CONSTRAINTS ON VARIABILITY IN THE VOICE ONSET TIME OF L2 ENGLISH STOP CONSONANTS

Eleanor Chodroff[1] and Melissa Baese-Berk[2]

[1]Department of Linguistics, Northwestern University
[2]Department of Linguistics, University of Oregon
eleanor.chodroff@northwestern.edu, mbaesebe@uoregon.edu

ABSTRACT

Non-native speech production is frequently characterized by its deviation from native pronunciation. Among segments, previous work has largely focused on describing the separation between native and non-native speakers at the level of individual phonetic categories. An additional hallmark of L1 pronunciation is the presence of systematic relationships within and among phonetic categories. For example, mean voice onset times (VOT) strongly covary among aspirated stop consonants across L1 speakers of American English. The present study examined whether L2 English speakers from various L1 backgrounds differ from native speakers in the relationship of VOT among word-initial /ptk/. Despite differences in the overall realization, L2 speakers resembled native English speakers in the degree of VOT covariation between stop-specific means and variances, as well as between /ptk/. These findings have important implications for the perception of accented speech, as listeners could employ structured relationships to facilitate adaptation despite non-native realizations of individual phonetic categories.

Keywords: voice onset time, L2 English, uniformity, speaker variation, corpus phonetics

1. INTRODUCTION

Conventional knowledge holds that L2 speech is more variable than L1 speech: contributing factors include unfamiliarity and uncertainty in the necessary gestures and phonetic targets for native-like realizations [4, 22, 24, 31]. Indeed, non-native speech production is frequently characterized by its deviation from native pronunciation [12, 27, 30]. Among segments, previous work has largely focused on describing the separation between native and non-native speakers at the level of individual phonetic categories as quantified by perceived production accuracy by native listeners [e.g., 10–11, 13] or by the phonetic degree of separation between native and non-native realizations [e.g., 1–2, 11, 14, 18].

More recent work has demonstrated that L2 speech is not always more variable than L1 speech. Variation can instead depend on factors such as the linguistic feature under analysis, as well as its instantiation and inherent variability in the speaker's L1 [29]. For example, the degree of variation in voice onset time (VOT) of voiced and voiceless consonants largely reflects the typical distribution of VOT in the native language: native Mandarin speakers produced Japanese voiceless stops with more variable VOT and Japanese voiced stops with less variable VOT relative to native Japanese speakers [29]. While Mandarin speakers primarily produced short-lag VOT values for Japanese voiced stops, Japanese speakers employed both lead and short-lag VOT.

These findings highlight the fact that variability in L2 speech is not random. Limited variation in part reflects properties of the L1 system, which may be inherently less variable than the L2 system. Moreover, to the extent that listeners adapt to L2 speech in perception, variability cannot be unconstrained within the phonetic system [1]. Mutual intelligibility between native and non-native speakers indicates that aspects of the L2 phonetic system must be preserved, even if individual segments have a non-native realization. These observations lead to the question as to how linguistic, structural factors in the phonetic system may constrain variability in L2 speech.

In the present study, we investigated constraints on variability in L2 English VOT in the realization of word-initial, prevocalic voiceless stop consonants. Previous research has identified strong, positive relationships between talker-specific mean VOTs and corresponding standard deviations, indicating increased variability with longer means [7, 28]. There are also well-known systematic relationships of VOT across place of articulation, in which VOT generally increases with more posterior places of articulation [6, 19]. Chodroff & Wilson [8] identified that the VOT relationship across place of articulation is much tighter than might be expected given a mere ordinal constraint requiring the VOT of /p/ to be less than the VOT of /k/: talker mean VOTs strongly covary among place of articulation, indicating mutual predictability between an individual's stop-specific VOT values. Systematic investigation of these phenomena, however, has largely been limited to L1 speech production.

The current study therefore examined whether L2 English speakers maintained systematic relationships in VOT that have been observed in the realization L1 English /ptk/. L1 English /ptk/ in word-initial position

is notably produced with aspiration, which manifests acoustically as long-lag VOT. The L2 speakers in the study come from varied L1 backgrounds: while all represented L1s have a voiceless stop series with all three places of articulation, several L1s are non-aspirating languages with short-lag VOT (e.g., French, Spanish). Most speakers produced long-lag VOTs, yet the variability in overall VOT was nonetheless greater across L2 speakers than across L1 English speakers. Critically, L2 speakers resembled native English speakers in the degree of VOT covariation between stop-specific means and variances, as well as among the voiceless stop categories.

