STRUCTURED SPEAKER VARIABILITY IN SPONTANEOUS JAPANESE STOP CONTRAST PRODUCTION

James Tanner[a], Morgan Sonderegger[a], Jane Stuart-Smith[b]

[a]Department of Linguistics, McGill University, Canada; [b]Glasgow University Laboratory of Phonetics (GULP), University of Glasgow, UK

james.tanner@mail.mcgill.ca, morgan.sonderegger@mcgill.ca, Jane.Stuart-Smith@glasgow.ac.uk

ABSTRACT

Studies of speaker variability in the realisation of stop voicing contrasts have demonstrated that differences across speakers are highly structured both within and across phonetic categories. These studies have focused on languages with similar voicing systems in scripted speech; it remains unclear how stop realisation varies in spontaneous speech more generally. This study examines speaker variability in two acoustic cues to stop voicing–Voice Onset Time and Voicing During Closure–in a corpus of spontaneous Japanese, a language undergoing change in its voicing contrast. Whilst speakers vary in both measures, this variability is highly structured: speakers with less aspirated stops are more likely to initiate voicing during the closure. However, no corresponding relationship is observed between how the two cues are used to mark the contrast. These findings extend previous work to demonstrate the structure of speaker variability in spontaneous speech.

Keywords: speaker variability, corpus phonetics, Japanese, spontaneous speech

1. INTRODUCTION

It is well established that the phonetic realisation of segments is highly variable across languages, phonological contexts, and speakers. Recent research has observed that this variability is *structured*, and exists in two predominant forms for speakers. First, speakers can systematically differ in the value of a particular phonetic parameter, such as Voice Onset Time (VOT) [1]. Second, speakers may covary in their values across phonetic categories: for example, variation in mean VOT values for [p^h] and [k^h] has been shown to be consistent across speakers of English [7] and German [12]. Furthermore, there is evidence that multiple phonetic dimensions can exist in covarying relationships, where variables exhibit correlations both within a given phonetic category as well as across phonological contrasts [9]. For stops, the focus of this paper, this work has largely focused on the relationship between VOT and F0 [24], as well as in a range of cues for a single phonetic category [8]. Covariation of this kind may facilitate speech perception by simplifying adaptation to novel speakers [7, 8]. Here, *structured variability* is used to refer to both types of variability across speakers.

Most of our prior understanding about the structure of speaker variablity comes from studies based on scripted speech [7, 8, 12] (on English, German). No work has investigated structured speaker variability in the relationship between multiple cues in spontaneous speech, much less a language using a non-West-Germanic voicing system. This empirical gap provides an opportunity to examine the realisation of the voicing contrast in spontaneous speech for languages that differ in phonetic implementation from English.

The context of this study is Japanese stop voicing, for which the contrast has been shown to exhibit variability in production. Voiced stops can be realised with either voicing lead (prevoiced) [26] or short lag VOT [30]. Voiceless stops are realised with VOT intermediate between unaspirated and aspirated stops [23]. In this sense, Japanese stop voicing appears to behave as a hybrid system between a 'true' voicing contrast and an English-style aspiration contrast. Japanese shows substantial regional variation in this contrast, and is undergoing a sound change wherein prevoiced stops are being lost in favour of using purely positive VOT [30, 31]. Given previous observations of cue variability in sound change [2] or dialect contact in other languages [25], it is possible that Japanese speakers may show a shift in the structure of the cues to the voicing contrast. This may include the incorporation of other cues (e.g. F0), but also change in the weightings of existing cues [2].

Working with spontaneous speech poses challenges for describing the realisation of stop voicing contrasts. It has been shown for English connected speech that the likelihood of producing voicing during the closure largely depends on the voicing of the preceding segment [10], which often results in voicing 'bleeding' from the preceding segment. In

Japanese, all relevant non-post-pausal stops are preceded by a voiced sonorant. As we may expect voicing during stop closure (VDC) to act as a cue to voicing (given the presence of voicing lead for voiced stops), it is entirely possible that Japanese speakers may produce VDC throughout the entire closure for voiced stops. In the absence of a clear understanding of how VDC is used in spontaneous speech, here we opt to apply a relatively broad first approximation, wherein the *presence* of VDC (i.e., VDC is either present or not, regardless of its duration) is analysed as an additional separate cue to VOT.

