Structured speaker variability in Japanese stops: relationships within and across cues to stop voicing*

James Tanner[†1], Morgan Sonderegger[1], and Jane Stuart-Smith[2]

[1]Department of Linguistics, McGill University, Canada
[2]Glasgow University Laboratory of Phonetics (GULP), University of Glasgow, UK

November 18, 2019

Abstract

A number of recent studies have observed that phonetic variability is constrained across speakers, where speakers exhibit limited variation in the signalling of contrasts in spite of overall speaker differences. This previous work has focused predominantly on controlled laboratory speech and exclusively on contrasts in English and German, leaving unclear how such speaker variability is structured in spotaneous speech and in linguistic contrasts which make use of more than one acoustic cue. This study attempts to both address these empirical gaps and expand the empirical scope of research investigating structured variability by examining how speakers vary in the use of positive voice onset time and closure voicing in marking the stop voicing contrast in Japanese spontaneous speech. Strong covarying relationships within each cue across speakers are observed, whilst such relationships in the combined use of both cues are substantially weaker, suggesting that structured variability is constrained by the language-specific phonetic implementation of linguistic contrasts.

1 Introduction

The acoustic realisation of segments can vary substantially across languages, phonological contexts, and speakers. Even within a single language, the realisation of a particular segment can differ as a function of phonological context (Cho and Ladefoged, 1999), speech rate (Allen et al., 2003), and of a range of other linguistic and social factors (e.g., Foulkes et al., 2001). For individuals, speakers may differ in the realisation of speech sounds as a result of a number of different properties: some speakers are more prone to hyperarticulation of segments (Lindblom, 1990; Johnson et al., 1993), differ in their anatomical characteristics (Peterson and Barney, 1952), or simply arrive at different acoustic targets as function of some probabilistic approximation of the speech sounds in their community (Bybee, 2001; Pierrehumbert, 2001). This kind of speaker-level variability poses a potential challenge for the perception of speech (Kleinschmidt, 2018), where the mapping from values in a multi-dimensional acoustic space to abstract phonetic categories (e.g., [+voice], [-high], etc.) is differently realised for individual speakers (Liberman et al., 1967; Lisker, 1986). How, then, do speakers successfully convey the presence of singular linguistic categories in spite of individual variation in those categories'

*This paper is an extended version of a preliminary report in Tanner et al. (2019). We would like to thank Meghan Clayards, Eleanor Chodroff, James Kirby, and the audience of the 19th International Congress of Phonetic Sciences (ICPhS) for their feedback on this research. The research reported here was supported by the Social Sciences and Humanites Research Council of Canada (#435-2017-0925).

[†]Corresponding author: james.tanner@mail.mcgill.ca

realisations? One way in which this individual variability may be constrained is by the existence of underlying *structure* in the realisation of speech sounds across speakers: namely, that speakers' individual productions are related in a way that is fundamentally non-random. For example, whilst speakers vary in the realisation of a single acoustic parameter such as Voice Onset Time (VOT) for stops, the differences between individual speakers' VOT values for different places of articulation are highly correlated (Chodroff and Wilson, 2017; Hullebus et al., 2018). Speakers may also show similar kinds of structured variation across *multiple* cues to the production of a speech sound, evidenced by observed covariation in VOT and F0 across voiced and voiceless stops (Bang, 2017; Chodroff and Wilson, 2018; Schultz et al., 2012; Clayards, 2018).

With the exception of one study (Sonderegger et al., 2020),[1] the majority of recent research on structured variation across individuals has focused on production in controlled laboratory speech, either as isolated words or reading sentences (Chodroff and Wilson, 2017; Hullebus et al., 2018; Schultz et al., 2012; Clayards, 2018). The realisation of stops and stop contrasts are well-established to be enhanced in laboratory speech (Lisker and Abramson, 1967) relative to conversational speech (Baran et al., 1977), and so it is less clear how variability is structured in less-controlled speech. Additionally, our understanding of structured speaker variability is derived from research which has exclusively examined languages such as English and German: both of these languages primarily use the length of the stop burst and aspiration (henceforth 'positive VOT') to signal a range of contrasts in word-initial stops (e.g., Lisker and Abramson, 1964, 1967). How speakers vary in languages where the stop contrasts involve the use of additional phonetic cues is not well-understood.

This study addresses both of these gaps by focusing on the acoustic realisation of stops in spontaneous Japanese. Japanese makes use of *both* positive VOT and the presence of voicing in the stop closure for marking the contrast between voiced and voiceless stops (Shimizu, 1996; Tsujimura, 2014, Section 2.1). Additionally, the Japanese stop voicing contrast has been observed to be undergoing change (Takada, 2011; Takada et al., 2015), and so may provide some insight into how speakers vary in the use of both VOT and the degree of voicing during the stop closure, as well as in how both parameters are used to define the voicing contrast. Thus, this study expands the search for structured speaker variability by examining the evidence for three kinds of such structure across speakers of spontaneous Japanese: (1) *within* a phonetic cue across different segments (e.g., VOT between voiced and voiceless stops); (2) the size of the voicing contrast *across* cues (i.e., the relative difference in voiced and voiceless stops); and (3) *across* phonetic cues across and within segments (i.e., the relationship between VOT and closure voicing in voiced and voiceless stops).

2 Background

2.1 Acoustic cues to stops & stop voicing

Voice Onset Time (VOT), referring to the time between the release of the stop and onset of glottal pulsing, has been well-established as the primary acoustic cue for the stop voicing contrast in a range of languages, where voiced stops have shorter average VOT than their voiceless counterparts (Liberman et al., 1958; Lisker and Abramson, 1964; Abramson and Whalen, 2017). Japanese maintains a two-way stop voicing contrast, distinguishing between 'voiced' {/b/, /d/, /g/} and 'voiceless' {/p/, /t/, /k/} categories: acoustically, Japanese voiced stops may be realised either with prevoicing

[1] Chodroff and Wilson (2017) also report a preliminary analysis of VOT covariation across speakers in the Buckeye corpus of spontaneous speech (Pitt et al., 2007).

(negative) or short-lag (positive) VOT (Shimizu, 1996; Nasukawa, 2005; Gao and Arai, 2019), and voiceless stops are realised with a VOT intermediate between short ('unaspirated', Tsujimura, 2014) and long-lag ('moderately aspirated', Shimizu, 1996; Riney et al., 2007). Whilst less is known about variability in Japanese stop production, a large body of work has focused on how stops are modulated in English: here it is assumed that these factors are to some extent language-independent and are thus relevant for examining stops in Japanese. With respect to VOT, stops are affected by a range of linguistic factors, such as place of articulation (Lisker and Abramson, 1964; Docherty, 1992), preceding phoneme manner (Docherty, 1992; Yao, 2009), vowel height (Klatt, 1975), phrasal position (Lisker and Abramson, 1964; Cho and Ladefoged, 1999; Yao, 2009; Kim et al., 2018), and speech rate (Allen et al., 2003). Most work on variation in English VOT has used controlled speech, though a number of studies have looked at variation in English spontanous speech and have confirmed the robust difference in VOT between voiced and voiceless stops (Baran et al., 1977; Yao, 2009; Sonderegger et al., 2014; Stuart-Smith et al., 2015; Sonderegger et al., 2020).

