STRUCTURED VARIABILITY IN ACOUSTIC REALIZATION: A CORPUS STUDY OF VOICE ONSET TIME IN AMERICAN ENGLISH STOPS

Eleanor Chodroff, John Godfrey, Sanjeev Khudanpur, and Colin Wilson

Johns Hopkins University
chodroff@cogsci.jhu.edu, godfrey.jack@gmail.com, khudanpur@jhu.edu, wilson@cogsci.jhu.edu

ABSTRACT

Talkers differ greatly in the acoustic realization of speech sounds, a source of signal variation that must be overcome by human and machine listeners. The present study examined talker variability in voice onset time (VOT) across the six word-initial stop consonant categories (/ptkbdg/) of American English. Employing a large corpus of productions from more than 100 speakers, we replicated previous findings of significant variation in overall and stop-specific VOT means. However, we also identified several statistical generalizations within and across phonetic patterns of individual talkers. Speaking rate accounted for a large portion of VOT variance, with talkers differing considerably in the strength of this relationship. Stop category means showed high pairwise correlations, particularly among /ptk/. Additionally, stop-specific means and variances were highly correlated. The structured variation present in VOT could be exploited by both listeners and automatic recognition systems to facilitate robust perceptual adaptation from limited exposure to novel talkers.

Keywords: talker variation, voice onset time, stop consonants, corpus phonetics

1. INTRODUCTION

The acoustic realization of a given speech category varies substantially across talkers, as demonstrated by studies of vowels [24], fricatives [16], and stop consonants [1, 31]. Talker variability presents a challenge to human and machine speech recognition: for example, humans show impaired identification of isolated vowels [2] and entire words [20] when stimuli are presented from multiple talkers. Nevertheless, listeners display a remarkable ability to adapt to specific talkers and generalize talker characteristics across sound categories. Models of talker adaptation aim to account for these abilities, and the presence of structured variation across talkers may play a crucial role in adaptation and generalization mechanisms.

Individual differences in speech production have been noted for various speech categories, with a clear example observed in vowel production. Differences in the frequencies of F1 and F2 across 10 vowels and 76 speakers were examined in [24]. This classic study found wide variation in the population, particularly with respect to gender and age, but relatively less variation internal to individual speakers. In particular, the "vowel systems" of different speakers form overlapping but generally congruent shapes in F1-F2 space. This result is well explained by the dependence of formant frequencies on talker-specific vocal tract length, shape, and size; crucially, this finding shows that a structured vowel space emerges within and across speakers. To the extent that an individual's vowel space resembles that of the population, a listener may abstract over talker-specific formant shifts to infer category labels (e.g. [12, 26, 27]).

Similar to findings in vowel production, individual differences have also been noted in fricatives. [16] assessed spectral center of gravity (COG) in the /s/-/ʃ/ distinction in CV syllables produced by 20 speakers. While talkers varied in their mean, variance, and distributional overlap of /s/ and /ʃ/ COG production, the COG for /s/ was found to be consistently higher than that for /ʃ/. Furthermore, the majority of talkers showed relatively little distributional overlap in their /s/-/ʃ/ production. These patterns reveal systematicity in the acoustic realization of the contrast, and support a mechanism of adaptation that targets the COG dimension as opposed to /s/ and /ʃ/ separately.

Individuals also vary significantly in average voice onset time (VOT) and associated durational properties such as overall speaking rate. [1] examined 8 speakers' production of word-initial, voiceless stop consonants in CVC words. Even after controlling for speaking rate, individual differences in VOT remained. Expanding on this study, [31] controlled for speaking rate and place of articulation. Ten speakers

produced repeated CV syllables (/pi/, /ti/, /ki/) at various rates of speech. The effect of speaking rate on VOT varied significantly across talkers; however, the magnitude of the effect was consistent across place of articulation within a speaker, suggesting at least some within-talker regularity in VOT production.

