-Johnson, M. (2007). Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics*, *35*(2), 180-209.

Dmitrieva, O., Llanos, F., Shultz, A. A., & Francis, A. L. (2015). Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English. *Journal of Phonetics*, *49*, 77-95.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224-238.

Ferguson, S. H. (2004). Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. *The Journal of the Acoustical Society of America*, *116*(4), 2365-2373.

Fox, N. P., Reilly, M., & Blumstein, S. E. (2015). Phonological neighborhood competition affects spoken word production irrespective of sentential context.*Journal of Memory and Language*, *83*, 97-117.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition*, *19*(5), 448-458.

Goldinger, S., & Summers, W. (1989). Lexical neighborhoods in speech production: A first report. *Research on Speech Perception Progress Report*. Psychology Department, Speech Research Laboratory, Indiana University, 15, 331–342.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hazan, V., Baker, R., Lee, W. S., & Zee, E. (2011, August). Is consonant perception linked to within-category dispersion or across-category distance. In *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 839-842).

Hazan, V., & Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, *116*(5), 3108-3118.

Holt, L., Lotto, A., & Kluender, K. (2001). Influence of fundamental frequency on stop-consonant voicing perception: A case of learned covariation or auditory enhancement? *The Journal of the Acoustical Society of America, 109*, 764-774.

Hombert, J. M., Ohala, J. J., & Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language*, 37-58.

Hoole, P., Honda, K., Murano, E., Fuchs, S., & Pape, D. (2004). Cricothyroid activity in consonant voicing and vowel intrinsic pitch. In *Proceedings of the Conference on Voice Physiology and Biomechanics, Marseille*.

Hoole, P., & Honda, K. (2011). Automaticity vs. feature-enhancement in the control of segmental F0. In Clements, G. Nick, and Ridouane, Rachid, Eds *Where do phonological features come from*: *Cognitive, physical and developmental bases of distinctive speech categories*. Amsterdam, NLD: John Benjamins Publishing Company, 131-171.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America, 25*(1), 105-113.

Johnson, K., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *The Journal of the Acoustical Society of America*, *94*(2), 701-714.

Kessinger, R. H., & Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French and English. *Journal of Phonetics, 25*, 143-168.

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.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language, 70*(2), 419-454.

Kingston, J., Diehl, R. L., Kirk, C. J., & Castleman, W. A. (2008). On the internal perceptual structure of distinctive features: The [voice] contrast. *Journal of Phonetics, 36*, 28-54.

Kirby, J., (2010). *Cue selection and category restructuring in sound change*. PhD thesis, University of Chicago.

Kirby, J., (2013) The role of probabilistic enhancement in phonologization. In: Yu, A., (ed), *Origins of sound patterns: approaches to phonologization.* Oxford: Oxford University Press, 228–246.

Kirby, J. P., & Ladd, D. R. (2015) Stop voicing and f0 perturbations: Evidence from French and Italian. *Proceedings of the 18th International Congress of Phonetic Sciences,* Glasgow, UK.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2009). Experimental effects and individual differences in linear mixed models: estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, 1, 238-238.

Kohler K. J. 1979. Dimensions in the perception of fortis and lenis plosives. *Phonetica* 36.332-43.

Krause, J. C., & Braida, L. D. (2009). Evaluating the role of spectral and envelope characteristics in the intelligibility advantage of clear speech. *The Journal of the Acoustical Society of America*, *125*(5), 3346-3357.

Lehiste, I., & Peterson, G. E. (1961). Some basic considerations in the analysis of intonation. *The Journal of the Acoustical Society of America*, *33*(4), 419-425.

Lindblom, B., Guion, S., Hura, S., Moon, S.-J., Willerman, R. 1995. Is sound change adaptive? *Rivista di linguistica* 7, 5–36.

Lisker, L. (1986). “Voicing” in English: a catalogue of acoustic features signaling/b/versus/p/in trochees. *Language and Speech*, *29*(1), 3-11.

Lisker, L., & Abrahmson, A. S. (1964). Cross-language study of voicing in initial stops. *Word, 20*, 384-422.