The degree of vocal fold vibration during the closure (Lisker 1986; Voicing During Closure, henceforth VDC) has been substantially less studied than English VOT, though it has been shown that voiced stops are more likely to contain VDC than their voiceless counterparts (Docherty, 1992; Sonderegger et al., 2020). Thus far, much of the research on VDC has focused on English read speech (e.g., Davidson, 2016, 2018; Kim et al., 2018). For both voiced and voiceless stops, VDC is more likely in phrase- or word-medial contexts (Docherty, 1992; Lisker and Abramson, 1964, 1967). VDC in phrase-intitial stops, sometimes referred to as 'negative VOT', has been observed for English (Lisker and Abramson, 1964, 1967; Hunnicutt and Morris, 2016) and other languages (Abramson and Whalen, 2017). Additionally, VDC is also more likely when the preceding segment is voiced (Docherty, 1992; Davidson, 2016, 2018), which has also been observed for spontaneous Glaswegian English (Sonderegger et al., 2020). With the exception of geminated consonants, all syllables in Japanese are either open (ending in a vowel) or have a nasal coda (Tsujimura, 2014): all segments preceding stops in these cases are underlyingly voiced, then, and so should affect the likelihood of a stop being realised with VDC. VDC is also used as a contrastive cue for voicing in Japanese (discussed as 'negative' VOT or 'prevoicing'), though recent studies have shown that the prevoiced variant of the voiced stop has become less common in phrase-initial position (Gao and Arai, 2019), and may represent a sound change towards the exclusive use of positive VOT coupled with a contrastive function for F0 (Takada, 2011; Kong et al., 2014; Takada et al., 2015; Gao et al., 2019; Gao and Arai, 2019).

2.2 Individual speaker variability in stops

Differences between individual speakers been noted since the earliest studies on stop acoustics (e.g., Lisker and Abramson, 1964). As opposed to being purely random variation, these differences between speakers are highly structured: speaker differences in VOT are consistent after controlling for other linguistic factors, such as speech rate (Allen et al., 2003; Theodore et al., 2009). Speaker mean VOTs for different places of articulation in voiceless stops have been shown to be highly correlated in both English (Chodroff and Wilson, 2017) and German (Hullebus et al., 2018): despite of overall differences in a given speaker's mean VOT, realisation of the contrasts between voiceless stops (i.e., /p/ $\sim$ /t/, /p/ $\sim$ /k/, /t/ $\sim$ /k/) exhibit strong linear relationships. With respect to speaker variability across *multiple cues*, Chodroff and Wilson (2018) show that American English speakers covary in use of three cues (VOT, F0, and spectral centre of gravity), and Glaswegian English speakers covary in the

relationship between postive VOT in closure voicing (Sonderegger et al., 2020). Bang (2017, Ch. 3) and Schultz et al. (2012) also show strong relationship between VOT and F0 in marking the voicing contrast in Korean and L2 Korean-English speakers respectively, whilst Clayards (2018) observed speaker differences in the correlated use of VOT, F0, and following vowel duration in English.

As a means of characterising the sources of structured variability within the phonological grammar of an individual, Chodroff and Wilson (2017, 2018, 2020) propose a 'principle of uniformity'. Uniformity in this sense seems to refer to a linear relationship in the acoustic production of two segments across speakers; this structure constrains the degree of variation in the difference between two speech sounds across speakers, and that the realisation of one such sound has a predctive relationship with the other. Whilst speakers may vary in their overall use of a given phonetic cue (i.e., where that speaker is situated on this line), the relative difference in between two segments with respect to that parameter is consistent across speakers.[2] Much of the empirical evidence for Chodroff & Wilson's proposition of uniformity is derived from studies on English which uses an aspiration-led phonetic implementation of stops, and predominantly examined in controlled laboratory speech.

By examining the structure of speaker variability in spontaneous Japanese, a new speech context and a new language with a different phonetic implementation of voicing, it is possible to consider further possible evidence for phonetic uniformity in a new empirical setting. This examination takes a range of forms in this study: the first is to consider how speakers modulate the stop voicing contrast within a given phonetic cue (VOT and closure voicing). The second concerns how these two phonetic cues are manipulated together in signalling this contrast. Whilst some research has examined speaker variability across multiple cues (e.g., Clayards, 2018; Chodroff and Wilson, 2018; Sonderegger et al., 2020), the predictions are less clear for a language like Japanese where the cues to stop voicing differ from English. These questions also address the extent to which phonetic uniformity across speakers might be constrained and whether such constraints may be related to language-specific properties.

3 Methods

3.1 Data

The data used in this study comes from the Core subset of the Corpus of Spontaneous Japanese (CSJ, Maekawa et al., 2000), constituting approximately 45 hours of speech recorded 1999-2001 from 137 speakers (58 female), born between 1930 and 1979. Within the CSJ, speaker birth years are grouped into increments of 5 years (e.g., 1930-34, 1935-39, 1940-44, etc); as this resulted in too few speakers per group to reliably control for age effects in this study, speakers were allocated into groups of 10 years (1930-39, 1940-49, etc). The variety of Japanese used in the CSJ is 'Common' Japanese: a standard variety that derives many of its linguistic features from the Tokyo dialect (Maekawa et al., 2000). Each recording is approximately 30 minutes in length, and is predominantly of academic interviews and informal public speaking, though a subset (approximately 5%) constitutes conversational dialogue and reading passages. This core subset contains extensive phonetic and prosodic annotation, such as hand-corrected segmental boundaries, presence of vowel devoicing, and voice quality (Kikuchi and Maekawa, 2003). As illustrated in Figure 1, stops are annotated as the beginning of the stop release until the onset of voicing of the following vowel (i.e. *positive* VOT), the stop closure, and

[2] If we assume that the linear relationship can be across the cue values (X $\sim$ Y) or the log-transofrmed cue values (log(X) $\sim$ log(Y)) (Chodroff and Wilson, 2020), then "relative difference in X and Y consistent across speakers" is equivalent to "X and Y are linearly related: $y = aX + b$".

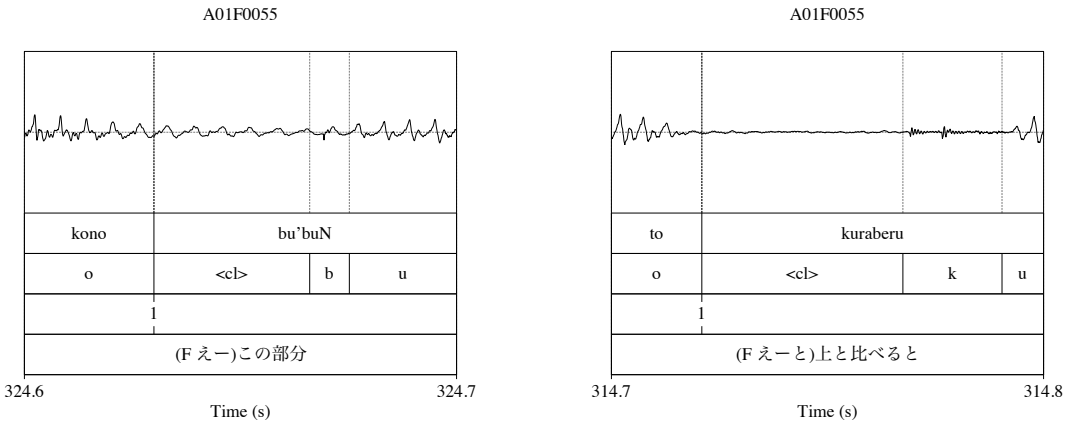

Figure 1: Waveforms and accompanying annotations for phrase-internal stops realised with and without closure voicing ('kon**o** **b**ubun', left; '**to** **k**uraberu', right, respectively) produced by a female speaker taken from a 100ms time window. Closure annotated as <cl>. Top tier represents word-level transcription, second tier contatins phone & sub-phone annotations, third tier marks prosodic boundaries via Break Index, and bottom tier contains utterance transcription.

whether the stop was fully realised, defined by whether a clear closure, burst, and voice onset could be visually observed (Kikuchi and Maekawa, 2003).

