THE INFLUENCE OF PROSODIC CONTEXT ON HIGH VOWEL DEVOICING IN SPONTANEOUS JAPANESE

Oriana Kilbourn-Ceron

McGill University
oriana.kilbourn-ceron@mail.mcgill.ca

ABSTRACT

In the description of phonological processes which apply across word boundaries, the notion of "pause" plays an important role. Pauses may be part of the trigger for an alternation, or they may block it. Both of these have been claimed to hold of the process of high vowel devoicing (HVD) in Tokyo Japanese. How could pauses both condition and inhibit the same sound change process? The present study addresses this question by breaking down the notion of pause into two different components: the physical, silent pause, and the subjective measure of "boundary strength". Analysis of a large corpus of spontaneous Japanese [9] shows that both of these factors have a major effect on the application of HVD, and they are crucially independent of one another.

Keywords: speech corpora, spontaneous speech, vowel devoicing, Japanese, prosodic boundaries

1. INTRODUCTION

In the description of phonological processes which apply across word boundaries, the notion of "pause" plays an important role. For example, the well-known phenomenon of coronal stop deletion in English, which deletes /t/ or /d/ when it is the second member of a word-final consonant cluster, is described as applying at different rates depending on whether a consonant, vowel or *pause* follows [2]. Pauses may be part of the trigger for an alternation, or they may block it. Both of these have been claimed to hold of the process of high vowel devoicing (HVD) in Tokyo Japanese [5, 6]. How could pauses both condition and inhibit the same sound change process? The present study addresses this question by breaking down the notion of pause into two different components: the physical, silent pause, and the subjective measure of "boundary strength". Analysis of a large corpus of spontaneous Japanese [9] shows that both of these factors have major effects on the application of HVD, and they are crucially independent of one another.

1.1. Context

High Vowel Devoicing (HVD) is a process characteristic of many varieties of Japanese, including the Tokyo standard. HVD applies to the high vowels *i, u* when they are preceded by a voiceless obstruent, and followed by either another voiceless obstruent or a pause. For example, *kisha* "journalist" is pronounced as [ki̥ʃa], with the empty ring below the *i* indicating voicelessness.

This paper focuses on the effect of prosodic information on HVD in spontaneous, connected speech. Previous studies have suggested that vowels at morphological boundaries [12, 14] and before pauses [5, 3] are less suceptible to HVD, but these effects have not been systematically examined. The study presented fills that gap by analysing data from the Corpus of Spontaneous Japanese [9]. With 201 monologues of academic presentation speech and simulated public speaking, and including expert prosodic annotation, this corpus is particularly well suited for examining the effect of boundary strength and silent pauses.

Most descriptions of HVD imply that a pause and a voiceless obstruent are more or less equivalent as a conditioning environement, with the caveat that devoicing is "less likely" before pauses [13]. However, there is no consensus in the literature on exactly what type of pause creates the C̥_# HVD environment. Kondo [5] suggests that the relevant pauses are only those at the end of utterances, citing Maekawa's (1989) observation that utterance-medial pauses after an adjunct phrase or between conjoined sentences do not trigger devoicing of preceding high vowels. Vance [12] states that word-internal pauses, as in careful syllable-by-syllable pronunciations, actually block devoicing that would occur at normal speech rates. If silent pauses block devoicing, it seems clear that the 'following pause' that triggers devoicing cannot simply be any following silent pause — rather, it may be that these pauses are being confounded with the edge of a prosodic unit, such as an utterance edge.

2. DATA

The data for this study was drawn from the Corpus of Spontaneous Japanese [9], a corpus of audio recordings of academic presentation speech and simulated public speaking. All speakers in the corpus are speakers of Standard Japanese from Tokyo and surrounding areas [7]. This study will focus on a subset of the CSJ referred to as "the Core" which, in addition to being transcribed and tagged morphologically, includes sub-segmental labelling and also X-JToBI labels [8] which mark prosodic information. The Core contains about 45 hours of speech from 201 different speakers.

All tokens of high vowel tokens were extracted, but for analysis were restricted to those which met the following criteria: the segments preceding and following the tokens are voiceless consonants (regardless of any intervening pauses). The token was not followed or preceded by another potentially devoiceable vowel, that is, the token is in the single devoicing environment. Finally, all tokens which were part of word fragments, mispronunciations or disfluencies (as annotated in the corpus) were excluded. The total number of token which met all these criteria was 37 338. All further discussion of the data refers only to this subset, unless otherwise noted.

The dependent variable for this study is the binary outcome of devoicing or no devoicing. This information was drawn directly from the segmental annotation in the CSJ, where it was determined by the human labellers preparing the corpus using information from "the wide-band spectrogram, speech waveform, extracted speech fundamental frequency, peak value of the autocorrelation function, in addition to audio playback." [7, :208].