The goal of this study is to examine speaker-level variability in the production of the stop contrast in an apparent-time corpus of spontaneous Japanese speech, describing how speakers differ in (1) their use of VOT and VDC as cues to the stop voicing contrast, and (2) the relationship between these two phonetic dimensions which cue the contrast. As this contrast has been shown to exhibit differences as a function of age in Japanese, how these cues are also conditioned by speaker birth year is also of interest.

2. METHODS

The data for this study come from 287,042 stops from the *Corpus of Spontaneous Japanese-Core* (CSJ), a corpus of 'Standard' Japanese, comprising around 650 hours of speech [20]. The Core, a subset of the CSJ, constitutes appoximately 500,000 words (44 hours of speech), from 137 speakers (58 female) born between 1930-1979, for which phonetic annotation exists from hand-correction of automatically-generated labels [15]. During hand-correction, stops were also annotated for: stop closure duration, positive VOT (defined as the difference between beginning of the burst on closure release until the onset of voicing), and whether voicing from the preceding segment persisted into the closure.

To ensure that only fully-realised stops were examined, several classes of tokens were excluded: (1) 56,667 stops with unobserveable bursts; (2) 11,938 stops followed by devoiced vowels [19], as it would not possible to determine the onset of voicing; (3) 19,785 tokens containing geminates, given that voiced geminates are often only partially voiced [13], and so not directly comparable to singletons; (4) 11,991 stops inside hesitations; (5) 4,790 tokens from non-spontaneous speech contexts (i.e., reading passages); and (6) 72,680 word-medial stops (i.e., those with no Break Index value). This last category was because prosodic position is known to affect the production of stops cross-linguistically [6], and so this study focuses on tokens that are minimally word-initial. Prosodic position was defined using the X-JToBI system [21], which locates prosodic boundaries through the presence of Break Indices. Within this system, word-medial stops are not marked with a Break Index value, and so were excluded.

In total, 109,119 stops (corresponding to 3,731 unique word types), spoken by 137 speakers (58 female) were used in the final analysis. The final dataset contained 796 tokens from each speaker on average (range: 198-3,823). In order to provide a first characterisation of VDC, discrete binary categories were used. VDC was calculated by converting Praat Pitch files into PointProcess objects, approximating the positions of voicing periodicity. If periodicity was observed within the annotated stop closure, VDC = 1, otherwise VDC = 0. These values were then checked through a manual inspection of 50 random tokens, which confirmed that the calculated VDC measure largely corresponded to the presence or lack of VDC in the spectrogram.

To investigate speaker-level differences after controlling for known influences on VOT and VDC [7, 11, 29], a mixed-effects linear regression model of log-transformed VOT (following [29]) was fit using *lmerTest* [17] in *R* [22]. VDC was fit with a mixed-effects logistic regression using *lme4* [5]. The fixed-effects structure of both models contained predictors for phoneme **voicing**, speaker **birth year** and **gender**, preceding phoneme **manner**, presence of a preceding **pause**, following **vowel height**, **place** of articulation, speaking **style** (lectures vs. public speaking vs. conversational dialogues), **frequency** (log-transformed), phrase **position** (using Break Index value), and **speech rate** (speaker mean rate and deviation from speaker's mean). To further control for factors known to condition stop voicing, the models also included interactions for **voicing : place**, **voicing : gender**, **voicing : birth year**, **voicing : position**, and **voicing : local speaking rate**, and **voicing : position**. The VOT model also included an interaction for **voicing : mean speaking rate**, a cross-linguistically expected effect [14].

Continuous predictors (speaking rates, frequency) were centred and divided by two standard deviations. Two-level factors (voicing, gender, vowel height, position, pause) were scaled and centred at 0 as numerical predictors. Predictors with three or more levels (all others) were sum-coded. Each model was fit with the most comprehensive random-effects structure that would enable convergence [4], without random-effect correlations:

- VOT model: By-word random intercept; by-speaker random intercept and random slopes for voicing, place, local speaking rate, pause, fre-

quency, position, vowel height, manner, and voicing:{speaking rate, position, pause, manner, place} interactions.
- VDC model: By-word random intercept; by-speaker random intercept and slopes for voicing, pause, phrase position, and voicing:pause.