In order to ensure that stops examined in this study were fully realised, a range of classes were excluded from further analysis. First, any stop marked in the corpus as not having a clear closure and burst (56,661 tokens); stops followed by a devoiced vowel, as the onset of voicing could not be ascertained (11,939 tokens); stops immediately following hesitations (11,991 tokens); geminate stops (19,785 tokens), as geminates in Japanese are not phonologically contrastive for voicing and undergo devoicing (Kawahara, 2015); stops from word-medial contexts (72,681 tokens), as stops are expected to undergo reduction in these contexts (Cho and Ladefoged, 1999; Kim et al., 2018); and stops from non-spontaneous read speech (4,790 tokens). Prosodic position is defined in the corpus using the X-JToBI prosodic labeling scheme (Maekawa et al., 2002), which numerically represents the perceived strength of a prosodic juncture through 'Break Indices' (BIs). The labelling of a BI is based on a range of perceptual cues including segmental lengthening, F0 reset, and changes in voice quality (Venditti, 2005). Junctures with a BI value of 1 typically represents an word boundary internal to an Accentual Phrase (AP), BI value typically represents the boundary between two APs, whilst BI values of 3 typically indicate the edge of an Intonational Phrase (IP). In this analysis, all tokens with *no* BI value (which are predominantly word-medial) were excluded. The set of stops analysed is therefore word-initial stops with any potentially-problematic cases excluded.

3.2 Voicing during closure (VDC)

The goal of the VDC measure is to characterise the presence of closure voicing, which plays a key part in signalling phonological voicing in Japanese. Traditionally this has been characterised in terms of 'negative VOT' (voicing beginning during the closure and continuing up to burst onset); it has been long known, however, that realisation of voicing within the stop closure is further complicated in connected speech, compared with realisation in isolated word productions (Lisker and Abramson,

1964, 1967; Abramson and Whalen, 2017). Voicing may continue for the entire stop closure ('full voicing'), or may subside ('bleed') and return just prior to the release ('trough') (Davidson, 2016). Cases like this make applying a traditional definition of 'negative VOT' difficult for the purposes of characterising the voicing pattern. Davidson (2016, 2018) observed in North American connected speech that closure voicing corresponding to negative VOT was incredibly rare, constituting only a handful of tokens. Moreover, Davidson noted the likelihood of producing closure voicing in English was closely tied to the voicing properties of the preceding segment: preceding voicing segments (vowels, sonorants) were more likely to induce closure voicing than voiceless segments. This is important for this study, where *all* preceding segments are voiced: Japanese syllables are either open (i.e., consonant-vowel) or contain a nasal coda (Tsujimura, 2014): as geminated stops have been excluded, this means that all stops are preceded by either a vowel or a nasal (potentially with an intervening pause). The presence of a preceding vowel does not guarantee the realisation of voicing in the stop closure, however: Figure 1 (left) shows a voiced stop realised with voicing throughout the whole stop closure ('full voicing'), whilst no such closure voicing is evident in a voiceless stop in the same phonetic context (Figure 1 right).

Within this study, the goal of the VDC measurement is to characterise the presence of phonetic voicing during closure in terms of the likely presence of an active closure voicing gesture. In order to capture this distinction, the presence of VDC is defined in binary terms between either the *presence* or *absence* of active closure voicing. This would aim to exclude the common cases of passive voicing which is often short and weak (less than 20 millieseconds) in amplitude, in contrast to an active voicing gesture, characterised by clear periodic voicing for a substantial portion of the closure and the presence of pitch. This somewhat deviates from previous studies on English using similar approaches (Davidson, 2016; Sonderegger et al., 2020) where closure voicing was trichotomised into 'no', 'partial', or 'full' voicing, determined by the relative portion of the observed voicing within the closure. The decision to use a binary voicing distinction in this study was based on the goal of restricting to cases whether an active voicing target was present or not, as well as on the empirical observation that both Davidson (2016) and Sonderegger et al. (2020) found that effects were more apparent in their respective binary ('no' versus 'full') models than comparing the relative degrees of voicing. This characterisation of closure voicing as distinct from a measure of VOT that may be either 'positive' or 'negative' enables both voicing presence and positive VOT to be examined as independent cues to stop production: given the observations that it is possible for speakers to produce stops with both closure voicing and positive VOT (Abramson and Whalen, 2017; Kim et al., 2018; Sonderegger et al., 2020), it would be important to know if speakers are able to modulate both VOT and the presence of closure voicing independently for the purposes of signalling the voicing contrast.

In order to calculate a measure of VDC, both the mean F0 and the 'fraction of unvoiced frames' were extracted from the labeled stop closure using Voice Report in Praat (Boersma and Weenink, 2017). As Voice Report has been known to produce inaccurate measurements of voicing in specific circumstances, the calculations in this study followed the recommendations of Eager (2015): specifically, the Voice Report measurement was performed inside of a Praat script without using the Editor window, gender-specific pitch ranges (70-250Hz for males; 100-300Hz for females), and a time step of 0.001 seconds. The percentage of voicing in the closure was calculated by subtracting 100 from Voice Report's proportion of the interval with *no* voicing: for example, if Voice Report returned an unvoiced closure value of 66%, then voicing % = $100 - 66 = 34$.

As noted above, the main goal involved determining which instances of voicing were most likely produced with targeted voicing gesture for the stop. For the purposes of this study, two criteria

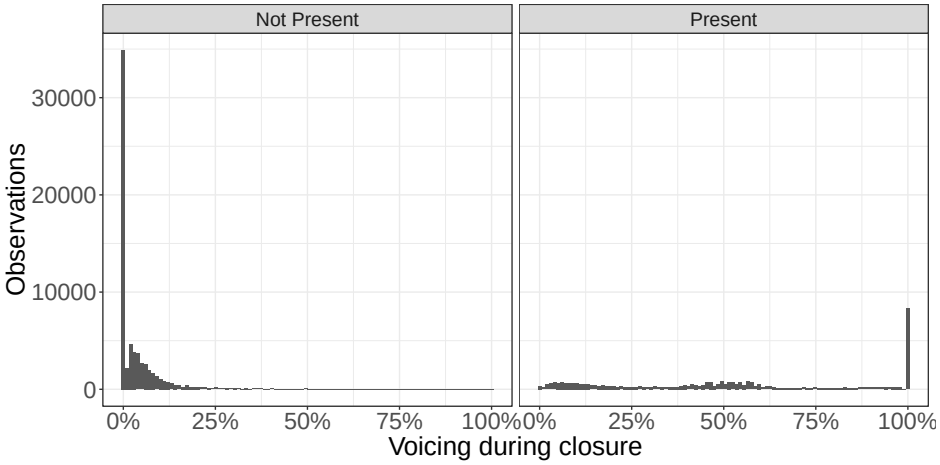

Figure 2: Histograms showing the distribution of the percentage of voicing during closure by whether F0 was also detected within the stop closure. 100 bins used within each histogram, meaning that each bar represents 1%.

were used to determine whether or not a closure contained an active voicing gesture, and tokens for which these criteria agreed were kept for the final analysis. The first was whether F0 was present in the closure, and the second was whether a significant portion of the closure contained voicing. Numerous values have been proposed in the literature for what proportion of the closure reflects active voicing, such as 'greater than 50%' (Abramson and Whalen, 2017) and 'greater than 10%' (Davidson, 2016). Here, decisions regarding the cutoffs were determined by examining the distribution of closure voicing percentages with and without the presence of F0. As shown in Figure 2, closure voicing with no accompanying F0 (left panel) ranges from 0% to approximately 15%, and so VDC (in terms of an active voicing gesture) was considered to be absent for such tokens. When F0 is present (right panel), a large number of tokens exhibited 100% closure voicing with a small cluster around 50%. To include these tokens, the 'present' VDC category was decided as tokens with the presence of F0 and at least 35% voicing in the closure. The other cases were taken to indicate that voicing was unreliable: F0 may have been present but the lack of substantial voicing % suggests potential voicing bleed. These unreliable tokens were excluded (18,960; 17.5%), meaning we have confidence that all remaining tokens are realised with either no closure voicing or an active voicing gesture. After all exclusions, the final dataset used for analysis contained 90,160 tokens (3,440 types) from 137 speakers (58 female), corresponding to an average of 658 tokens per speaker (range: 149-2,913).