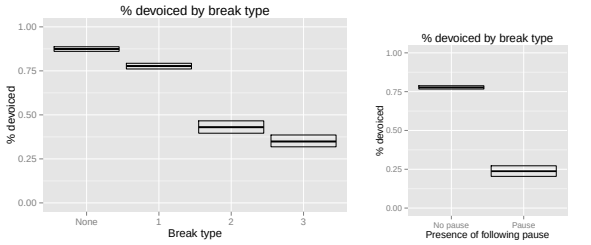
2.1. Empirical trends

Overall, the devoicing rates were 91.5% for /i/-tokens and 88.2% for /u/-tokens, slightly higher rates than those reported in [7], which were 89.15% and 84.25% respectively. However, the mean when averaged by lexical type is lower, at 81.7% for /i/ and 81.5% for /u/. This discrepancy may be due to some high-frequency lexical items which idiosyncratically more prone to devoice and therefore inflate the mean devoicing rate calculated for tokens. Many previous works have noted that the highly frequent morphemes *masu* "[polite non-past]" and *desu* "[polite copula]" are devoiced at a much higher rate than other lexical items [11, 5]. The use of a mixed-effects model for the present study will allow us to control for varying inter-word devoicing rates by fitting a random intercept for each word. In light of this, all the following figures in this section plot mean devoicing rates averaged by lexical type rather than by token.

One of the major gaps in the literature on high vowel devoicing is the lack of investigation into the effects of boundaries within the HVD environment. As discussed in §1, there have been a number of scattered observations suggesting that there is more variability when target vowels are at word edges, but the source and structure of this variability has not been investigated. Factors such as morpheme boundaries, utterance edges, and utterance-internal pauses have been noted to affect devoicing, but there has been no study that focuses on differentiating between prosodic factors which could be at play in these environments.

This study tracks two different variables potentially corresponding to "pause": the presence of a following silent pause of at least 200 ms, and the strength of the prosodic boundary between the potential target and the following triggering segment. The data have been labelled according to the strength of the X-JToBI [8] break index that follows the token (*None, 1, 2* or *3*), and whether or not there was a following silent pause (*No pause* vs *Pause*). The empirical rates of devoicing within these categories is shown in Figure 1.

Figure 1: Relationship between break type and mean devoicing percentage by lexical type.



	Mean % devoiced	No. of lexical types	No. of tokens
None	89.8	1 806	13 880
BI 1	83.2	1 428	19 205
BI 2	44.6	415	1 819
BI 3	36.9	370	2 554
No pause	77.7	5 224	35 441
Pause	23.7	546	2 018

The rate of devoicing is the highest, at 89.8% when the break type is *None*, that is, when the token is morpheme-internal. The rate at `breakType` = *1*, morpheme boundaries with no following pause, is only slightly lower at 83.2%. Devoicing rates at breaks *2* and *3* were lower , at 44.6% and 36.9% respectively. The effect of a following silent pause looks quite significant, going from 77.7% for tokens with no following pause to 23.7% when a pause is present.

The fact that devoicing rates are so much higher

when a token is word-internal or at a phrase-internal morpheme boundary immediately suggests that prosodic boundaries, represented here by the X-JToBI break indices, play a role in determining the application of HVD. However, following silent pauses also appear to have a strong inhibiting effect – if these are mostly associated with higher break indices, it may be pauses that are responsible for the apparent effect of break indices. Whereas previous literature has not explicitly distinguished between these two factors, the statistical model in this study will address track these two variables separately and be able to determine whether the influence of each is independently significant once other factors, including by-word variability, are accounted for.

3. MODEL

The mixed-effects logistic regression model was fit using R [10] and `lme4` package [1]. In order to examine the two separate aspects of "pause", two different variables were tracked: whether or not the token was followed by a silent pause of at least 200 ms, and the strength of the prosodic boundary (if any) between the token and the following segment (a voiceless consonant in all cases). Since there were no cases of word-internal pauses, these two variables were combined into a single predictor, `breakType`, for the purpose of fitting the statistical model.

Based on previous findings, the following predictors were also included: manner of previous and following consonants, presence of high tone associated with target vowel, quality of vowel ([i] or [u]), and speaking rate (mean of each speaker, and the local SR relative to the speakers' mean). These are included as controls, and will not be analysed in detail in this paper.

In addition to the fixed effects described above, the model included random effects for two groups: lexical types, of which there were 3211, and speakers, of which there were 201. Each member within these groups had fitted a random intercept, and a random slope for the effect of high tone presence.