2. METHODS

2.1. Corpus description

The present analysis employed connected speech in English from L1 and L2 English speakers. The data were obtained from the Archive of L1 and L2 Scripted and Spontaneous Transcripts and Recordings (ALLSSTAR) Corpus, which contains connected and spontaneous English speech samples from over 100 L1 and L2 speakers of English [3]. Though not analyzed here, the corpus also includes matched L1 recordings for each L2 English speaker.

The connected speech subset of the corpus contains five production tasks: two Hearing in Noise tasks (60 sentences each), a sample from the United Nations Declaration of Human Rights (20 sentences), a sample from the short story Le Petit Prince (30 sentences), and The North Wind and the Sun passage (4 sentences). These tasks were completed by 140 speakers from the Northwestern University community. There were 26 native, monolingual speakers of American English (14 female) and 102 L2 English speakers (36 female), representing 19 language backgrounds. These were Cantonese (14 speakers), Farsi (3), French (1), German (2), Gishu (1), Greek (1), Gujarati (1), Modern Hebrew (4), Hindi (5), Indonesian (1), Japanese (3), Korean (11), Mandarin (16), Brazilian Portuguese (5), Runyankore (1), Russian (5), Spanish (11), Turkish (13), and Vietnamese (4).

2.2. Stop segmentation and measurement

The script for each task was automatically aligned to the audio files using the FAVE-align system [23]. The boundaries of word-initial, prevocalic voiceless stop consonants were further refined with AutoVOT, which returns the optimal boundaries corresponding to the stop release and following vowel onset within a pre-specified window of analysis [15]. To set the window of analysis, the original boundaries from FAVE were extended an additional 30 ms in each direction, and the minimum permissible VOT was set to 10 ms. The word 'to' was omitted from analysis because of its tendency to undergo reduction.

An estimate of the automatic boundary error was obtained through manual measurement of 638 stop consonants, randomly selected (~5% of stops under analysis). Of the manually measured stops, 12 instances did not contain a stop release, and were therefore removed. The RMS error between the automatically and manually aligned VOTs was 16 ms (/p/: 27 ms, /t/: 20 ms, /k/: 14 ms). An additional 567 VOTs with values equal to the minimum permissible duration of 10 ms were hand-corrected. Of those, 59 instances were removed for lacking a stop release.

VOT was measured as the duration between the AutoVOT-defined boundaries or when available, the manually defined boundaries. For each speaker, values 2.5 standard deviations beyond the mean were considered potential measurement errors and excluded (145 values). A total of 12,624 stops was available for analysis. Per speaker, there was a maximum of 28 /p/s, 26 /t/s, and 47 /k/s, and the median number per speaker was within one to two tokens of the maximum.

3. RESULTS

The results section is organized as follows: we first present the summary statistics for talker VOT means and standard deviations (SD), followed by the analysis of VOT relationships. The VOT relations considered here were the covariation between talker mean and SD, the ordinal relationship of VOT means across place of articulation, and the covariation of talker mean VOTs among place of articulation.

The mean and range of talker VOT means for L1 and L2 speakers are presented in Table 1 and shown in Figure 1. L2 speakers produced a range of VOT means, spanning traditional short-lag to long-lag VOT values. Several, though not all, L2 speakers which lack aspiration in their L1 were nevertheless capable of producing VOT within the long-lag region. For example, two native Spanish speakers had average VOTs around 68 ms, even though Spanish voiceless stops are produced with short-lag VOT. Exactly 15 L2 speakers had average VOTs below 35 ms. (We used 35 ms as a rough boundary between short-lag and long-lag VOT as this value reflects a general auditory boundary in human and animal stop perception [16].) The native languages of those speakers were Greek (1 speaker), Hindi (4), Japanese (1), Brazilian Portuguese (3), Runyankore (1), Russian (1), Spanish (3), and Turkish (1). Apart from Hindi and to some extent Japanese, these languages are typically described as having unaspirated stop consonants. The segmental inventory of Hindi contains aspirated stop consonants [20]; however, four out of five Hindi speakers had average English VOTs less than 35 ms. Japanese stop consonants have

variably been described as unaspirated or moderately aspirated with intermediate VOT [21].

Table 1: Mean and range of speaker VOT means (ms) for L1 and L2 English speakers.