3.3 Models

The goal of this study is to examine the evidence of structured speaker variability (1) within individual acoustic cues; (2) in the voicing contrast across cues; and (3) across cues within individual phonetic categories. In order to address these questions, VOT and VDC were statistically modelled to characterise individual speaker differences whilst controlling for a range of factors known to influence both cues (Section 2.1). VOT (log-transformed) and VDC were jointly modelled using a multivariate Bayesian mixed model using *brms* (Bürkner, 2018), an R front-end for the Stan programming

language (Carpenter et al., 2017).[3] A Bayesian model returns a *distribution* of potential values for all model parameters, which makes it possible to estimate correlations across speakers as well as the uncertainty associated with each correlation. This is ideal for addressing all three research questions, as it means that the strength of relationships across speakers can be characterised formally in terms of both the strength of the correlations and the range of possible correlations consistent with the data. As both VOT and VDC are fit within the same model, it is possible to also directly estimate the speaker correlations *across* phonetic cues, which is crucial for research questions (2) and (3). Finally, the use of a statistical model to estimate speaker correlations, rather than estimating correlations from empirical data as in most previous work on structured speaker variability, allows for correlations (and individual speaker values for each cue) to be estimated whilst controlling for the range of other factors known to affect both VOT and VDC (Sec. 2.1).

The model consists of a sub-model predicting VOT and a sub-model predicting VDC, and terms linking these sub-models together. We first describe the terms in each sub-model, which were identical. Each sub-model included the following population-level ('fixed-effect') predictors for stop **voicing**, previous phoneme **manner**, speaker **birth year** and **gender**, stop **place of articulation**, speech **style**, prosodic **position**, log-transformed word **frequency**, speaker **mean** and **local** (relative to mean) speech **rate** (Sonderegger et al., 2014; Stuart-Smith et al., 2015), the presence of a preceding **pause**, following vowel **height** and **duration**. To control how each predictor influenced the realisation of the voicing contrast, two-way interation terms between stop voicing and all other predictors were also included in the model. Continuous predictors (speaking rates, frequency, vowel duration) were centred and divided by two standard deviations (Gelman and Hill, 2007). Two-level factors (voicing, accent, gender, vowel height, pause) were converted into binary (0/1) measures, scaled, and centred. Predictors with three or more levels (birth year, place of articulation, phoneme manner) were coded with sum contrasts. For group-level ('random-effect') predictors, the model was fit with a random intercept for words; speaker-level effects consisted of a random intercept and random slopes for all population-level predictors (with the exception of style, age, and gender). As the relationship between a speaker's overall value for VOT/VDC and the size of their voicing contrast is of direct interest to this study, both models included a correlation term between the speaker-level intercept and the voicing predictor. The VOT and VDC sub-models were tied together by three correlations between the key speaker-level effects: intercepts, voicing, and the correlation between them. For example, the correlation term between the VOT intercept and the VDC intercept captures the extent to which speakers with higher mean VOT are more likely to use VDC. The model used 8000 samples across 4 Markov chains and was fit with weakly-informative 'regularising' priors (Nicenboim and Vasishth, 2016; Vasishth et al., 2018b) of normal distributions with a mean of 0 and standard deviations of 1 and 0.5, and 0.5 for VOT intercept, VDC intercept, and fixed effect parameters respectively. The default prior in *brms* for group-level effects was used: a half Student's t-distribution with 3 degrees of freedom and a scale parameter of 10. Correlations used the LKJ prior (Lewandowski et al., 2009) with $\zeta = 2$, in order to give lower prior probability to perfect (1/-1) correlations, as recommended by Vasishth et al. (2018b).[4]

[3] See Nicenboim and Vasishth (2016); Vasishth et al. (2018a,b) for further information on the application of Bayesian regression modelling for linguistic and phonetic research.

[4] To ensure that the correlations reported were not due to the choice of a specific prior, an identical model with a weaker 'flat' prior ($\zeta = 1$) was also fit. The correlations estimated from this model were near identical (within 0.01) to those from the stronger model, indicating that the evidence for the correlations in the data is strong enough as to not be affected by the subjective choice to use a more informative prior.

Correlation	ρ	2.5% CrI	97.5% CrI	Pr($\rho <> 0$)
Voiceless VOT, Voiced VOT	0.77	0.709	0.821	1
Voiceless VDC, Voiced VDC	0.664	0.594	0.729	1

Table 1: Median correlation, 95% credible intervals (CrI), and posterior probability of within-cue correlations (Spearman's ρ) across speakers sampled from the model posterior with all other predictors held at their 'average values' (e.g., mean word frequency, mean across all places of articulation, etc).

4 Results

The research questions concern the relationships observed across speakers both *within* each cue (1) as well as *across* both cues (2, 3), and so correlations were calculated for each of the 8000 draws from the posterior sample and reported as the median, 95% credible interval (CrI), and the posterior probability of the parameter not including 0, using `fitted_draws` and `median_qi`, respectively, from the *tidybayes* package (Kay, 2019).[5] Speaker-level variability is first examined *within* VOT and VDC separately (4.1) before examining the relationships *between* both cues across speakers (4.2). Following the suggestions of Nicenboim and Vasishth (2016), we consider there to be strong evidence for a non-null effect if the 95% CrI for the parameter do not include 0; if 0 is within the 95% CrI but the probability of the parameter not changing direction is at least 95%, this is considered to represent weak evidence for a given effect. Crucially the strength of evidence for an effect is distinct from its magnitude, and so the strength of a given predictor's effect on VOT/VDC is considered alongside its relative evidence. The size or magnitude of a given correlation is assessed in terms of Cohen's (1988) conventions: correlations with sizes between 0 and 0.1 (in either direction) are considered to be *negligible*; those with sizes between 0.1 and 0.3 to be *small*; between 0.3 and 0.5 to be *medium*, and *strong* correlations have values larger than 0.5. Cohen's conventions are considered to be heuristic and should be considered relative to previous effect sizes observed for a given phenomenon. Given the relatively small body of work examining the speaker-level relationships across speakers, it is considered that Cohen's conventions allow us to have some initial benchmark with which to evaluate the relative relationships within and across phonetic cues.

4.1 Within-cue variability

The effects of the population-level parameters on VOT were as expected, including the size of voicing contrast (Table 3, Appendix A). As the VOT voicing contrast is maintained across all population-level effects (i.e., no parameter neutralised or reversed the voicing contrast) and it is speaker-level variability which is of interest for our research questions, these parameters provide controls for the speaker-level variability, and thus the fixed-effects will not be discussed further. Figure 3 (left) demonstrates the strong correlation between speakers' voiced and voiceless VOTs (95% CrI = [0.709, 0.821]; Table 1, row 1): each point represents a speaker's median estimated voiceless (x-axis) and voiced (y-axis) VOT value. All individual speakers have higher VOTs for voiceless than voiced stops, indicated by all points appearing on one side of the dashed $y = x$ line. Speakers differ in their particular VOT

[5]These are conceptually related to an estimated value, confidence interval, and p-value that one would typically use to report a correlation using a (non-Bayesian) hypothesis test, though differ somewhat in their interpretation (see Nicenboim and Vasishth, 2016).

