TRAJECTORIES OF VOICE ONSET TIME IN SPONTANEOUS SPEECH ON REALITY TV

Morgan Sonderegger

Department of Linguistics, McGill University
morgan.sonderegger@mcgill.ca

ABSTRACT

Do speakers' accents change from day to day? This paper examines this question through the lens of voice onset time (VOT). We examine whether VOT within individual speakers shows time dependence—daily fluctuations, longer-term time trends, or both—by examining spontaneous speech from 11 English speakers on three months of the reality TV show *Big Brother UK*. We build statistical models of time dependence in VOT for each speaker, controlling for a range of other factors, and find that all speakers show daily fluctuations in VOT, for both voiced and voiceless stops, while longer-term time trends (weeks–months) are present in about half of cases. Together with previous work, these results suggest that pronunciation (at least VOT) does change from day to day, possibly due to accumulated accommodation effects, but that these shifts often do not accumulate into longer-term change.

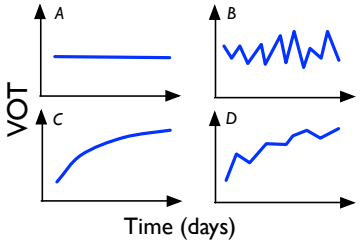
Keywords: Longitudinal variation, spontaneous speech, phonetic corpora, voice onset time

1. INTRODUCTION

How a speaker realizes sounds in spontaneous speech varies greatly as a function of many factors, such as properties of the context (e.g. coarticulation), the utterance (e.g. speaking rate), and the social setting. Less is known about variability over *time* within individuals, the subject of this paper.

Shifts in phonetic variables (such as VOT or vowel formants) over time have primarily been examined in previous work on two timescales. Phonetic accommodation/imitation studies have shown that *short-term* shifts (minutes–hours) in phonetic parameters under exposure to the speech of others, in laboratory or conversational settings [3, 14, 21], are ubiquitous: speakers tend to robustly show some shift in the parameter under study. In *long-term* studies in phonetic and sociolinguistics, phonetic variables are measured in speech from the same individual(s) at a few time points, years apart [13, 20]. While some individuals show remarkable flexibility over the lifespan [9], as in a study of VOT in a Portuguese-English bilingual [18], the general finding is of significant heterogeneity among individuals and variables in whether (and to what extent) change occurs [19], often with a minority showing *any* change. The primary theoretical issue addressed by this paper is: how can the ubiquity of short-term accent change in individuals be reconciled with the heterogeneity of long-term accent change?

Figure 1: Schematic of possible types of time dependence of VOT within indvidual speakers: none (A), by-day variability (B), time trend (C), BDV & time trend (D).



We address this issue by examining the timescale in between, asking how a phonetic variable (VOT) varies in individuals from day to day, over a "medium-term" timescale (months). Almost nothing is known about variability over this timescale (c.f. [10, 15]), as noted by [13]. Fig. 1 shows schematics of four possible kinds of dynamics of a variable on this timescale:

- A No change over time, a null hypothesis analogous to the "apparent-time hypothesis" of little change over the post-adolescent lifespan [4].
- B Variability from day to day (*by-day variability*; BDV), but with no systematic change in mean value over time.
- C Systematic change in mean value over time (*time trend*), with no (further) BDV.
- D Both by-day variability and time trend.

A phonetic variable could show any of these patterns within a given individual. The main empirical question addressed in this paper is, what kind of time dependence does VOT show in individual speakers, for voiced and voiceless stops?

We address this question using a corpus of speech from a reality television show, where contestants live

in an isolated house for (up to) 3 months and are constantly recorded, allowing us to examine an individual's speech on different days. Because contestants constantly interact, accommodation effects might be expected to build up over time. We examine the dynamics of VOT in a large dataset ($\sim$21k tokens), allowing us to detect small fluctuations, and find that change over time is widespread. A second question asked is whether this kind of change is importan, which we address by comparing the magnitude of change observed to that of contextual effects on VOT. Finally, we suggest that the observed dynamics of VOT suggest an answer to the theoretical issue raised above.