Lisker, L. & Abrahmson, A. S. (1967). Some effect of context on Voice Onset Time in English Stops. *Language and Speech*, 10, 1-28.

Mack, M. (1982). Voicing-dependent vowel duration in English and French: Monolingual and bilingual production. *The Journal of the Acoustical Society of America*, *71*(1), 173-178.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the /sh/-/s/ distinction. *Perception and Psychophysics, 28*(3), 213-228.

Matuschek, H., Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Balancing type I error and power in linear mixed models. http://arxiv.org/abs/1511.01864.

McMurray, B., Clayards, M. A., Tanenhaus, M. K., & Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin & Review*, *15*(6), 1064-1071.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118* (2), 219.

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

Nearey, T. (1997). Speech perception as pattern recognition. *The Journal of the Acoustical Society of America, 101,* 3241-3254.

Nearey, T. & Hogan, J.T. (1986). Phonological contrast in experimental phonetics: Relating distributions of production data to perceptual categorization curves. In J.J. Ohala & J.J. Jaeger (Eds.) *Experimental Phonology* (pp. 121-162). Orlando: Academic Press.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3), 1181-1196.

Noiray, A., Iskarous, K., & Whalen, D. H. (2014). Variability in English vowels is comparable in articulation and acoustics. *Laboratory Phonology,* 5, 271-288. doi: 10.1515/lp-2014-0010

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech.*Cognitive Psychology*, *47*(2), 204-238.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355-376.

Ohde, R. N. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *The Journal of the Acoustical Society of America, 75*, 224.

Port, R. F. (1981). Linguistic timing factors in combination. *The Journal of the Acoustical Society of America, 69*(1), 262-274.

Port, R. F., & Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception and Psychophysics, 32*, 141-152.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raphael, L. J. (2005). Acoustic cues to the perception of segmental phonemes. *The Handbook of Speech Perception*, 182-206.

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

Scobbie, J. M. (2006). Flexibility in the face of incompatible English VOT systems. In: *Laboratory Phonology 8 Varieties of Phonological Competence. Phonology and Phonetics* 4-2 . Mouton de Gruyter, Berlin, pp. 367-392. ISBN 978-3110176780

Shultz, A. A., Francis, A. L., & Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *The Journal of the Acoustical Society of America*, *132*(2), EL95-EL101.

Smiljanic´, R., and Bradlow, A. R. (2009). "Speaking and hearing clearly: Talker and listener factors in speaking style changes," Linguist. Lang. Compass 3, 236–264.

Slis, I. H., & Cohen, A. (1969). On the complex regulating the voiced-voiceless distinction. *Language and Speech, 12*(2), 80-102.

Sonderegger, M. (2015) Trajectories of voice onset time in spontaneous speech on reality TV. *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK

Sussman, H. M., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: the orderly output constraint. *The Behavioral and Brain Sciences*, *21*(2), 241–259

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *Journal of the Acoustical Society of America*, *125*(6), 3974-3982.

Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. The Journal of the Acoustical Society of America, 128(4), 2090-2099.

Theodore, R. M., Myers, E. B., & Lomibao, J. A. (2015). Talker-specific influences on phonetic category structure. *The Journal of the Acoustical Society of America*, *138*(2), 1068-1078.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434-464

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, *74*(6), 1284-1301.

Umeda, N. (1981). Influence of segmental factors on fundamental frequency in fluent speech. *The Journal of the Acoustical Society of America*, *70*(2), 350-355.

Wang, W., & Fillmore, C. (1961). Intrinsic cues and consonant perception. *Journal of Speech and Hearing Research, 4*(2), 130-136.

Figure Legends

Figure 1: Hypothetical distributions of two cues to two categories showing three possible relationships between the cues [A] token-by-token within-category correlations with a possible intrinsic source, [B] uncorrelated within-category, [C] within-category correlations with "cue trading". Ellipses represent equal probability lines of multivariate distributions. Distributions for each individual dimension are represented as continuous histograms above and to the right of the three panels.

Figure 2: Density plots smoothed with a Gaussian kernel showing the distribution of values for each cue for each category pooled across all speakers. Black lines are /b/ productions, grey lines are /p/ productions. Onset f0 is normalized relative to talker mean onset f0. Vowel duration is normalized for voicing of the following consonant and talker mean duration.