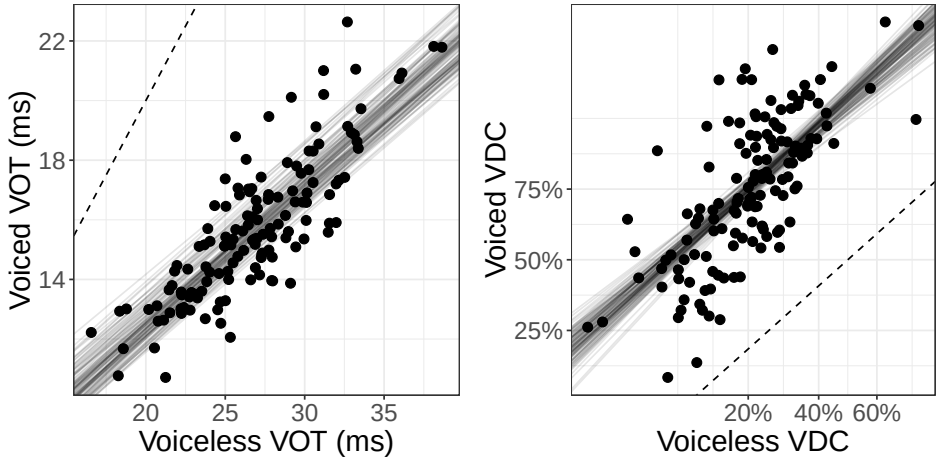

Figure 3: Model-estimated cue values for VOT (left) and VDC (right) for voiceless (x-axis) and voiced (y-axis) stops. One point per speaker. Black lines are 100 lines of best fit drawn from the model posterior to show direction and uncertainty in the correlation. Dashed line is $y = x$, where the value for voiceless stops equals that for voiced stops. VOT plot in linear (millisecond) scale; VDC plot is in logit-scaled probability scale to illustrate differences at extreme upper and lower probabilities.

values, but the relative difference between their voiced and voiceless VOTs (i.e., the voicing contrast) is consistent: the regression lines demonstrate this linear relationship, where speakers both maintain the contrast between stops, and speakers with long VOTs for voiceless stops also have long VOTs for voiced stops.

As for VDC, no population-level effect neutralised or reversed the VDC voicing contrast (Table 4, Appendix B), meaning that VDC is always predicted to be more likely for voiced than voiceless stops ($\hat{\beta}$ = 2.99, CrI = [2.76, 3.21], Pr($\hat{\beta} > 0$) = 1). Note, however, that the large effect of the presence of a preceding pause on VDC, which suggests that speakers producing spontaneous Japanese are substantially less likely to produce VDC directly following a pause ($\hat{\beta} = -3.24$, CrI = $[-3.51, -2.97]$, Pr($\hat{\beta} < 0$) = 1), consistent with recent experimental findings (Gao and Arai, 2019). Comparing across voicing categories, Figure 3 (right) illustrates that speakers maintain a strong positive relationship between their voiced and voiceless VDCs (95% CrI = $[0.594, 0.729]$; Table 1, row 2). No speaker has a reversed voicing contrast for VDC, reflected by all speaker values (represented as points) appears above the $y = x$ line. The multiple regression lines illustrate that, as with VOT, speakers who are more likely to produce VDC for voiced stops are also more likely, on average, to produce voiceless stops with VDC.

4.2 Across-cue variability

In the previous section, research question (1) was addressed by examining how speakers vary *within* a single cue (VOT, VDC) between voiced and voiceless stops. We now address whether speakers vary *across* cues in production, where speakers may coordinate both cues in signalling the stop voicing contrast (question 2), or specific segments (question 3). Comparing the size of the voicing contrast for each cue, a weak positive relationship across speakers can be observed (95% CrI = $[-0.001, 0.346]$; Table 2, row 1): this can be interpreted as meaning that the voicing contrast sizes across

	Correlation	ρ	2.5% CrI	97.5% CrI	Pr($\rho <> 0$)
Voicing contrast	VOT contrast, VDC contrast	0.198	−0.001	0.346	0.974
Within-category	Voiced VOT, Voiced VDC	−0.348	−0.423	−0.27	1
	Voiceless VOT, Voiceless VDC	0.135	0.038	0.228	1
Across-category	Voiceless VOT, Voiced VDC	−0.152	−0.233	−0.066	0.99
	Voiced VOT, Voiceless VDC	0	−0.092	0.093	0.5

Table 2: Median correlation, 95% credible intervals (CrI), and posterior probability of across-cue correlations (Spearman's ρ) across speakers sampled from the model posterior with all other predictors held at their 'average values' (e.g., mean word frequency, mean across all places of articulation, etc). VOT contrast = voiceless VOT − voiced VOT; VDC contrast = voiced VDC − voiceless VDC.

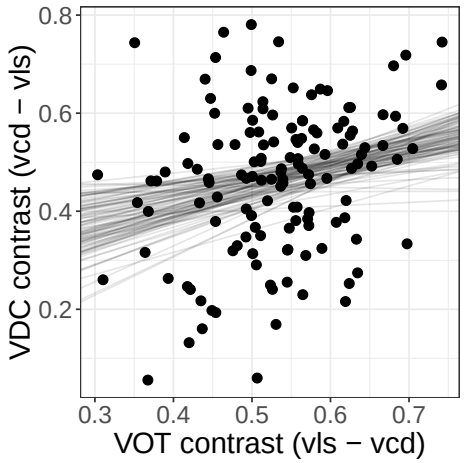

Figure 4: Model-estimated voicing contrast sizes for VOT (x-axis) and VDC (y-axis). One point per speaker. Black lines are 100 lines of best fit drawn from the model posterior to show direction and uncertainty in the correlation.

cues are somewhat linked, with speakers differing in precisely how they realise the voicing contrast simultaneously across both VOT and VDC (Figure 4).

Given the strong correlations across speakers in single use of a given cue (Figure 3) and the observation that speakers only weakly vary in the size of their voicing contrast across both cues (Figure 4), the question remains as to how speakers covary in the use of VOT and VDC within specific phonetic categories. In other words, do speakers' values for one cue (e.g., VOT) within a category (e.g., voiceless stops) correlate with their values for the other cue (VDC) in that same category? Figure 5 demonstrates this combination of cues and voicing categories, and illustrates an asymmetry in the VOT-VDC relationship between voiced and voiceless stops. Speakers strong evidence for a negative relationship of medium strength between VOT and VDC in voiced stops (Figure 6, top left), meaning that speakers with larger voiced VOTs have a lower voiced VDC likelihood (95% CrI = [−0.423, −0.27]; Table 2, row 2). For voiceless stops, however, there is strong evidence for a weak *positive* relationship (95% CrI = [0.038, 0.228]; Figure 6, bottom left; Table 2, row 3), though even the upper

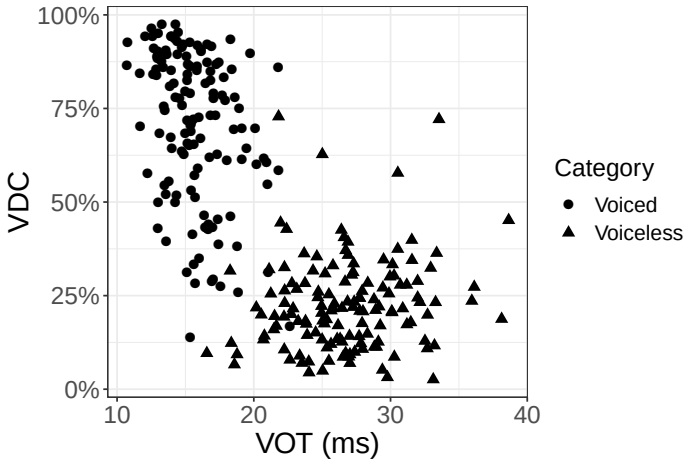

Figure 5: Model-estimated cue values for VOT (x-axis) and VDC (y-axis). Voicing category of the stop is represented by shape (points = voiced; triangles = voiceless). One value per speaker-category pair, meaning each speaker is represented by one point and one triangle.

credible interval value is considered weak by Cohen's (1988) conventions. A negative relationship is also observed between speakers' voiced VDC rate and their voiceless VOTs, though this is much smaller in magnitude than the voiced VOT-voiced VDC relationship (95% CrI = $[-0.233, -0.066]$; Figure 6, bottom right; Table 2, row 4); voiceless VDC does not show a meaningful correlation with voiced VDC across speakers (95% CrI = $[-0.092, 0.093]$; Figure 6, top right; Table 2, row 5).