Previous studies have provided data on population VOT values [6, 15, 32, 34]; however, many studies do not examine talker-specific variation (cf. [9, 32]). While [6] surveyed over 600 speakers in the TIMIT corpus [10], that corpus provides too little data from individual speakers to allow for robust estimation of talker-specific VOT parameters (cf. estimation effects of gender). In [32], VOT was automatically measured from 19 speakers in the Buckeye corpus [25]. Similarly, [9] measured VOT from four speakers in the Boston University Radio News corpus [21]. While both studies observed considerable individual differences, generalizations of cross-talker phonetic patterns were limited by the statistical analyses and/or sample size. In contrast to the previous studies, the current study examined both talker variability as well as systematicity in the VOT of word-initial, prevocalic stop consonants in American English read speech. We used a large corpus and mostly automatic methods. This allowed us to measure approximately 68,000 tokens from more than 100 speakers, including all six stops (/ptkbdg/) in many sentential contexts.

The present more extensive study provides further insight into the range and limits of talker variability across stop consonant categories. Significant individual differences were identified in stop-specific VOT means and in the effect of speaking rate on VOT. Additionally, structured variation was observed in pairwise correlations of stop-specific talker means and standard deviations. The presence of this structure transcends individual variability and may facilitate talker adaptation by both humans and machines.

2. METHODS

2.1. Corpus description

The Mixer 6 Corpus contains speech from over 600 talkers, recorded over three separate sessions [5]. In each session, an interview, transcript reading, and phone call were recorded. The following analysis employs the transcript portion of 129 native English speakers, for which three sessions were recorded. The transcript contains randomly selected utterances from previously collected spontaneous speech in the Switchboard corpus [11]. These are therefore naturally occurring sentences, not specifically constructed for this study. For each session, the talker recorded up to 15 minutes of 335 different utterances, read in order. The median utterance length was 7 words (range: 1-17).

All of the speakers studied here were born in the United States, and approximately half in the Philadelphia region. Sixty-eight speakers were from Pennsylvania, 32 from other mid-Atlantic and New England regions and 29 from other areas of the United States. Ages ranged from 19 to 87 years old (median: 27). Speaker gender was roughly balanced (60 male).

2.2. Corpus preparation

The transcript portion of the corpus was audited for reading and recording errors with automatic and manual methods. Each session was submitted to HTK-based automatic speech recognition with acoustic models trained on the Wall Street Journal corpus [7]. The output was evaluated against the intended speech of the transcript using the NIST Score-Lite system [19], which produced a word error rate for each time-aligned sentence. Sentences with less than 100% accuracy were manually audited through listening.

The cleaned version of the transcript was force-aligned to its corresponding audio segment with the Penn Phonetics Lab Forced Aligner (PFA) [33]. Word-initial, prevocalic stop consonants were located and further processed with AutoVOT [13, 29]. AutoVOT uses subphonemic processing to identify the stop release and vocalic offset of stop consonants, corresponding to standard boundary locations for positive VOT measurements. The window of analysis was extended from the PFA boundaries 30 ms in both directions for voiceless stops and 10 ms in both directions for voiced stops. The minimum VOT duration was set to 15 ms for the voiceless stops and 4 ms for the voiced stops. In addition to the automatic measurements, the VOT of approximately 3,000 stop consonants was manually measured. Comparison with the AutoVOT output yielded a root mean square deviation of 12.9 ms (cf. [28]). As described in section 2.3, outliers were removed to minimize noise in the dataset. The automatic measurement may be further improved by training a corpus-specific acoustic classifier.

2.3. Acoustic analysis

VOT was measured as the duration between the AutoVOT-defined stop consonant boundaries or, if

available, manually placed boundaries. Speaking rate was measured as the mean word duration per utterance from the PFA boundaries using Praat [4]. Only stop consonants followed by a primary-stressed vowel were analyzed, and with the exception of '*to*', function words were retained in the analysis. VOTs 2.5 standard deviations away from the group category mean were considered outliers and excluded from analysis. There was a total of 68,456 stop consonants, or an average of 531 stop consonants per talker (range: 320-726). For each talker there were 46-100 tokens of /p/ (median: 72), 18-78 tokens of /t/ (median: 46), 56-117 tokens of /k/ (median: 90), 72-133 tokens of /b/ (median: 100), 68-191 tokens of /d/ (median: 140), and 59-118 tokens of /g/ (median: 92).