2. DATA AND METHODS

2.1. Data

The data come from a corpus of speech from 22 contestants on the 2008 season of *Big Brother UK*, which lasted 93 days [22, 23]. We consider speech from the 11 contestants (henceforth *speakers*) who were on the show for at least 50 days, and seem to be native English speakers: four female and seven male (labeled F1–F4, M1–M7). Speech comes from 588 segments (or *clips*) of several minutes, where speakers were in the "diary room" speaking to Big Brother (remotely, without seeing him or her); the clips thus have a constant recording environment and consist of similar types of speech. Each clip corresponds to a day of the season; on any given day there are sometimes several or no clips for a given speaker (average: 1 clip/1.39 days). One contestant per week is evicted from the show; speakers thus vary widely in the amount of data they contribute (811–3313 tokens/speaker, 32–80 clips/speaker). Most speakers speak British dialects; M2 is from the US, F4 is from Australia, and M5 has accented but fully grammatical speech. These differences are unimportant for the questions addressed in this paper.

The dataset consists of the 20822 word-initial stops in the corpus where a burst was present (voiced: 10656 tokens, 709 types; voiceless: 10166 tokens, 893 types). VOT for these stops was measured semi-automatically: following forced alignment of transcription and signal using FAVE [17], automatic VOT measurements were made using AutoVOT [11, 24], then manually corrected by three research assistants. Deciding how to define "voice onset time" in spontaneous speech is complex; we used similar criteria to [25]. In particular, no attempt was made to account for prevoicing or voicing during closure; thus, all our VOT measurements are positive, and for some (phonologically) voiced stops might be better characterized as "burst +aspiration duration".

VOT in this dataset is modeled as a function of a range of variables (in SMALL CAPS). Of primary interest are CLIP and the DAY a clip is from, which characterize time dependence. The models also control for a range of non-time variables (which we call *static factors*) that greatly influence VOT [2, 7, 12], especially in highly-variable spontaneous speech [22, 25, 27] (expected effect on VOT listed for each): stop PLACE OF ARTICULATION (labial<alveolar<velar), following PHONE TYPE (V<C), following VOWEL HEIGHT (non-high<high) FREQUENCY (in corpus, log-transformed: high<low), syllable STRESS (N<Y), position in phrase[1] (non-initial<initial) SPEAKING RATE (sylls/sec in phrase: high<low).

2.2. Models

We modeled time dependence of VOT within each speaker using a two-step process.[2]

First, two linear mixed-effects models were fit (using the `lme4` package in `R`; [6, 16]) of log(VOT) for voiced and voiceless stops, as a function of static factors only, across data from all speakers. These models included fixed effects for all static factors listed above, including interactions based on exploratory data analysis, and all possible by-word and by-speaker random intercepts and slopes. The *residuals* of these models thus capture VOT variability after accounting for static factors.

We then modeled time dependence in these residuals within individual speakers, for each type of stops (voiced, voiceless), using generalized additive mixed models (GAMMs), built using the `mgcv` package in `R` [26]. GAMMs allow for incorporation of two types of terms which conceptually correspond to the two types of time dependence which differentiate possibilities A–D (Fig. 1): (1) a random intercept of CLIP, which captures *by-day variability*; (2) an arbitrary smooth function of DAY, conceptually similar to a nonlinear smoother, which captures any *time trend*. We built four GAMM models for each speaker/voicing subset, one for each combination of presence/absence of terms (1) and (2); the best model was then selected using AIC.

3. RESULTS

This process resulted in 22 models of time dependence (11 speakers $\times$ {voiced, voiceless}) of VOT within individual speakers for voiced and voiceless stops, summarized in Table 1. All cases show some by-day variability, and are thus of Type B or D, de-

pending on whether mean VOT changes over time. (For example, speaker M1 shows BDV but no time trend in voiced stops.) We discuss BDV and time trend results in turn.