5 Discussion

The phonetic realisation of a given segment is well-known to differ across languages, dialects, phonetic contexts, and individual speakers. Recent research has observed that this variability across individual speakers is *structured*: whilst speakers may differ in the overall value of a particular phonetic cue (e.g., stop VOT), they also demonstrate covariation in the use of one or more cues in the marking of linguistic contrasts (e.g., Theodore et al., 2009; Chodroff and Wilson, 2017, 2018; Sonderegger et al., 2020). Such constraints on phonetic realisation aid in speaker normalisation (Kleinschmidt, 2018) by reducing the range and dimensions in which a given speaker can vary, and may represent some form of inherent constraint as a property of the speaker's grammar (Chodroff and Wilson, 2017, 2018). Much of the previous empirical work on structured speaker variability has focused on controlled speech styles: the degree to which such structure is reliably maintained in less-controlled speech is still an open question (Sonderegger et al., 2020). In addition, the majority of previous studies have focused on stop contrasts in English and German, which both share a similar aspiration-driven phonetic implementation. Given that the English stop system is largely structured around varying degrees of positive VOT, it is not known how speaker variability may be structured in a language that phonetically and phonologically differs from English in the signalling of linguistic contrasts.

This study has attempted to both address these empirical gaps and broaden the search for structured speaker variability through the examination of more than one cue to word-initial stop voicing in spontaneous Japanese speech. Specifically, this study has investigated how speakers systematically

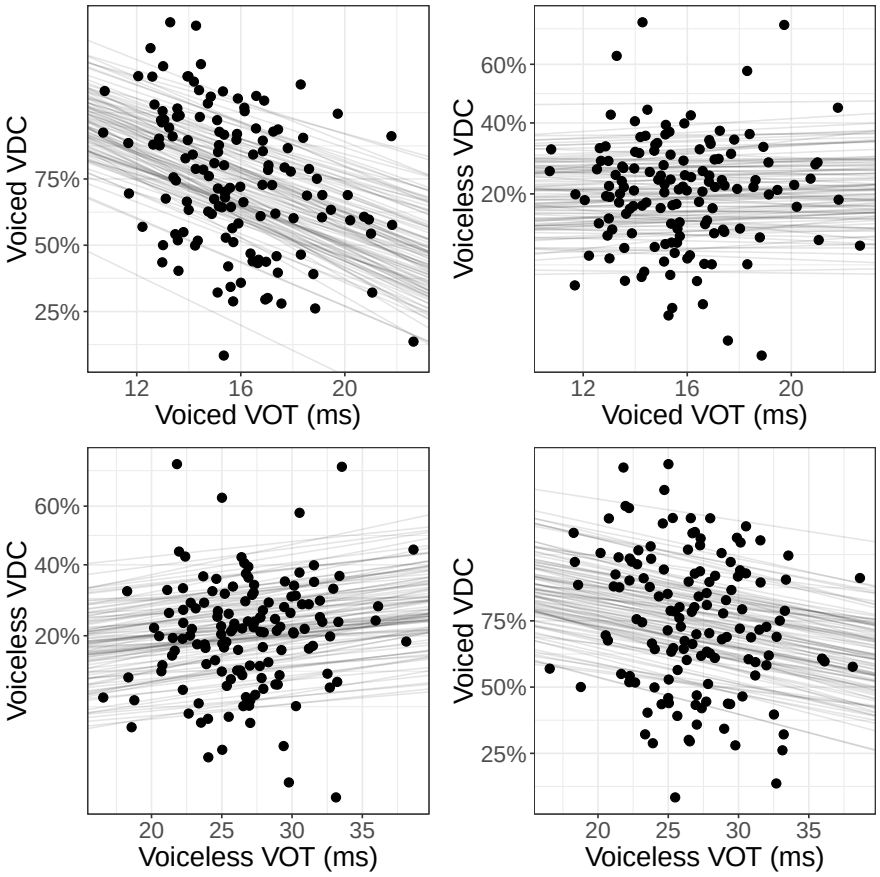

Figure 6: Model-estimated cue values for VOT (x-axis) and VDC (y-axis), comparing relationship between cues either within (left) or across (right) a given stop category. One point per speaker. Black lines are 100 lines of best fit drawn from the model posterior to show direction and uncertainty in the correlation. VOT in linear (ms) scale; VDC in logit-scaled probabilities to show differences at extreme probabilities (near 0% or 100%).

vary in the separate and combined use of VOT and the presence of voicing during the stop closure (VDC) as cues to the word-initial voicing contrast. Strong within-cue relationships are observed across speakers between voiced and voiced stops: whilst speakers differ in their overall of VOT or VDC, speakers are consistent in the relative *difference* of VOT or VDC in marking the voicing contrast. These within-cue relationships are of comparable magnitude to the strongest correlations observed for English stops in both laboratory (Chodroff and Wilson, 2017, 2018) and spontaneous English speech (Sonderegger et al., 2020), demonstrating that structured speaker variability is present in voicing systems beyond the English aspiration-system, and in more than one independent cue to a single contrast in spontaneous speech.

Here, most of the predictable variability across individual speakers is *within* a given phonetic cue (4.1), as compared with variability *across* the two cues (4.2): no across-cue relationship (Table 2) is as strong as either of the within-cue correlations (Table 1).[6] The size of the voicing contrasts

[6]For every within-cue correlation r_w and cross-cue correlation r_b, the posterior probability $P(r_w > r_b)$ is > 0.99.

between VOT and VDC is positively correlated across speakers (Figure 4). This could be interpreted as evidence that speakers vary in the degree of 'clarity' in their speech: speakers align multiple cues to a voicing contrast simultaneously for the purposes of maximising the acoustic distinctiveness between the categories, as opposed to emphasising one cue over another (Bang, 2017; Clayards, 2018). An explanation in terms of speech clarity does not straightforwardly apply in this data, however, for two reasons. First, the size of the correlation itself is small (Table 2, row 1), reflecting only a weak relationship between the two cue contrast sizes. Second, this predictive pattern for the use of VOT and VDC is observed only for voiced stops: whilst the VOT-VDC relationship is negatively correlated in voiced stops, no clear relationship between the cues is observed for their voiceless counterparts (Table 2; Figure 5). This suggests that the VOT-VDC cue relationship is *asymmetric* between stop voicing categories. Such an observation could be interpreted as restriction on structured speaker variability for only segments in a series (i.e., voiced and voiceless stops) that have some form of featural specification. It has been previously argued that Japanese is a 'voiced' language (Ito and Mester, 1995; Nasukawa, 2005) in the sense of being specified exclusively for a monovalent [voice] feature on voiced stops and no featural specification for voiceless stops (e.g., Iverson and Salmons, 1995; Salmons, 2019).

The within-cue findings (Section 4.1) suggest that speakers are able to use cues independently for the purposes of marking a linguistic contrast *without* maintaining the same cross-category relationships across more than one phonetic cue. This supports a restricted form of structured variability, constraining the predictability of speakers of spontaneous Japanese in their realisation of phonological categories along a single phonetic dimension. Crucially, speakers use two cues to *separately* realise the same phonological contrast. In this sense, the structured variability is *constrained*: in this study, speaker variability is present within the use of a single acoustic cue, but speakers are less consistent in simultaneous use of multiple cues to the stop voicing contrast.

Whe considered from the perspective of the 'principle of uniformity' on phonetic variation (Chodroff and Wilson, 2017, 2018, 2020), the results reported here provide some evidence for uniformity across speakers: namely, speakers are highly consistent *within* cues in signalling stop voicing contrast. These results also demonstrate that the notion of phonetic uniformity must be assumed to be subject to constraints: here we find evidence of speakers covarying within individuals cues, as opposed to covarying across more than one cue in marking the same contrast. Japanese differs from English in how the stop voicing contrast is specified and realised: Japanese maintains a 'hybrid' stop voicing system involving the use of both positive VOT and closure voicing (Shimizu, 1996; Nasukawa, 2005; Tsujimura, 2014). Thus our evidence for covariation from this stop voicing stystem suggests that that phonetic uniformity is constrained by language-specific properties. Our findings emphasise the importance for examining the evidence for uniformity in a range of empirical contexts, and especially across languages which differ in their phonetic implementation for a given linguistic contrast.