3. RESULTS

This study is interested in structured speaker-level variability in (1) the use of VOT and VDC cues to the Japanese stop voicing contrast, and (2) the relationship *between* VOT and VDC, as well as their patterning with speaker age. The speaker differences for each cue will be examined separately, followed by an exploration of how speakers modulate both cues together. Rather than reporting full regression tables for the two models, factors of interest are reported as the estimated marginal means, computed using *emmeans* [18]. These can be interpreted as the predictions made from each model (i.e., predicted VOT or VDC values) at each level of the factor of interest (e.g., voiced vs. voiceless stops) whilst holding other predictors at their average levels.

3.1. VOT

The directions and size of the fixed effects largely resemble those observed in previous studies of Japanese VOT [30, 23]. For example, VOT for voiced stops is approximately 13ms, compared to 36ms for voiceless stops ($\hat{\beta}$ = 0.999, $p < 0.001$). Speakers significantly differ in their overall use of VOT ($\chi^2(1) = 773.3$, $p < 0.001$) and the size of their VOT slopes ($\chi^2(9) = 1744.8$, $p < 0.001$), as assessed by likelihood ratio tests comparing models with and without random speaker intercepts and slopes respectively. The degree of interspeaker variability in log-transformed VOT is 0.155. This means speakers vary in their overall VOT value between 16ms and 30ms. Examining variability in the size of the VOT voicing contrast shows that speakers also vary: the voiced:voiceless VOT ratio ranges between 0.5 and 0.73. Crucially, this variability is *structured*: as Figure 1 (left) illustrates, the relationship *between* each speaker's voiced and voiceless VOTs is highly consistent. Speakers with long VOTs for voiceless stops also show long VOTs for voiced stops (Spearman's ρ = 0.461, $p < 0.001$). With respect to age-related differences in the use of positive VOT, a pairwise comparison of speaker birth year does not predict a significant change in the size of the contrast, though an increase in voiceless VOT between the youngest (1970-1979) and oldest (1930-39) speakers approaches significance ($\hat{\beta}$ = 0.236, $p = 0.057$), consistent with previous work [30].

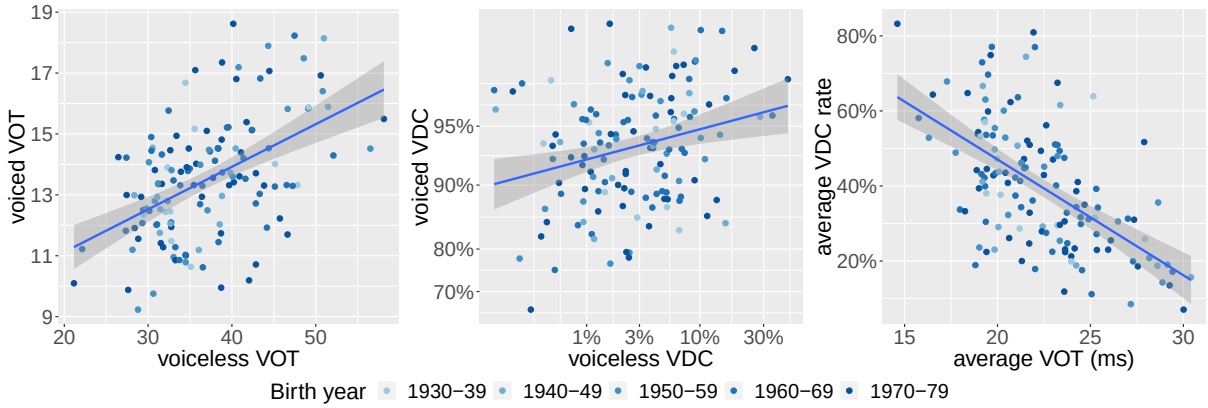
3.2. VDC

As expected for Japanese, where voiced stops typically show voicing lead, the probability of voicing during closure is more likely for voiced than voiceless stops ($\hat{\beta} = -6.192$, $p < 0.001$). Specifically, voiced and voiceless stops are predicted to exhibit very different VDC probabilities: voiceless stops show only 3% VDC, while that for voiced stops approaches ceiling (94%). As with VOT, speakers also differ significantly in their overall use of VDC ($\chi^2(1) = 1948$, $p < 0.001$). The degree of interspeaker variability in overall VDC log-odds (i.e., how often speakers produce VDC regardless of the voicing category) is 0.831, relative to a group-level VDC intercept of -0.437, meaning that speakers vary widely in their general VDC probability (range: 11% to 77%). Speakers also vary in the size of their VDC slope ($\chi^2(2) = 954.4$, $p < 0.001$), the voiced/voiceless difference in VDC, such that they are at least 5.7 more likely to show VDC for their voiced than voiceless stops. Figure 1 (centre) illustrates that VDC use in voiced and voiceless stops is signficantly related (ρ = 0.25, $p < 0.005$). There is no effect of age on average VDC, but a clear age-related pattern can be observed in the *size* of the voicing slope: the three youngest age groups each have significantly smaller VDC slopes compared to the 1930-39 group (1970-79: $\hat{\beta} = -1.261$, $p < 0.001$; 1960-69: $\hat{\beta} = -1.269$, $p < 0.001$; 1950-59: $\hat{\beta} = -0.965$, $p < 0.01$)[1]. This shows that younger speakers have a smaller difference in VDC between voiced and voiceless stops than older speakers.