3.1. Results

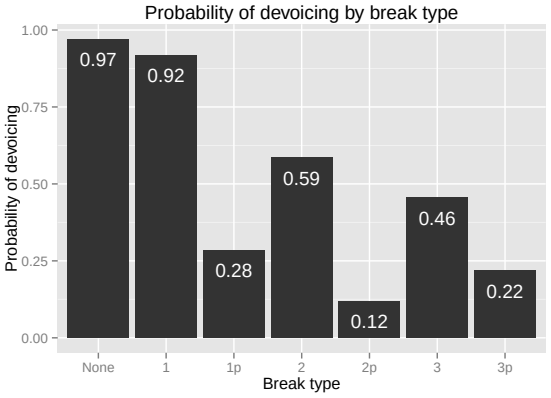
Table 1 shows the model's estimates for the fixed effects. All of the effects have estimates which are significantly different from 0 ($p < 0.001$), except for `logFreq` which is significant at $p = 0.01$, and `vowel`, which is not significant ($p > 0.05$). The probability of devoicing with all predictors held at their mean values is predicted to be 56.7%.

Figure 2 shows that the devoicing rate for the *None* condition, where there is no prosodic break

Table 1: Model output — Fixed effects (Intercept omitted)

	Estimate	Std. Error	z value	Pr(>|z|)
breakType = *1 vs None*	-0.53	0.05	-9.92	<0.001
= *1p vs prev. levels*	-1.29	0.08	-16.20	<0.001
= *2 vs prev. levels*	-0.33	0.03	-10.40	<0.001
= *2p vs prev. levels*	-0.67	0.05	-12.45	<0.001
= *3 vs prev. levels*	-0.14	0.02	-6.79	<0.001
=*3p vs prev. levels*	-0.26	0.02	-12.64	<0.001
prevManner = *stop vs afr*	0.27	0.07	3.88	<0.001
= *fric vs prev. levels*	0.18	0.04	4.94	<0.001
follManner = *stop vs afr*	-0.30	0.08	-3.94	<0.001
= *fric vs prev. levels*	-0.39	0.03	-12.12	<0.001
= *gem vs prev. levels*	-0.66	0.04	-17.70	<0.001
highTone	-3.63	0.15	-24.74	<0.001
vowel	0.05	0.10	0.45	0.65
meanSR	0.35	0.11	3.14	<0.001
srDev	0.35	0.06	5.91	<0.001
logFreq	0.43	0.16	2.62	0.01

Figure 2: Predicted probability of devoicing at each break type with all other predictors held at mean values.



within the HVD environment, vowels are predicted to devoice in almost 97% of cases, all other things being equal. Vowels followed by a prosodic break index of *1* but with no following pause are also predicted to devoice at a high rate, 92%. For break indices *2* and *3*, the devoicing rates are predicted to be 59% and 46% respectively. However, when there is a physical pause between the vowel and the following voiceless consonant, predicted devoicing rates are much lower, between 28% and 12%.

Post-hoc tests were used to determine if the differences in predicted devoicing rates in Figure 2 are indeed significant differences in log-odds of devoicing in the model. The tests determined whether tokens that were followed by a silent pause (`breakType` = 1p, 2p, 3p) differed significantly from tokens that were not followed by a pause (`breakType` = 1, 2, 3, N), and whether within each break type, the presence of a pause had a significant effect (1p vs 1; 2p vs 2; 3p vs 3). They also tested whether there was a significant difference in the estimates between tokens which were followed by no pause, but different

Table 2: Results of post-hoc tests for `breakType`. Negative estimates indicate that probability is lower for the first group than the second group.

Linear Hypotheses:

	Estimate	Std. Error	z value	Pr(>|z|)	
(1p, 2p, 3p) - (1, 2, 3, N)	-10.0044	0.4490	-22.281	<0.001	***
1p - 1	-3.3433	0.2382	-14.036	<0.001	***
2p - 2	-2.3459	0.2707	-8.666	<0.001	***
3p - 3	-1.0975	0.1559	-7.041	<0.001	***
2 - 1	-2.0758	0.1077	-19.270	< 0.001	***
3 - 2	-0.5223	0.1245	-4.196	< 0.001	***

break indices (2 vs 1; 3 vs 2). These hypotheses were tested with a general linear hypothesis test, using `glht` in the `multcomp` package [4] in R. Table 2 reports the estimated difference (in log-odds of deletion rate) and p-value for each hypothesis.

All of these comparisons were found to have significantly different estimates ($p < 0.001$). The first hypothesis test in Table 2 confirms that the estimate for all tokens with a following pause is much lower than for tokens with no following pause by a very large margin, confirming the empirical trend seen in Figure 1.

The next three comparisons show that the inhibitory effect of a following pause is greatest when the break index is 1, and that its magnitude decreases as the value of the break index, i.e. prosodic boundary strength, increases. Put another way, the greater the strength of the prosodic boundary, the less inhibitory the effect of a physical pause becomes.