3.1. By-day variability

All cases show some by-day variability. The amount of BDV in each case is reported in two ways, labeled *%increase* and *range* in Table 1. Each model predicts the magnitude of fluctuations of VOT around the mean as a parameter σ, such that 95% of days have VOT within $\pm 2\sigma$ of the mean. Since log(VOT) is being modeled, this means that the *ratio* of the VOT on a "high day" ($+2\sigma$) and a "low day" (-2σ) is $e^{4\sigma}$; this, converted into percent increase, is the quantity reported as *% increase*. (Ex: voiced stops for speaker F1 are 43% larger on high days than on low days.) Because VOT is commonly reported in msec, we also give the predicted magnitude of fluctuations, in msec, between high and low days, evaluated at the speaker's mean VOT value. (For example, the difference between the mean VOT on $+2\sigma$ and -2σ days for voiced stops for F1 is 8 msec.) This quantity, reported as *range*, is analogous to the "range" measure of effect size for factors. We generally use range below, for easier interpretability.[3]

For different speakers, by-day fluctuations in VOT are between 7–17 msec for voiced stops, and 6–26 msec for voiceless stops. We can get a sense of how important these fluctuations are by comparison to the effect size of PLACE OF ARTICULATION, the contextual variable which has the largest effect on VOT in our models. The difference in VOT among places of articulation (i.e. "range") is 9 msec for voiced stops and 27 msec for voiceless stops. Thus, the by-day fluctuations in VOT are of similar magnitude to contextual effects.

The size of the fluctuations can also be compared with the magnitude of variability over time observed in three short-term studies of shifts in VOT in production, when subjects were exposed to voiceless stops with lengthened VOT: subjects increased VOT (of voiceless stops) by 12 msec on average in a shadowing task [21]; different subjects increased VOT by about 0–30 msec in one imitation task [14] and changed VOT (up or down) by about 0–17 msec in another imitation task [28]. All these shifts are comparable with the magnitude of fluctuations of VOT in voiceless stops in our data.

3.2. Time trends

Half of cases (voiced: 4/11 speakers; voiceless: 7/11 speakers) show change in mean VOT over time.

Figure 2: Time trajectories of mean VOT over time for individual speakers for voiced and voiceless stops, based on model selected by AIC. (Flat lines indicate a model had no time trend.)

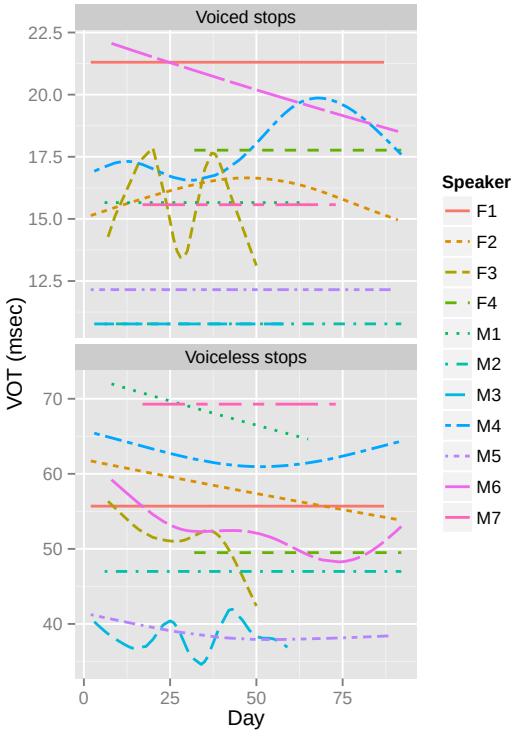


Fig. 2 shows the model-predicted time trajectories of mean VOT over time for each speaker. The shape and magnitude of change in mean VOT over time differ greatly between speakers, but a few general observations are possible. First, change over time is never large enough to result in overlap of VOT between voiced and voiceless stops (note y-axis ranges in Fig. 2), in line with the status of VOT as the primary cue to the voicing contrast for English stops. Second, there is no clear pattern of overall convergence in VOT, a point we return to below. Finally, the amount of change in mean VOT (the vertical range spanned by each curve) is smaller than the effect size of place of articulation (discussed above), though still of a similar order of magnitude.

4. DISCUSSION

We discuss how our findings bear on the primary question of interest, how VOT varies on a timescale of days-months, as well as possible theoretical implications.