3. RESULTS

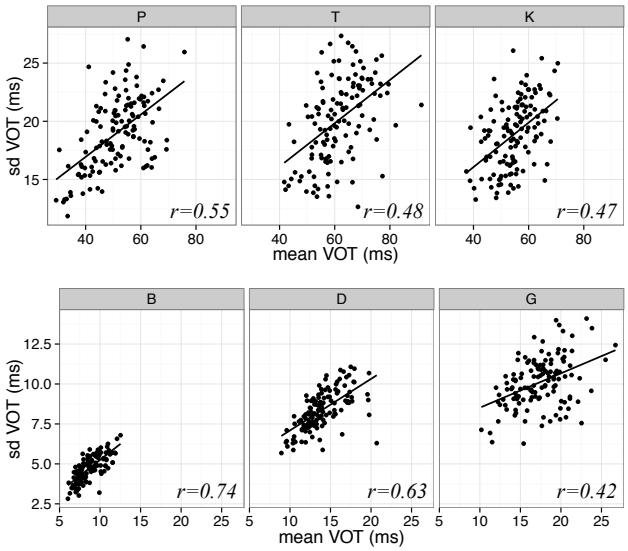
Average VOT was calculated both across and within talkers for each stop category. Population values are shown in Table 1. The ranking of VOT means across stop consonants was largely in agreement with early studies on VOT [15, 34], larger corpus studies [6, 32], as well as cross-linguistic patterns [8]. However, [6, 15, 34] found that /k/ had a marginally longer mean than /t/ (see also [8]), whereas the present study reports the opposite pattern (see also [32]). In detail, 74 speakers had an increasing VOT ranking of /b-d-g-p-k-t/, 23 had /b-d-g-k-p-t/, and 14 had /b-g-d-p-k-t/, with the remaining 16 showing various other rankings. As seen in Figure 1, considerable individual differences were also found in stop-specific means and standard deviations; however, these two parameters are strongly correlated within each stop category (/p/: $r = 0.55$; /t/: $r = 0.48$; /k/: $r = 0.47$; /b/: $r = 0.74$; /d/: $r = 0.63$; /g/: $r = 0.42$; all categories collapsed: $r = 0.93$; ps < 0.001). These correlations suggest structured rather than unrestricted variation in VOT patterns.

Table 1. Population means and standard deviations of VOT (ms).

Stop	Mean	*SD*
P	50.8	21.1
T	60.5	21.8
K	54.4	20.2
B	8.7	5.0
D	13.8	8.7
G	17.2	10.6

A linear mixed-effects model [3] was used to analyze the influence of voice, place of articulation, and speaking rate on VOT in ms (Model 1). Model 1 included fixed and interacted effects of voice (sum-coded, voiceless = +1) and place of articulation (sum-coded, labial reference category), a main effect of speaking rate (in s), maximal random-effect structure for talker, and a random intercept for word. For the mixed-effects analyses, a t-value greater than 2.0 was considered significant (but many effects of interest far surpassed that criterion). There were significant main effects of voice ($\beta = 21.54$, $t = 31.23$), place (*dorsal*: $\beta = 3.62$, $t = 4.10$), and speaking rate ($\beta = 21.50$, $t = 13.38$); however, the main effect of coronal place was not significant ($\beta = 1.50$, $t = 1.67$). As estimated by the model, voiceless stops were approximately 42 ms longer than voiced, dorsal stops were 8 ms longer than labials, and coronal stops were 3 ms longer (n.s.). VOT also increased by about 2.1 ms for every 100 ms increase in average word duration. Interactions of place and voice were not significant (*dorsal* × *voice*: $\beta = -1.15$, $t = -1.31$; *coronal* × *voice*: $\beta = 1.12$, $t = 1.26$).

Figure 1. Talkers' mean VOT (ms) plotted against corresponding sd.



Additional mixed-effects models of VOT were used to further explore individual variability. Each subsequent model had the same fixed structure as Model 1, differing only in the talker random effects structure. Model 2 eliminated all random effects for talker; Model 3 allowed only the VOT intercept to vary by talker; Model 4 estimated an intercept and separate slopes of place and voice for talker; and Model 5 added the interaction of place and voice to the random structure of Model 4, and thus differed

from Model 1 only in excluding the random slope for speaking rate.

Model 1 significantly outperformed all of the other models on the likelihood ratio test (M2: $\chi^2(28) = 9620$; M3: $\chi^2(27) = 5860$; M4: $\chi^2(18) = 582.5$; M5: $\chi^2(7) = 86$; ps < 0.0001) as well as the more conservative BIC comparison (M1: 540940; M2: 550248; M3: 546499; M4: 541322; M5: 540948). This result reveals significant individual variation in the effect of each stop category (place x voice) on VOT, as well as individual variation in the effect of speaking rate. (Note however that the practical significance of each random effect remains to be investigated.)