Figure 3: Scatter plots of individual tokens for VOT, talker-relative f0 and final-voicing and talker-relative vowel duration. Black circles are /b/ tokens and grey circles are /p/ tokens.

Figure 4: Talker means for each voicing category for VOT, final-voicing-and-talker-relative vowel duration and talker-relative f0. Black circles are /b/ productions and grey circles are /p/ productions.

Figure 5: Cohen's D values of each cue for each talker.

Table Headings

Table 1: Means for each of the cues by voicing category and talker gender. Mean for vowel duration also grouped by word-final voicing.

Table 2: Pearson R correlations for VOT, onset f0 and final voicing and talker corrected vowel duration for each voicing category as well as both categories together. VOT values for /b/ are split into positive (+ve) and negative (-ve) VOT values.

Table 3. Results from model comparison tests of talker differences for each cue and category for /p/ productions (Left) and /b/ productions (Right). Models were lmer(X ~ {vowel_length} + rep + (1+{vowel_length} +rep ||talker) + (1|word)). /b/ productions include only positive VOTs (no pre-voicing)

Table 4: Model output from lmer(VOT ~ vowel duration mean + vowel duration deviation + onset f0 mean + onset f0 deviation + repetition + (1 + vowel duration deviation + onset f0 deviation||talker) + (1|word)) for data from each voicing category.

Table 5: Model output from lmer(onset f0 ~ mean_vowel_length + vowel_length_dev + mean VOT + VOT deviation + rep + (1 + vowel_length_dev+VOT deviation+rep||talker) + (1|word)) for data from each voicing category (positive VOTs only).

Table 6: Model output from lmer(relative vowel duration ~ mean f0 + f0 deviation + mean VOT + VOT deviation + repetition + (1 + f0 deviation + VOT deviation ||talker) + (1|word)) for data from each voicing category (positive VOTs only).

Table 7: Correlations between Cohen's D values or LDA Coefficients calculated for each talker and each cue.

Tables

Table 1: Means for each of the cues by voicing category and talker gender. Mean for vowel duration also grouped by word-final voicing.

Gender		VOT (ms)	f0 (Hz)	Vowel (ms)	Final Voicing		Vowel (ms)
/b/	f	2.4	225.4	206.1	/b/	Voiced	296.1
	m	7.5	116.2	182.4		Voiceless	144.0
/p/	f	76.1	245.9	172.4	/p/	Voiced	266.9
	m	68.0	125.3	162.5		Voiceless	119.8

Table 2: Pearson R correlations for VOT, onset f0 and final voicing and talker corrected vowel duration for each voicing category as well as both categories together. VOT values for /b/ are split into positive (+ve) and negative (-ve) VOT values.

	Data	Pearson r	*p*	N
f0/VOT	All	0.309	0.000	1271
	/p/	0.000	0.995	652
	/b/+ve	-0.111	0.008	573
	/b/-ve	-0.208	0.160	47
Vowel duration /VOT	All	-0.280	0.000	1366
	/p/	0.054	0.158	686
	/b/+ve	-0.155	0.000	635
	/b/-ve	-0.022	0.886	46
f0/ Vowel duration	All	-0.079	0.005	1269
	/p/	0.116	0.003	652
	/b/	0.019	0.630	617

Table 3. Results from model comparison tests of talker differences for each cue and category for /p/ productions (Left) and /b/ productions (Right). Models were lmer(X ~ {vowel_length} + rep + (1+{vowel_length} +rep ||talker) + (1|word)). /b/ productions include only positive VOTs (no pre-voicing)

/p/		*χ2 (df)*	*p*	/b/		*χ2 (df)*	*p*
VOT	Intercept	210.1 (1)	0.00	VOT	Intercept	66.6 (1)	0.00
	Vwl len.	1.4 (1)	0.23		Vwl len.	0 (1)	1.00
	Rep.	0 (1)	1.00		Rep	0.2 (1)	0.68
f0	Intercept	0 (1)	1.00	f0	Intercept	4.3 (1)	0.04
	Vwl len.	8.8 (1)	0.00		Vwl len.	4.5 (1)	0.03
	Rep.	121.6 (1)	0.00		Rep	25.8 (1)	0.00
Vwl	Intercept	0 (1)	1.00	Vwl	Intercept	0 (1)	1.00
	Rep.	53.3 (1)	0.00		Rep	57.3 (1)	0.00

Table 4: Model output from lmer(VOT ~ vowel duration mean + vowel duration deviation + onset f0 mean + onset f0 deviation + repetition + (1 + vowel duration deviation + onset f0 deviation||talker) + (1|word)) for data from each voicing category.