First, we can definitively reject the null hypothesis of no change (Type A in Fig. 1): *all speakers showed* some *change in VOT, for both voiced and voiceless stops*. Our other main finding is that *by-day variability in VOT is ubiquitous*. In all cases, a speaker's

Table 1: Summary of time dependence of VOT within individual speakers for voiced and voiceless stops, based on model selected by AIC. The magnitude of by-day variability is described by the difference predicted in VOT between days -2 SD and +2 SD from the mean, reported as (1) *% increase*, and (2) predicted difference in msec evaluated at the speaker's overall mean (*range*). Time trend = Y/N corresponds to the presence/absence of change in mean VOT over time.

		F1	F2	F3	F4	M1	M2	M3	M4	M5	M6	M7
Voiced stops												
By-day	*% increase*	43%	64%	84%	156%	96%	142%	68%	79%	180%	42%	66%
variability	*range*	8 ms	8 ms	10 ms	17 ms	11 ms	10 ms	6 ms	11 ms	13 ms	7 ms	8 ms
Time trend?		N	Y	Y	N	N	N	N	Y	N	Y	N
Voiceless stops												
By-day	*% increase*	13%	48%	22%	67%	46%	42%	18%	42%	23%	45%	25%
variability	*range*	7 ms	23 ms	10 ms	26 ms	26 ms	17 ms	6 ms	22 ms	8 ms	19 ms	15 ms
Time trend?		N	Y	Y	N	Y	N	Y	Y	Y	Y	N

use of VOT fluctuates from day to day, after controlling for a host of covariates (such as speaking rate and place of articulation) that affect VOT. This pattern is in line with the robustness across individuals of shifts in phonetic parameters in conversation and laboratory experiments (timescale of minutes–hours), suggesting that phonetic parameters may be flexible on a timescale of up to days.

We cannot say much about the source of the fluctuations in VOT; though we controlled for a range of covariates, they could be due to factors not in the models, such as differences in speaking style (beyond rate). However, the fact that the observed fluctuations are of similar magnitude to those observed for VOT in imitation and shadowing studies invites the speculation that the fluctuations are due to accommodation during conversation: speakers "bounce around" from speaking with each other, and thus have different baseline VOT values when recorded in a constant environment (the diary room) on different days. Regardless of their source, the existence of robust and sizable fluctuations in phonetic parameters has important implications for studies of phonetic change over the lifespan, which generally only measure speakers on a *single* day at each time point years apart; they could find a spurious difference between two time points or fail to find a difference (when there is one), due to "noise" from by-day variability. The current study establishes this as a worry for VOT; future work should examine other phonetic parameters as well.

On the other hand, time trends in VOT over weeks–months are much more sporadic: in 50% of cases, VOT did not (measurably) show any time trend.[4] This is in line with the huge variability in whether and how much phonetic parameters change in long term studies, over years. It is also striking, given the extreme proximity and constant interaction of contestants on *Big Brother*, that there is no clear overall pattern (i.e. of convergence or divergence) in the time trends that are found (Fig. 2), as might be expected under theories in which sound change is driven by unconscious and automatic phonetic imitation [8]. Depending on *why* speakers show different time trends (or none), our findings could support theories of sound change in which socially-mediated accommodation plays a role [1]. Future work should examine whether speakers' different dynamics can be grounded in social interaction.

More work is needed to check if other phonetic parameters show the type of time dependence within individuals observed here.[5] However, assuming the current results are to some extent representative of plasticity in speech in general, the way that different parts of our results line up with the short-term and long-term change literatures suggests a speculative answer to the question posed above, of how the ubiquity of short-term accent change can be reconciled with the heterogeneity of long-term accent change: short-term shifts persist on timescales of days, and hence could accumulate into longer-term change. However, these by-day fluctuations often do not accumulate into longer-term trends, perhaps in part because individuals differ significantly in how long accommodation effects persist—as can already be observed here, on a timescale of months. Long-term accent change is then (correctly) predicted to be very heterogeneous.

Acknowledgements This work was partially funded by SSHRC #430-2014-00018 and FRQSC #183356. We thank M. Bane and P. Graff for collaboration in an earlier stage of this research, as well as L. Bassford, T. Knowles, M. Labelle, and M. Schwartz for research assistance.