6 Conclusion

By examining structured variability in this context, this study has demonstrated that structured variability is both present in a new empirical setting, but also that structured variability is constrained in ways which are not straightforwardly predicted from previous observations in English-based studies. Specifically, this constraint arises from the linguistic specification and phonetic implementation of stop voicing in Japanese which requires a different configuration of acoustic cues than English. Such a finding motivates an expanded search for structured speaker variability across a range of languages

and cues. Within Japanese, for example, this could mean the inclusion of F0 as an acoustic cue, given recent research illustrating its increasing importance to the stop voicing contrast (Kong et al., 2014; Gao et al., 2019; Gao and Arai, 2019). This study has provided the first sketch of a more complex picture for the structure of speaker variability, and motivates the expanded search of structured variability across a range of languages, cues, and contrasts.

References

Abramson, A. S. and Whalen, D. H. (2017). Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics*, 63:75–86.

Allen, S. J., Miller, J. L., and DeSteno, D. (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America*, 113:544–552.

Bang, H.-Y. (2017). *The structure of multiple cues to stop categorization and its implications for sound change*. PhD thesis, McGill University.

Baran, J., Laufer, M., and Daniloff, R. (1977). Phonological contrastivity in conversation: a comparative study of voice onset time. *Journal of Phonetics*, 5:339–350.

Boersma, P. and Weenink, D. (2017). Praat: doing phonetics by computer (version 6.0.36).

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1):395–411.

Bybee, J. B. (2001). *Phonology and Language Use*. Cambridge University Press, Cambridge.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Cho, T. and Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics*, 27:207–229.

Chodroff, E. and Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, 61:30–47.

Chodroff, E. and Wilson, C. (2018). Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice. *Linguistics Vanguard*, 4.

Chodroff, E. and Wilson, C. (2020). Uniformity in phonetic realization: Evidence from sibilant place of articulation in American English. unpublished manuscript.

Clayards, M. (2018). Individual talker and token covariation in the production of multiple cues to stop voicing. *Phonetica*, 75:1–23.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Earlbaum Associates, Hillsdale, NJ.

Davidson, L. (2016). Variability in the implementation of voicing in American English obstruents. *Journal of Phonetics*, 54:35–60.

Davidson, L. (2018). Phonation and laryngeal specification in American English voiceless obstruents. *Journal of the International Phonetic Association*, 48:331–356.

Docherty, G. (1992). *The timing of voicing in British English obstruents*. Foris, Berlin & New York.

Eager, C. (2015). Automated voicing analysis in Praat: statistically equivalent to manual segmentation. In *Proceedings of the 18th International Congress of Phonetic Sciences*. University of Glasgow.

Foulkes, P., Docherty, G., and Watt, D. (2001). The emergence of structured variation. *University of Pennsylvania Working Papers in Linguistics*, 7:67–84.

Gao, J. and Arai, T. (2019). Plosive (de-)voicing and f0 perturbations in Tokyo Japanese: positional variation, cue enhancement, and contrast recovery. *Journal of Phonetics*, 77:1–33.

Gao, J., Yun, J., and Arai, T. (2019). VOT and F0 coarticulation in Japanese: production-biased or misparsing? In *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge.

Hullebus, M. A., Tobin, S. J., and Gafos, A. I. (2018). Speaker-specific structure in German voiceless stop voice onset times. In *Proceedings of Interspeech 2018*, pages 1403–1407, Hyderabad.

Hunnicutt, L. and Morris, P. A. (2016). Prevoicing and aspiration in Southern American English. In *Proceedings of the 39th Annual Penn Linguistics Conference*, volume 22, Philadelphia. University of Pennsylvania.

Ito, J. and Mester, A. R. (1995). Japanese phonology. In Goldsmith, J. A., editor, *The Handbook of Phonological Theory*, pages 817–838. Blackwell.

Iverson, G. and Salmons, J. (1995). Aspiration and laryngeal representation in Germanic. *Phonology*, 12:369–396.

Johnson, K., Ladefoged, P., and Lindau, M. (1993). Individual differences in vowel production. *Journal of the Acoustical Society of America*, 94:701–714.

Kawahara, S. (2015). Geminate devoicing in Japanese loanwords: Theoretical and experimental investigations. *Language and Linguistics Compass*, 9:181–195.

Kay, M. (2019). *tidybayes: Tidy Data and Geoms for Bayesian Models*. R package version 1.0.4.

Kikuchi, H. and Maekawa, K. (2003). Performance of segmental and prosodic labeling of spontaneous speech. In *Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo. Tokyo Institute of Technology.

Kim, S., Kim, J., and Cho, T. (2018). Prosodic-structural modulation of stop voicing contrast along the VOT continuum in trochaic and iambic words in American English. *Journal of Phonetics*, 71:65–80.

Klatt, D. (1975). Voice onset time, frication and aspiration in word-initial consonant clusters. *Journal of Speech, Language and Hearing Research*, 18:686–706.

Kleinschmidt, D. F. (2018). Structure in talker variability: How much is there and how much can it help? *Language, Cognition, and Neuroscience*, 34:1–26.

Kong, E. J., Yoneyama, K., and Beckman, M. E. (2014). Effects of a sound change in progress on gender-marking cues in Japanese. In *Proceedings of LabPhon 14*, Tokyo. NINJAL.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100:1989–2001.

Liberman, A. M., Delattre, P. C., and Cooper, F. S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, 1:153–167.

Liberman, M. A., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74:431–461.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the h&h theory. In Hardcastle, W. J. and Marchal, A., editors, *Speech Production and Speech Modelling*, volume 4 of *NATO ASI Series*, pages 403–439. Kluwer Academic Publishers.

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

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

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

Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). Spontaneous speech corpus of Japanese. In *Proceedings of the Second International Conference of Language Resources and Evaluation (LREC)*, volume 2, pages 946–952.

Maekawa, K., Koiso, H., Igarashi, Y., and Venditti, J. (2002). X-JToBI: An extended J_ToBI for spontaneous speech. In *Proceedings of the 7th International Conference on Spoken Language Processing*, pages 1545–1548, Denver.

Nasukawa, K. (2005). The representation of laryngeal-source contrasts in Japanese. In van de Weijer, J., Nanjo, K., and Nishihara, T. ., editors, *Voicing in Japanese*, pages 71–87. De Gruyter Mouton.

Nicenboim, B. and Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas - part II. *Language and Linguistics Compass*, 10:591–613.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24:175–184.

Pierrehumbert, J. B. (2001). Exemplar dynamics: word frequency, lentition, and contrast. In Bybee, J. and Hopper, P., editors, *Frequency and the Emergence of Linguistic Structure*, pages 137–157. John Benjamins.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Fosler-Lussier, E. (2007). *Buckeye Corpus of Spontaneous Speech*. Ohio State University, Columbus, 2 edition.

Riney, T. J., Takagi, N., Ota, K., and Uchida, Y. (2007). The intermediate degree of VOT in Japanese initial stops. *Journal of Phonetics*, 35:439–443.

Salmons, J. (2019). Laryngeal phonetics, phonology, assimilation and final neutralization. In Page, R. and Putnam, M. T., editors, *Cambridge Handbook of Germanic Linguistics*, pages 119–142. Cambridge University Press, Cambridge.

Schultz, A. A., Francis, A. L., and Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *Journal of the Acoustical Society of America*, 132.

Shimizu, K. (1996). *A cross-language study of the voicing contrasts of stop consonants in Asian languages*. Seibido, Tokyo.

Sonderegger, M., Bane, M., and Graff, P. (2014). The medium-term dynamics of accents on reality television. *Language*, 93:598–640.