In spite of significant individual differences, talker-specific category means are strongly correlated (Table 2). This is particularly true for the voiceless stops, and for homorganic coronal and dorsal pairs, indicating that talker-specific effects are similar for these categories. While it may benefit listeners to estimate stop-specific VOTs for individual talkers, adaptation at a more general level (e.g., at the level of overall VOT or speaking rate) could also be effective and would possess greater statistical strength.

Table 2. Correlations of talkers' mean VOTs. The asterisk indicates $p < 0.003$, alpha-corrected for multiple comparisons.

	P	T	K	B	D	G
P	1	0.81*	0.82*	0.09	0.47*	0.25
T	0.81*	1	0.80*	-0.01	0.59*	0.14
K	0.82*	0.80*	1	0.11	0.50*	0.42*
B	0.09	-0.01	0.11	1	0.05	0.43*
D	0.47*	0.59	0.50*	0.05	1	0.43*
G	0.25*	0.14	0.42*	0.43*	0.43*	1

4. CONCLUSION

Talkers vary considerably in their production of VOT in word-initial stop consonants, extending findings from [1] and [31] to more realistic speech and all six stop consonant categories. The best-performing model indicates that full perceptual adaptation would require estimating talker-specific effects on each stop category, as well as on the relationship between VOT and speaking rate. Phoneme identification, however, may not require such precision. Population estimates of stop-specific effects or talker-specific estimates of the individual's offset (i.e., intercept) may be sufficient for accurate classification. The correlations of talker means across stop categories, and with corresponding stop-specific standard deviations, also reveal important structure within talker variation. Recent models of talker adaptation propose a form of incremental adaptation that is generally in line with this proposal (e.g., [14, 18]). As information about a particular speaker accumulates, a listener may refine a default model of VOT using knowledge of population variation and covariation across categories.

Previous perception studies have also demonstrated listeners' ability to generalize talker characteristics across stop consonant categories. Listeners are able to identify that a long /k/ is more characteristic of a talker with a long /p/ even without hearing the talker produce the /k/ category [30]. Furthermore, in imitation, listeners extrapolate a talker's characteristically long VOT of /p/ to /k/, again without prior exposure [17]. While [30] found that listeners are sensitive to the correlation between a talker's short-lag VOTs across categories in perception, [17] did not find this effect in imitation. In the case of imitation, a speaker may be limited in producing a shortened VOT. Nonetheless, the strong correlations of talker means across voiceless stop categories are highly compatible with the perceptual findings of [30] and [17], and with models of generalization across categories [18, 22, 23].

While the present study documented significant structured variation in stop consonant VOTs, there are some limitations to these findings. First, only positive VOTs were analyzed in the corpus. Further measurements may reveal additional structure in both positive and negative VOT (particularly in the voiced stops). Secondly, because VOT is not the only cue to stop consonant voice, correlations with other acoustic-phonetic cues may facilitate talker adaptation and subsequent categorization. Third, the present results were limited to word-initial stops in stressed syllables, and structured variability within and across other context should be investigated. Finally, further research will be necessary to determine how individual variation and regularity in stop consonant VOT are best integrated into formal models of talker adaptation.

5. ACKNOWLEDGMENTS

The authors would like to thank Matthew Maciewejski, Wade Shen, Sharon Tam, Chloe Haviland, Elsheba Abraham, Alessandra Golden, Spandana Mandaloju, and Benjamin Wang for their assistance in data processing. We would also like to thank Matt Goldrick for his useful input on acoustic analysis, and Paul Smolensky for helpful questions and comments. Finally, we acknowledge the DHS-USSS Forensic Services Division for supporting this research.

6. REFERENCES

[1] Allen, J.S., Miller, J.L., DeSteno, D. (2003). Individual talker differences in voice-onset-time. *J. Acoust. Soc. Amer.*, 113(1):544–552.

[2] Assmann, P.F., Nearey, T.M., Hogan, J.T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *J. Acoust. Soc. Amer.*, 71(4):975–989.

[3] Baayen, R.H., Davidson, D.J., Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59(4):390–412.

[4] Boersma, P., Weenink, D. (2014). Praat: Doing phonetics by computer [Computer program]. Version 5.4.01.