The last two lines in Table 2 show that even when no pause is present, there is still a significant difference between consecutive values of the break indices. The difference between BI 1 and no break index (word-internal context) is already represented in the first row of Table 1, with an estimated log-odds difference of -0.53. This means that devoicing is only 0.59 times as likely to occur if a BI 1 follows than in the word-internal case. The estimated difference between BI 2 and BI 1 tokens is of greater magnitude, with the odds differing by a factor of 0.125. The difference between a following BI 3 and BI 2 is comparable to BI 1 vs no BI, with a difference in factor of 0.593.

In other words, among tokens with no following pause, the tokens that are followed by BI 1 are about 1.7 times less likely to be devoiced than those which are word-internal. Those with a following BI of 2 are 8 times less likely to be devoiced than those followed by BI 1, and tokens followed by BI 3 are 1.7 times less likely to be devoiced than the BI 2 tokens.

In sum, the model indicates that both physical pauses as well as break indices are very good predictors of devoicing probability. The presence of a pause was found to significantly decrease the overall probability of devoicing, and this inhibitory effect was also found to be significant within each value of break index (1, 2 or 3). When there was no following pause, the difference between each successive value of break index was also found to differ significantly. In other words, the probability differences illustrated in Figure 2 are almost all confirmed to be significant by these post-hoc tests.

4. DISCUSSION

The statistical model fit to this spontaneous speech data confirms that, all else being equal, devoicing is significantly less likely when the vowel is followed by a *physical pause*. However, the *strength of the boundary* was also shown to affected devoicing probability independently.

Something that has not been previously noted in the literature on HVD is that the blocking effect of pauses differs depending on the strength of the prosodic boundary that the pause co-occurs with. The inhibitory effect of a physical pause is strongest BI 1, where it reduces the odds of devoicing by 28 times. By contrast, at BI 2 the presence of a pause reduces the odds by about 10 times, and at BI 3 by only 3 times. Hence, the effect of pause is cumulative, but not strictly additive.

5. CONCLUSIONS

Statistical analysis of high vowel devoicing in the Corpus of Spontanous Japanese has revealed that (at least) two different ways of defining a "pause" are relevant for this process: silent pauses and prosodic boundary strength. Future work will include more detailed measures for both of these predictors, including pause duration and some concrete acoustic cues as correlates of boundary strength. Such work could help define cases where abstract prosodic structure must be posited to define the domain of a process, or where such domains are best described by reference to more phonetic contexts such as silent pauses.

6. REFERENCES

[1] Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. *lme4: Linear mixed-effects models using Eigen and S4*, 2014. R package version 1.1-7.

[2] Andries W Coetzee and Joe Pater. The place of variation in phonological theory. In Alan C. L. Yu John A. Goldsmith, Jason Riggle, editor, *The Handbook of Phonological Theory, Second Edition*, pages 401–434. Wiley-Blackwell, 2011.

[3] Manami Hirayama. *Postlexical Prosodic Structure and Vowel Devoicing in Japanese*. PhD thesis, University of Toronto, 2009.

[4] Torsten Hothorn, Frank Bretz, and Peter Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008.

[5] Mariko Kondo. *Mechanisms of vowel devoicing in Japanese*. PhD thesis, University of Edinburgh, 1997.

[6] Laurence Labrune. *The phonology of Japanese*. Oxford University Press, 2012.

[7] Kikuo Maekawa and Hideaki Kikuchi. Corpus-based analysis of vowel devoicing in spontaneous Japanese: an interim report. In Jeroen van de Weijer, Kensuke Nanjo, and Tetsuo Nishihara, editors, *Voicing in Japanese*, number 84 in Studies in generative grammar, pages 205–228. Berlin: Mouton de Gruyter, 2005.

[8] Kikuo Maekawa, Hideaki Kikuchi, Yosuke Igarashi, and Jennifer J Venditti. X-JToBI: an extended J-ToBI for spontaneous speech. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP2002)*, pages 1545–1548. Denver, 2002.

[9] Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. Spontaneous speech corpus of Japanese. In *Proceedings of the Second International Conference of Language Resources and Evaluation (LREC)*, volume 2, pages 947–952, 2000.

[10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[11] Timothy J Vance. *An introduction to Japanese phonology*. Albany: State University of New York Press, 1987.

[12] Timothy J Vance. Lexical phonology and Japanese vowel devoicing. In Gary N Larson, Lynn A MacLeod, James D McCawley, and Diane Brentari, editors, *The joy of grammar: a festschrift in honor of James D. McCawley*, pages 337–350. Amsterdam: J. Benjamins Pub. Co., 1992.

[13] Timothy J Vance. *The Sounds of Japanese*. Cambridge University Press, 2008.

[14] John Kevin Varden. *On high vowel devoicing in standard modern Japanese: implications for current phonological theory*. PhD thesis, University of Washington, 1998.