3.3. VOT & VDC

The second research question concerns how speakers modulate VOT and VDC as cues to the voicing contrast. The two previously discussed models provide a statistical representation of how speakers differ in their realisation in the stop contrast along two dimensions for each cue: the *intercept* value (the overall use of VOT and VDC) and the *slope* (how much VOT/VDC differs between voiceless vs. voiced stops). One way of determining the relationship between the cues is to examine whether the cues are *correlated* with another in production. A relationship between cues *within* a single phonetic category might indicate that these parameters are intrinsically linked, perhaps sharing a single articulatory source [9]. A relationship *across* categories could enhance

Figure 1: Left: model-predicted voiceless vs. voiced VOT (ms) by speaker. Centre: model-predicted voiceless vs. voiced VDC (logit-scaled probability) by speaker. Right: model-predicted average VOT and VDC values by speaker. Lines/shading are linear smooths (95% confidence intervals); colours indicate speaker birth year.



the realisation of the contrast through the availability of multiple cues. Here, there is a strong negative relationship between the sizes of VOT and VDC intercepts across speakers ($\rho = -0.567$, $p < 0.001$). Figure 1 (right) illustrates this effect, which can be interpreted as both a within- and across-category relationship: the less aspirated a stop is, the more likely that stop is to be realised with VDC. Correlations between categories for both cues are all significant, indicating that this relationship holds across any pair of stops (Voiceless VOT, Voiceless VDC: $\rho = -0.413$, $p < 0.001$; Voiced VOT-Voiced VDC: $\rho = -0.36$, $p < 0.001$; Voiceless VOT-Voiced VDC: $\rho = -0.301$, $p < 0.005$; Voiced VOT-Voiceless VDC: $\rho = -0.427$, $p < 0.001$). Looking at the relationship across categories (i.e., the covariance relation in the voicing contrast), a null relationship is observed across speakers (Spearman's $\rho = -0.05$, $p = 0.562$), suggesting that the size of a speaker's VOT contrast does not predict their VDC contrast.

4. DISCUSSION

Stop contrasts–and their language-specific implementation–have been extensively studied, though only a handful of phonetic studies have focused on their realisation in spontaneous speech [3, 28, 29, 32, 27]. Here, multiple cues to a stop contrast undergoing sound change [31, 16] have been examined. Whilst the Japanese stop contrast has been well studied at the dialectical and generational level [30], there is less work on how this contrast is realised at the level of *individuals*. The questions here asked how speakers (1) vary in the use of acoustic parameters in stop realisation, and (2) modulate these cues to signal the contrast.

Speakers were found to vary from each other in their use of VOT and VDC, and this variability is highly *structured*. VOT values for voiced and voiceless stops were correlated across speakers: speakers with large voiceless VOTs also have large voiced VOTs, consistent with previous findings of other languages [7, 8, 12]. VDC also exhibited a similar degree of speaker structure, despite the wide range of speaker variability in the size of the VDC contrast. Younger speakers also had smaller VDC contrasts than older speakers, consistent with previous claims about this sound change [30, 16, 31]. Finally, a strong relationship between the *overall* use of VOT and VDC was observed, meaning that speakers with more aspirated stops are less likely to produce voicing during stop closures. This can be interpreted as a correlated difference in the use or lack of voicing used in the production of any stop regardless of the voicing category, and perhaps provides a baseline from which category-specific VOT/VDC values can be generated. One explanation for this observation is that both cues share an intrinsic articulatory link, possibly conditioned by relative oral pressure during the stop closure. However, such a covarying relation was not observed for the two cues in the voicing contrast. This suggests that these cues do not engage in a complementary or enhancement relationship for Japanese stops. An interesting next step in developing this work on structured variability in Japanese stops will be to include F0 with VOT and VDC.