5. REFERENCES

[1] Auer, P., Hinskens, F. 2005. The role of interpersonal accommodation in a theory of language change. In: Auer, P., Hinskens, F., Kerswill, P., (eds), *Dialect change: convergence and divergence in European languages*. Cambridge: Cambridge University Press 335–357.

[2] Auzou, P., Ozsancak, C., Morris, R., Jan, M., Eustache, F., Hannequin, D. 2000. Voice onset time in aphasia, apraxia of speech and dysarthria: a review. *Clin. Linguist. Phonet.* 14(2), 131–150.

[3] Babel, M. 2011. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *J. Phon* 40(1), 177–189.

[4] Bailey, G. 2002. Real and apparent time. In: Chambers, J., Trudgill, P., Schilling-Estes, N., (eds), *The Handbook of Language Variation and Change*. Malden, MA: Blackwell 312–331.

[5] Bane, M., Graff, P., Sonderegger, M. 2010. Longitudinal phonetic variation in a closed system. Baker, A., Baglini, R., Grinsell, T., Keane, J., Thomas, J., (eds), *Proc. CLS 46* 43–58.

[6] Bates, D., Maechler, M., Bolker, B., Walker, S. 2014. *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7.

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

[8] Delvaux, V., Soquet, A. 2007. The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica* 64(2), 145–173.

[9] Harrington, J., Palethorpe, S., Watson, C. 2000. Does the Queen speak the Queen's English? *Nature* 408(6815), 927–928.

[10] Heald, S. 2012. *Tuning the Targets of Speech Production: Effects of Articulatory and Perceptual Experience on Speech Production*. PhD thesis University of Chicago.

[11] 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. https://github.com/mlml/autovot/.

[12] Kessinger, R. H., Blumstein, S. E. 1997. Effects of speaking rate on voice-onset time in Thai, French, and English. *J. Phon* 25(2), 143–168.

[13] Nahkola, K., Saanilahti, M. 2004. Mapping language changes in real time: A panel study on Finnish. *Lang. Var. Change* 16(2), 75–92.

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

[15] Pisoni, D. 1980. Variability of vowel formant frequencies and the quantal theory of speech: a first report. *Phonetica* 37(5-6), 285–305.

[16] R Core Team, 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

[17] Rosenfelder, I., Fruehwald, J., Evanini, K., Yuan, J. 2011. FAVE (Forced Alignment and Vowel Extraction) Program Suite [Computer program].

[18] Sancier, M., Fowler, C. 1997. Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *J. Phon* 25, 421–436.

[19] Sankoff, G. 2013. Longitudinal studies. In: Bayley, R., Cameron, R., Lucas, C., (eds), *The Oxford Handbook of Sociolinguistics*. Oxford: Oxford University Press.

[20] Sankoff, G., Blondeau, H. 2007. Language change across the lifespan: /r/ in Montreal French. *Language* 83(3), 560–588.

[21] Shockley, K., Sabadini, L., Fowler, C. A. 2004. Imitation in shadowing words. *Percept. Psychophys.* 66(3), 422–429.

[22] Sonderegger, M. 2012. *Phonetic and phonological dynamics on reality television*. PhD thesis University of Chicago.

[23] Sonderegger, M. 2014. Trajectories of phonetic variability in spontaneous speech on reality TV. Poster presented at LabPhon 14 (Tachikawa, Japan).

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

[25] Stuart-Smith, J., Sonderegger, M., Rathcke, T., Macdonald, R. in press. The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology*.

[26] Wood, S. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* 99(467), 673–686.

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

[28] Yu, A., Abrego-Collier, C., Sonderegger, M. 2013. Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and "autistic" traits. *PLOS ONE* 8(9), e74746.

[1] Defined as speech between force-aligned pauses of at least 100 msec.

[2] VOT was first log-transformed, because its distribution is right-skewed and non-negative.

[3] However, it is important to bear in mind that the models do not actually predict VOT differences in msec, but in % increase.

[4] This supercedes the conclusion of an extremely preliminary version of this work (806 voiceless tokens; [5]), which found time trends for 4/4 speakers, but failed to consider the possibility of by-day variability.

[5] Vowel formants seem to show qualitatively similar time dependence, in the same corpus [22, 23].