		Estimate	Std.Error	t.value	p
/p/	Intercept	74.94	3.08	24.36	0.00
	Vowel duration mean	-5.15	2.17	-2.38	0.02
	Vowel duration deviation	1.51	0.58	2.60	0.01
	Onset f0 mean	-1.20	2.50	-0.48	0.63
	Onset f0 deviation	-1.05	1.11	-0.94	0.35
	Repetition	-0.48	0.73	-0.65	0.51
/b/	Intercept	12.23	0.66	18.40	0.00
	Vowel duration mean	-0.18	0.51	-0.35	0.72
	Vowel duration deviation	-0.57	0.22	-2.61	0.01
	Onset f0 mean	0.08	0.56	0.15	0.88
	Onset f0 deviation	-0.53	0.32	-1.66	0.10
	Repetition	0.05	0.27	0.19	0.85

Table 5: Model output from lmer(onset f0 ~ vowel duration mean + vowel duration deviation + VOT mean + VOT deviation + rep + (1 + vowel duration deviation+VOT

deviation+rep||talker) + (1|word)) for data from each voicing category (positive VOTs only).

		Estimate	Std.Error	t.value	p
/p/	Intercept	0.52	0.09	5.56	0.00
	VOT mean	-0.08	0.06	-1.31	0.19
	VOT deviation	-0.07	0.10	-0.66	0.51
	Vowel duration mean	-0.04	0.06	-0.69	0.49
	Vowel duration deviation	0.05	0.06	0.86	0.39
	Repetition	-0.59	0.21	-2.88	0.00
/b/	Intercept	-0.86	0.12	-7.26	0.00
	VOT mean	-0.06	0.09	-0.59	0.56
	VOT deviation	-0.20	0.09	-2.20	0.03
	Vowel duration mean	-0.12	0.09	-1.26	0.21
	Vowel duration deviation	-0.05	0.09	-0.55	0.58
	Repetition	-0.33	0.18	-1.86	0.06

Table 6: Model output from lmer(relative vowel duration ~ onset f0 mean + onset f0 deviation + VOT mean + VOT deviation + repetition + (1 + onset f0 deviation + VOT deviation ||talker) + (1|word)) for data from each voicing category (positive VOTs only).

		Estimate	Std.Error	t.value	p
/p/	Intercept	-12.10	1.42	-8.55	0.00
	VOT mean	-1.82	1.41	-1.29	0.20
	VOT deviation	5.10	2.49	2.05	0.04
	Onset f0 mean	0.50	1.47	0.34	0.73
	Onset f0 deviation	1.62	1.55	1.05	0.30
	Repetition	3.50	3.08	1.14	0.26
/b/	Intercept	13.30	2.37	5.61	0.00
	VOT mean	-0.12	1.43	-0.08	0.93
	VOT deviation	-3.86	3.47	-1.11	0.27
	Onset f0 mean	-2.35	1.53	-1.53	0.13
	Onset f0 deviation	-0.01	1.79	0.00	1.00
	Repetition	2.47	2.20	1.12	0.26

Table 7: Correlations between Cohen's D values or LDA Coefficients calculated for each talker and each cue.

Cohen's D	Pearson *R*	*p*	*N*
VOT/f0	0.270	0.249	20
VOT/Vowel duration	0.109	0.646	20
Vowel duration/f0	-0.192	0.417	20
Linear Discriminant Analysis	Pearson *R*	*p*	*N*
VOT/f0	0.408	0.075	20
VOT/Vowel duration	0.352	0.128	20
Vowel duration/f0	0.510	0.021	20