Acknowledgements This work was funded by SSHRC #435-2017-0925 to MS. We thank M. Clayards for comments.

5. REFERENCES

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

[2] Bang, H. Y., Sonderegger, M., Kang, Y., Clayards, M., Yoon, T.-J. 2018. The emergence, progress, and impact of sound change in progress in Seoul Korean: implications for mechanisms of tonogenesis. *J. Phon.* 66, 120–144.

[3] Baran, J., Laufer, M., Daniloff, R. 1977. Phonological contrastivity in conversation: a comparative study of voice onset time. *J. Phon.* 5, 339–350.

[4] Barr, D. J., Levy, R., Sheepers, C., Tily, H. J. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* 68(3), 255–278.

[5] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1–48.

[6] Cho, T., Kim, S.-A. 2000. Domain-initial strengthening as enhancement of laryngeal features: Aerodynamic evidence from Korean. *UCLA Working Papers in Phonetics* 99, 57–70.

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

[8] Chodroff, E., 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.

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

[10] Davidson, L. 2016. Variability in the implementation of voicing in American English obstruents. *J. Phon* 54, 35–60.

[11] Docherty, G. 1992. *The timing of voicing in British English obstruents*. Berlin & New York: Foris.

[12] Hullebus, M. A., Tobin, S. J., Gafos, A. I. 2018. Speaker-specific structure in German voiceless stop voice onset times. *Proc. Interspeech 2018* Hyderabad. 1403–1407.

[13] Kawahara, S. 2015. Geminate devoicing in Japanese loanwords: Theoretical and experimental investigations. *Lang. Ling. Compass* 9, 181–195.

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

[15] Kikuchi, H., Maekawa, K. 2003. Performance of segmental and prosodic labeling of spontaneous speech. *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* Tokyo.

[16] Kong, E. J., Yoneyama, K., Beckman, M. E. 2014. Effects of a sound change in progress on gender-marking cues in Japanese. *Proc. LabPhon 14* Tokyo.

[17] Kuznetsova, A., Bruun Brockhoff, P., Haubo Bojesen Christensen, R. 2016. *lmerTest*. R package.

[18] Lenth, R. 2018. *emmeans*. R package version 1.2.3.

[19] Maekawa, K., Kikuchi, H. 2005. Corpus-based analysis of vowel devoicing in spontaneous Japanese: an interim report. In: van de Weijer, J., Nanjo, K., Nishihara, T., (eds), *Voicing in Japanese*. De Gruyter Mouton 205–228.

[20] Maekawa, K., Koiso, H., Furui, S., Isahara, H. 2000. Spontaneous speech corpus of Japanese. *Proc. 2nd LREC* volume 2 946–952.

[21] Maekawa, K., Koiso, H., Igarashi, Y., Venditti, J. 2002. X-JToBI: An extended J_ToBI for spontaneous speech. *Proc. 7th ICSLP* Denver. 1545–1548.

[22] R Core Team, 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

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

[24] Schultz, A. A., Francis, A. L., Llanos, F. 2012. Differential cue weighting in perception and production of consonant voicing. *J. Acoust. Soc. Am.* 132.

[25] Scobbie, J. 2006. Flexibility in the face of incompatible English VOT systems. In: Goldstein, L., Whalen, D., Best, C., (eds), *Laboratory Phonology 8*. Berlin: Mouton de Gruyter 367–392.

[26] Shimizu, K. 1996. *A cross-language study of the voicing contrasts of stop consonants in Asian languages*. Tokyo: Seibido.

[27] Smiljanic, R., Bradlow, A. R. 2008. Stability of temporal contrasts across speaking styles in English and Croatian. *J. Phon.* 36, 91–113.

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

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

[30] 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]*. Tokyo: Kurosio.

[31] Takada, M., Kong, E. J., Yoneyama, K., Beckman, M. E. 2015. Loss of prevoicing in Modern Japanese /g, d, b/. *Proc. 18th ICPhS* Glasgow.

[32] Yao, Y. 2009. Understanding VOT variation in spontaneous speech. *UC Berkeley Phonology Lab Annual Report* 29–43.

[1] Tukey-adjusted p-values for a pairwise comparison between a set of 5 estimates.