Group	/p/	/t/	/k/
L1	54 (34-65)	69 (46-93)	60 (41-79)
L2	42 (12-83)	56 (11-100)	59 (24-90)

While L2 speakers exhibit a much larger range of VOT means than L1 speakers, previous research has indicated structure in the relationship between the mean and SD in L1 English VOT [7]. Talker-specific means and SDs were moderately correlated within each stop category for L1 speakers, though the correlation for /t/ did not reach significance at a Bonferroni-corrected alpha of 0.008 (*r*s: /p/ 0.56, p = 0.003, /t/ 0.46, p = 0.04, /k/ 0.56 p = 0.003). Correlations were moderate to strong and significant for L2 speakers (*r*s: /p/ 0.75, /t/ 0.63, /k/ 0.72, *p*s < 0.008). Though the correlations are numerically stronger across L2 speakers, no significant difference was observed in the correlational magnitudes between L1 and L2 speakers for /p/ (p = 0.14), /t/ (p = 0.28), or /k/ (p = 0.22), as assessed using Fisher's r-to-z transformation. Overall, correlations between the mean and SD may be stronger among short-lag productions (or for lower VOTs more generally), which accords with the strong correlations found for English short-lag stops /bdg/ [7].

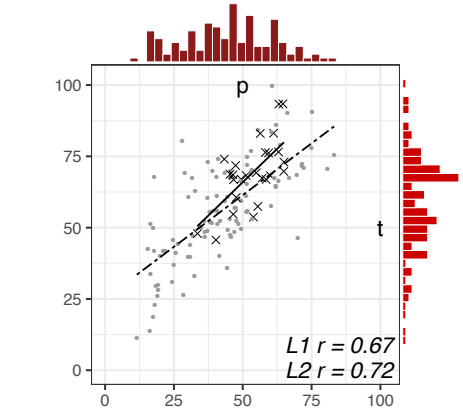
Table 2: Percent adherence to canonical ordinal relationships in VOT among L1 and L2 English speakers.

Group	/p/ < /t/	/t/ < /k/	/p/ < /k/
L1	96%	15%	85%
L2	87%	64%	93%

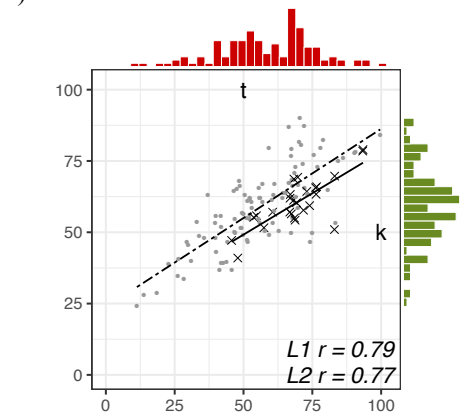
To examine the relationships of talker VOT means *across* place of articulation, we first assessed the degree to which speakers adhered to the expected ordinal relationship in which VOT increases with more posterior places of articulation (/p/ < /t/ < /k/; Table 2). Among L1 speakers, the VOT of /k/ was most often shorter than the VOT of /t/. Previous studies have reported deviation from the canonical ranking between /t/ and /k/ for several varieties of English (American: [8, 32]; British: [9, 26]). The ranking between /t/ and /k/ among L2 speakers was relatively more variable than between other stop pairs, but unlike native English speakers, most speakers preserved the canonical ranking of /t/ < /k/.

Though L1 English speakers had a longer VOT for /t/ than /k/, the ordinal rankings fail to reveal the extent to which /k/ exceeds /t/, and whether this difference is realized consistently across speakers. Moreover, L2 English speakers may nevertheless preserve relationships among /ptk/ that are similar to L1 speakers even if the ordinal rankings differ slightly. As shown in Figure 1, correlations of talker mean VOTs between place of articulation were quite strong (*r*s > 0.70; *p*s < 0.001), and did not significantly differ in magnitude between L1 and L2 speakers (*z*s: /p/-/t/ 0.39, p = 0.70; /t/-/k/ 0.87, p = 0.87; /p/-/k/ -0.29, p = 0.77).

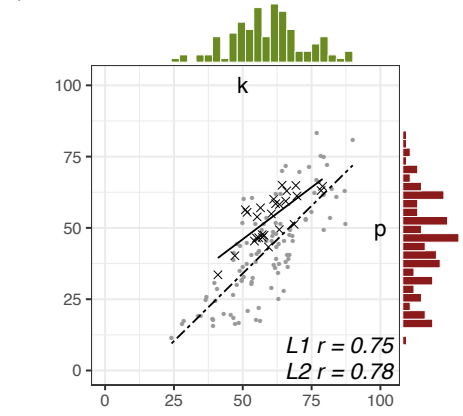
Figure 1: Correlations of talker VOT means between a) /p/ and /t/, b) /t/ and /k/, and c) /p/ and /k/ across L1 English speakers (×) and L2 English speakers (·) with the best fit linear regression lines for each group (L1 = solid, L2 = dashed). The marginal histograms reflect the range of talker VOT means for each stop category.



a)



b)



c)

Simple linear regressions were also fit between place-specific talker means for L1 and L2 speakers to further assess the relation between place-specific talker means (Table 3). Among L1 and L2 speakers, the regressions indicated that with longer VOT means, the VOT of the canonically lower stop approached the VOT of the higher one at increasing rates (slopes < 1). In some cases, and especially between /t/ and /k/, the VOT of the canonically lower stop often surpassed the canonically higher one.