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

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

[7] Charniak, E., Blaheta, D., Ge, N., Hall, K., Hale, J., Johnson, M. (2000). BLLIP 1987-89 WSJ Corpus Release 1 LDC2000T43. DVD. Philadelphia: Linguistic Data Consortium.

[8] Cho, T., Ladefoged, P. (1999). Variation and universals in VOT: Evidence from 18 languages. *J. Phon.*, 27(2):207–229.

[9] 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. *J. Phon.,* 35(2):180-209.

[10] Garofolo, J., Lamel, L., Fisher, M., Fiscus, J., Pallett, D., Dahlgren, N. (1993). DARPA TIMIT acoustic phonetic continuous speech corpus. NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD.

[11] Godfrey, J., Holliman, E. (1993). Switchboard-1 Release 2 LDC97S62. DVD. Philadelphia: Linguistic Data Consortium.

[12] Jansen, A., Niyogi, P. (2006). Intrinsic Fourier analysis on the manifold of speech sounds. *Proc. ICASSP,* 1:I-I.

[13] Keshet, J., Sonderegger, M., Knowles, T. (2014). AutoVOT: A tool for automatic measurement of voice onset time using discriminative structured prediction [Computer program]. Version 0.91, retrieved November 2014 from https://github.com/mlml/autovot/.

[14] Kleinschmidt, D.F., Jaeger, T.F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psych. Review*, in press.

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

[16] Newman, R.S., Clouse, S.A., Burnham, J.L. (2001). The perceptual consequences of within-talker variability in fricative production. *J. Acoust. Soc. Amer.*, 109(3):1181–1196.

[17] Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *J. Phon.*, 39(2):132–142.

[18] Nielsen, K., Wilson, C. (2008). A hierarchical Bayesian model of multi-level phonetic imitation. *27th WCCFL*, 335–343.

[19] NIST (2009). Speech recognition scoring toolkit (SCTK). Version 2.4.0. http://www.nist.gov/speech/tools.

[20] Nygaard, L.C., Sommers, M.S., Pisoni, D.B. (1994). Speech perception as a talker-contingent process. *Psych. Sci.*, 5(1):42–46.

[21] Ostendorf, M., Price, P.J., Shattuck-Hufnagel, S. (1995). The Boston University Radio News Corpus. Philadelphia: Linguistic Data Consortium.

[22] Pajak, B., Bicknell, K., Levy, R. (2013). A model of generalization in distributional learning of phonetic categories. *Proc. 4th Workshop on Cog. Modeling and Comp. Ling.*, 11-20.

[23] Pajak, B., Levy, R. (2014). The role of abstraction in non-native speech perception. *J. Phon.,* 46:147-160.

[24] Peterson, G.E., Barney, H.L. (1952). Control methods used in a study of the vowels. *J. Acoust. Soc. Amer.*, 24(2):175–184.

[25] Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Comm.*, *45*(1):89-95.

[26] Plummer, A.R. (2012). Aligning manifolds to model the earliest phonological abstraction in infant-caretaker vocal imitation. *INTERSPEECH*, 2481–2484.

[27] Plummer, A.R., Beckman, M.E., Belkin, M., Fosler-Lussier, E., Munson, B. (2010). Learning speaker normalization using semi-supervised manifold alignment. *INTERSPEECH*, 2918–2921.

[28] Sonderegger, M., Keshet, J. (2012). Automatic measurement of voice onset time using discriminative structured prediction. *J. Acoust. Soc. Amer.*, *132*(6):3965-3979.

[29] Stuart-Smith, J., Sonderegger, M., Rathcke, T., Macdonald, R. (to appear). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Lab. Phon.*

[30] Theodore, R.M., Miller, J.L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *J. Acoust. Soc. Amer.*, 128(4):2090–2099.

[31] Theodore, R.M., Miller, J.L., DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *J. Acoust. Soc. Amer.*, 125(6):3974–3982.

[32] Yao, Y. (2007). Closure duration and VOT of word-initial voiceless plosives in English in spontaneous connected speech. *UC Berkeley Phonology Lab Annual Report*, 183-225.

[33] Yuan, J., Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Proc. Acoustics*.

[34] Zue, V. W. (1976). Acoustic characteristics of stop consonants: A controlled study (No. MIT-TR-523). *MIT LLL.*