Sonderegger, M., Stuart-Smith, J., Knowles, T., MacDonald, R., and Rathcke, T. (2020). Structured heterogeneity in Scottish stops over the twentieth century. *Language*. Accepted.

Stuart-Smith, J., Sonderegger, M., Rathcke, T., and Macdonald, R. (2015). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology*, 6:505–549.

Takada, M. (2011). *Nihongo no gotou heisa'on no kenkyuu: VOT no kyoujiteki bunpu to tsuujiteki henka [Research on the word-initial stops of Japanese: synchronic distribution and diachronic change in VOT]*. Kurosio, Tokyo.

Takada, M., Kong, E. J., Yoneyama, K., and Beckman, M. E. (2015). Loss of prevoicing in Modern Japanese /g, d, b/. In *Proceedings of the 18th International Congress of Phonetic Sciences*. University of Glasgow.

Tanner, J., Sonderegger, M., and Stuart-Smith, J. (2019). Structured speaker variability in spontaneous Japanese stop contrast production. In *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia.

Theodore, R. M., Miller, J. L., and DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *Journal of the Acoustical Society of America*, 126:3974–3982.

Tsujimura, N. (2014). *Introduction to Japanese Linguistics*. Wiley-Blackwell, Oxford.

Vasishth, S., Mertzen, D., Jäger, L. A., and Gelman, A. (2018a). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, 103:151–175.

Vasishth, S., Nicenboim, B., Beckman, M., Li, F., and Kong, E. J. (2018b). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, 71:147–161.

Venditti, J. (2005). The J_ToBI model of Japanese intonation. In Sun-Ah, J., editor, *Prosodic Typology*, pages 172–200. Oxford University Press, Oxford.

Yao, Y. (2009). Understanding VOT variation in spontaneous speech. *UC Berkeley Phonology Lab Annual Report*, pages 29–43.

Appendix A Population-level effects (VOT)

Predictor	$\hat{\beta}$	Error	2.5% CrI	97.5% CrI
Intercept	3.11	0.02	3.08	3.15
Voicing	-0.51	0.02	-0.54	-0.48
Gender	-0.09	0.03	-0.15	-0.03
Previous phoneme manner (long)	0.03	0.00	0.03	0.04
Previous phoneme manner (nasal)	0.03	0.00	0.02	0.04
Birth year (1960-69)	0.04	0.02	-0.01	0.09
Birth year (1950-59)	0.03	0.02	-0.02	0.08
Birth year (1940-49)	0.00	0.03	-0.06	0.06
Birth year (1930-39)	-0.02	0.04	-0.09	0.05
Place of articulation (alveolar)	-0.18	0.01	-0.20	-0.15
Place of artciulation (velar)	-0.12	0.01	-0.14	-0.10
Speech style (public speaking)	-0.10	0.00	-0.11	-0.09
Style style (dialogue)	0.01	0.00	0.00	0.02
Break Index (2)	0.05	0.00	0.05	0.06
Break Index (3)	0.05	0.00	0.04	0.05
Frequency (log)	-0.04	0.01	-0.05	-0.03
Speech rate (mean)	-0.06	0.03	-0.12	0.01
Speech rate (local)	-0.03	0.00	-0.04	-0.02
Preceding pause	0.04	0.01	0.02	0.05
Vowel height	0.14	0.01	0.11	0.16
Voicing : Gender	0.08	0.02	0.03	0.12
Voicing : Previous phoneme manner (long)	0.02	0.01	0.01	0.03
Voicing : Previous phoneme manner (nasal)	0.02	0.01	0.00	0.04
Voicing : Birth year (1960-69)	-0.03	0.02	-0.07	0.00
Voicing : Birth year (1950-59)	-0.05	0.02	-0.08	-0.01
Voicing : Birth year (1940-49)	0.04	0.02	0.00	0.08
Voicing : Birth year (1930-39)	-0.03	0.03	-0.08	0.02
Voicing : Place of articulation (alveolar)	0.05	0.02	0.02	0.09
Voicing : Place of articulation (velar)	0.06	0.01	0.04	0.09
Voicing : Speech style (public speaking)	0.03	0.01	0.01	0.04
Voicing : Speech style (dialogue)	-0.02	0.01	-0.03	0.00
Voicing : Break Index (2)	-0.06	0.01	-0.07	-0.05
Voicing : Break Index (3)	-0.04	0.00	-0.05	-0.03
Voicing : Frequency (log)	0.01	0.01	-0.01	0.04
Voicing : Speech rate (mean)	0.05	0.03	0.00	0.10
Voicing : Speech rate (local)	0.00	0.01	-0.02	0.01
Voicing : Preceding pause	-0.06	0.02	-0.10	-0.03
Voicing : Vowel height	-0.08	0.02	-0.13	-0.03

Table 3: Estimate ($\hat{\beta}$), error, and 95% credible intervals for all population-level ('fixed effect') predictors for log-transformed VOT.

Appendix B Population-level effects (VDC)

Predictor	$\hat{\beta}$	Error	2.5% CrI	97.5% CrI
Intercept	-1.13	0.12	-1.36	-0.90
Voicing	2.99	0.14	2.72	3.25
Gender	0.12	0.18	-0.23	0.48
Previous phoneme manner (long)	0.01	0.03	-0.06	0.07
Previous phoneme manner (nasal)	-0.17	0.05	-0.27	-0.08
Birth year (1960-69)	0.33	0.14	0.04	0.61
Birth year (1950-59)	0.40	0.15	0.10	0.69
Birth year (1940-49)	-0.01	0.18	-0.35	0.34
Birth year (1930-39)	-0.36	0.21	-0.77	0.06
Place of articulation (alveolar)	0.00	0.07	-0.14	0.13
Place of articulation (velar)	0.13	0.05	0.04	0.22
Speech style (public speaking)	0.13	0.04	0.04	0.21
Speech style (dialogue)	-0.42	0.05	-0.52	-0.33
Break Index (2)	0.39	0.03	0.32	0.45
Break Index (3)	0.53	0.02	0.49	0.58
Frequency (log)	0.17	0.04	0.09	0.26
Speech rate (mean)	-0.57	0.19	-0.95	-0.20
Speech rate (local)	-0.16	0.04	-0.23	-0.09
Preceding puase	-3.24	0.16	-3.56	-2.93
Vowel height	0.12	0.07	-0.02	0.26
Voicing : Gender	0.06	0.20	-0.34	0.45
Voicing : Previous phoneme manner (long)	-0.20	0.06	-0.32	-0.07
Voicing : Previous phoneme manner (nasal)	-0.09	0.07	-0.22	0.04
Voicing : Birth year (1960-69)	-0.03	0.17	-0.36	0.29
Voicing : Birth year (1950-59)	0.04	0.17	-0.30	0.38
Voicing : Birth year (1940-49)	0.05	0.20	-0.34	0.44
Voicing : Birth year (1930-39)	-0.32	0.24	-0.78	0.14
Voicing : Place of articulation (alveolar)	0.24	0.12	0.01	0.46
Voicing : Place of articulation (velar)	0.13	0.09	-0.04	0.31
Voicing : Speech style (public speaking)	0.52	0.07	0.38	0.67
Voicing : Speech style (dialogue)	0.13	0.08	-0.03	0.28
Voicing : Break Index (2)	-0.54	0.06	-0.64	-0.42
Voicing : Break Index (3)	-0.57	0.03	-0.63	-0.50
Voicing : Frequency (log)	0.14	0.08	-0.02	0.30
Voicing : Speech rate (mean)	0.11	0.22	-0.31	0.55
Voicing : Speech rate (local)	0.10	0.06	-0.02	0.22
Voicing : Preceding pause	2.00	0.21	1.58	2.41
Voicing : Vowel height	0.60	0.15	0.31	0.90

Table 4: Estimate ($\hat{\beta}$), error, and 95% credible intervals for all population-level ('fixed effect') predictors for VDC (logit-scale).