Table 3: Simple linear regressions predicting talker-specific VOT means for one place of articulation (left of tilde) from a second place (right of tilde). Asterisks reflect $p < 0.001$.

Group	Parameter	/t/ ~ /p/	/k/ ~ /t/	/k/ ~ /p/
L1	Intercept	19.14	20.15	18.65
	Slope	0.94*	0.58*	0.78*
L2	Intercept	25.34*	27.02*	33.45*
	Slope	0.78*	0.56*	0.62*

4. DISCUSSION

L2 English stop VOT means encompassed a greater range than those found in L1 English. While several L2 speakers produced VOT values within a native-like long-lag range, several speakers produced short-lag VOTs. Highly comparable relationships among VOT parameters were observed between L1 and L2 English speakers. Talker means and SDs were weakly correlated across L1 speakers and moderately to strongly correlated across L2 speakers. The relationship between the mean and SD may simply be stronger among short-lag values, which is consistent with previous findings for English short-lag /bdg/ [7]. Moderate differences were observed between groups in the ordinal relations, particularly with respect to the ranking between /t/ and /k/, in which the VOT of L1 English /k/ was somewhat lower than expected given the canonical rank relationship. Nevertheless, strong covariation of talker mean VOT was observed among places of articulation with near identical magnitudes for both L1 and L2 speakers. The simple linear regressions revealed a general tendency for differences among place-specific means to decrease with longer VOT values.

Several studies on VOT examine how L2 speakers differ in the realization of VOT from L1 English speakers [1, 5, 17, 25]. Previous work, however, has largely focused on describing the separation between native and non-native speakers at the level of individual phonetic categories, with relatively minimal discussion with regards to the relation between those segments. Moreover, exploration of systematic relationships among L2 VOT has been quite limited. While a few studies have examined rank relationships in L2 VOT [17, 25], investigation of the degree of systematicity in VOT through correlation analysis is a novel aspect of the present study and offers important insight into the structure of the L2 grammar.

These findings have implications for the phonetics-phonology interface and structure of the L2 grammar, as well as for perceptual adaptation to accented speech. Systematic relationships of VOT among stop categories for both L1 and L2 speakers indicate that the phonetic specification for these differing segments cannot be independent of one another. Covariation could arise from a uniformity constraint on the laryngeal specifications giving rise to the measured VOT. A similar glottal spreading duration and temporal alignment to the oral constriction for each place of articulation would result in minor differences in VOT [6, 19], and simultaneously account for covariation due to underlying identity across talkers [8]. The uniformity constraint would target the phonetic implementation of the natural class by constraining the phonetic targets for the shared distinctive feature value or gestural constellation to be uniform regardless of co-occurring features or gestures within each segment. The findings presented here indicate that uniformity may also constrain phonetic implementation in the L2 grammar to some degree.

Assuming a uniformity constraint, it may appear that L1 English deviates from the cross-linguistic tendency for the VOT of /t/ to be less than that of /k/. Given the patterns observed for L2 English, English /k/ is indeed somewhat lower than expected, and /t/ may also be slightly higher than expected. English may marginally violate the uniformity constraint; however, the strength of the covariation indicates that the relationship between the laryngeal specifications must still be strongly constrained among these segments.

Structured relations in L2 speech could facilitate perceptual adaptation in that listeners could use information from one segment to refine adaptation to other related segments. Critically, listeners could use this structure even if the absolute phonetic realization deviates substantially from native norms. Further research is necessary to investigate perceptual generalization in accented speech.

5. ACKNOWLEDGEMENTS

The authors would like to thank Kayla Walker for her assistance in data processing, Ann Bradlow and Chun Liang Chan for access to the data, Jennifer Cole for her support, as well as two anonymous reviewers for their feedback.

6. REFERENCES

[1] Antoniou, M., Best, C. T., Tyler, M. D., & Kroos, C. (2011). Inter-language interference in VOT production by L2-dominant bilinguals: Asymmetries in phonetic

code-switching. *J. Phon.*, *39*(4), 558–570.

[2] Baker, W., & Trofimovich, P. (2005). Interaction of native- and second- language vowel system(s) in early and late bilinguals. *Lang. Speech*, *48*(1), 1–27.

[3] Bradlow, A. R., Ackerman, L., Burchfield, L. A., Hesterberg, L., Luque, J., & Mok, K. (2011). Global features of native and non-native speech. *Proc. 17th ICPhS*. Hong Kong, 356–359.

[4] Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.

[5] Chionidou, A., & Nicolaidis, K. (2015). Voice onset time in bilingual Greek-German children. *Proc. 18th ICPhS*. Glasgow, 0595.

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

[7] Chodroff, E., Godfrey, J. J., Khudanpur, S., & Wilson, C. (2015). Structured variability in acoustic realization: A corpus study of voice onset time in American English stops. *Proc. 18th ICPhS*. Glasgow, 0632.

[8] Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *J. Phon.*, *61*, 30–47.

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

[10] Flege, J. E. (1988). Factors affecting degree of perceived foreign accent in English sentences. *J. Acoust. Soc. Am.*, *84*(1), 70–79.

[11] Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *J. Phon.*, *25*(4), 437–470.

[12] Flege, J. E., & Hillenbrand, J. (1986). Differential use of temporal cues to the /s/-/z/ contrast by native and non-native speakers of English. *J. Acoust. Soc. Am.*, 79(2), 508–517.

[13] Flege, J. E., Takagi, N., & Mann, V. A. (1995). Japanese adults can learn to produce English /r/ and /l/ accurately. *Lang. Speech*, *38*(1), 25–55.

[14] Guion, S. G. (2003). The vowel systems of Quichua-Spanish bilinguals: Age of acquisition effects on the mutual influence of the first and second languages. *Phonetica*, *60*(2), 98–128.

[15] 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 August 2014 from https://github.com/mlml/autovot/.

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

[17] Lopez, V. G., & Counselman, D. (2013). L2 acquisition and category formation of Spanish voiceless stops by monolingual English novice learners. In J. C. Amaro (Ed.), *Proc. 16th Hisp. Ling. Symp.*(pp. 118–127). Somerville, MA: Cascadilla Proceedings Project.

[18] MacLeod, A. A. N., Stoel-Gammon, C., & Wassink, A. B. (2009). Production of high vowels in Canadian English and Canadian French: A comparison of early bilingual and monolingual speakers. *J. Phon.*, *37*(4), 374–387.

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

[20] Ohala, M. 1983. Aspects of Hindi Phonology. Motilal Banarsidass. Delhi.

[21] Riney, T. J., Takagi, N., Ota, K., & Uchida, Y. (2007). The intermediate degree of VOT in Japanese initial voiceless stops. *J. Phon.*, *35*(3), 439–443

[22] Romero-Rivas, C., Martin, C. D., & Costa, A. (2015). Processing changes when listening to foreign-accented speech. *Front. Hum. Neuro.*, *9*, 1–15.

[23] Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., & Yuan, J. (2014). FAVE Program Suite v1.2.2.

[24] Schmale R., Hollich G., & Seidl A. (2011): Contending with foreign accent variability in early lexical acquisition. *J. Child Lang., 38*(5), 1096–1108.

[25] Simon, E. (2010). Acquiring a Second Language Laryngeal System. Gent: Academia Press.

[26] Suomi, K. (1980). *Voicing in English and Finnish Stops: A Typological Comparison with an Interlanguage Study of the Two Languages in Contact*. PhD Dissertation. University of Turku.

[27] Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Stud. Sec. Lang. Acq.*, *28*(1), 1–30.

[28] Turk, A. E., & Shattuck-Hufnagel, S. (2014). Timing in talking: What is it used for, and how is it controlled? *Phil. Trans. Royal Soc. London. Series B, Bio. Sci.*, *369*, 20130395.

[29] Vaughn, C., Baese-Berk, M., & Idemaru, K. (2018). Re-examining phonetic variability in native and non-native speech. *Phonetica*, 1–32.

[30] Wade, T., Jongman, A., & Sereno, J. (2007). Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. *Phonetica*, 64(2–3), 122–144.

[31] Witteman, M. J., Weber, A., & McQueen, J. M. (2014). Tolerance for inconsistency in foreign-accented speech. *Psych. Bull. Rev.*, *21*(2), 512–519.

[32] Yao, Y. (2009). Understanding VOT variation in spontaneous speech. In M. Pak (Ed.), *Current Numbers in Unity and Diversity of Languages* (pp. 1122–1137). Seoul: Linguistic Society of Korea.