Acoustic and articulatory correlates of Japanese devoiced vowels

Marco Fonseca[1], Maria Cantoni[2], Thaïs Cristófaro-Silva[2]

The University of Tokyo, Federal University of Minas Gerais
marcofon@illinois.edu, mmcantoni@gmail.com, thaiscristofaro@gmail.com

ABSTRACT

In Japanese, high vowels may devoice between two unvoiced obstruents (*k[ɯ̥]kaku* 'division', *k[i̥]kaku* 'plan'). Recent studies have shown that the devoicing of non-high vowels (*k[ḁ]karu* 'take', *k[e̥]ta* 'digit, *k[o̥]tae* 'answer') may also occur. This paper evaluates the role of vowel duration and vocal folds gestures involved in vowel devoicing in Japanese. An experiment using an electroglottograph (EGG) was conducted, and the results showed different duration values and different vocal folds gestural patterns, depending on vowel quality. However, different duration values were not observed for partially devoiced vowels. Therefore, it is argued that Japanese vowel devoicing is a gradual phenomenon involving the reduction of time and glottal gesture magnitude. Additionally, it is suggested that Japanese devoiced vowels emerge as a targetless vowel that cannot be characterized in terms of duration and vocal folds gestures.

Keywords: Vowel devoicing, electroglottography, vowel duration, vocal folds, gesture.

1. INTRODUCTION

Japanese vowel devoicing has been extensively studied under several phonetic and phonological approaches [1, 2]. It is typically assumed that a devoiced vowel is deleted. This paper suggests that Japanese vowel devoicing is a gradient phenomenon involving temporal reduction and glottal gestures changing instead. Most of the previous research considers only the devoicing of high vowels. However, more recent research shows that non-high vowels ([a], [e], [o]) may also be devoiced, such as *k[ḁ]karu* 'take', *k[e̥]ta* 'digit, and *k[o̥]tae* 'answer' [3, 4, 5].

Results gathered from experimental data have shown that high vowels are shorter than non-high vowels [6] and that devoiced vowels are shorter than voiced vowels [2]. This paper takes previous studies further by taking into account the duration of both high and non-high vowels. This study also advances on previous research by employing an EGG in the analysis in order to evaluate vowel duration and vocal folds gestures configurations in both voiced and devoiced vowels. EGG is a technique in which two electrodes are applied to the subject's neck around the location of the vocal folds. Its main appeal is the fact that it is relatively non-expensive, easy to use, and non-invasive [7]. The signal of the EGG represents the variation of the area of contact of the vocal folds over time and it suffers little influence from vocal tract resonance. Thus, EGG data offer reliable data regarding vocal folds contact which is a known correlate of voicing.

2. METHODS

2.1. Participants

Twenty-six undergraduate students, ranging from 18 to 21 years old, native speakers of Tokyo (standard) Japanese took part in the experiment. In order to obtain the best quality EGG signal, only male participants with little adipose tissue or hair on their neck were recorded [7].

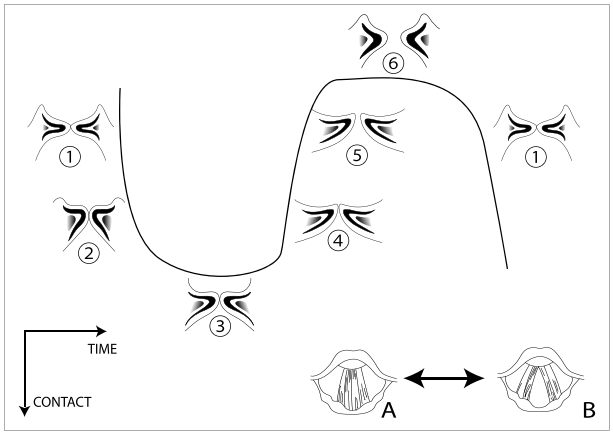
2.2. Stimuli

The stimuli words were inserted in the following carrier-sentence: *watashi wa ___ to iimashita* 'I said ___'. The subjects were asked to read the sentences at a normal speed as naturally as possible. In this experiment, a total of 136 words were presented to the subjects. Forty words correspond to the target words analyzed in this paper. They are comprised mainly of words containing three or four morae. Only the vowel located in the first mora, which is unaccented, was considered for this analysis. Overall, 1040 tokens of the target vowels were obtained (40 tokens x 26 subjects) for the analysis.

2.3. Processing of the EGG signal

The EGG signal represents the variation of the area of contact of the vocal folds over time. Figure 1 is a schematic representation of the EGG signal. Locations (1) and (2) represent the closing movement of the vocal folds. The valley of the wave (3) represents vocal fold contact. Locations (4) and (5) represent the opening movement while (6), the peak of the wave, represents the vocal fold opening.

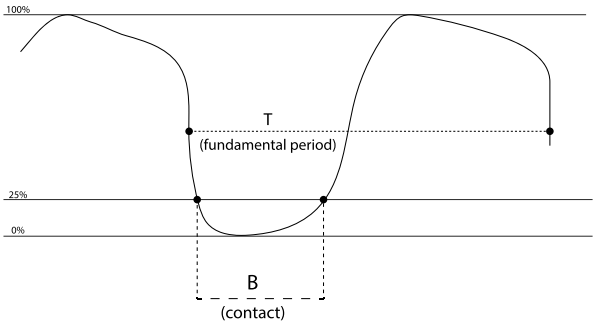
Figure 1: Schematic representation of the EGG signal.

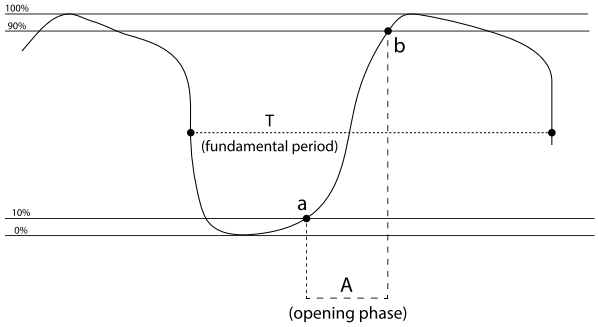


The EGG signal may fluctuate due to movements of the neck, head, or electrodes, or due to the influence of the electrical grid [7]. Thus, a high-pass Hann filter (cut off frequency: 60 Hz, smoothing: 5 Hz) was applied to the EGG signal to eliminate these fluctuations, using the software Praat [8] stop band filter function. After filtering the EGG signal, the target vowel of each one of the 1040 tokens was manually segmented and annotated through Praat's TextGrid tool.

Two articulatory measures suggested by [7] for the EGG signal were adopted in this paper: the contact quotient (CQ) and the opening phase (OP). Consider the following reference points of the EGG signal based on [7]:

Figure 2: Reference points used in the glottal measurement from the EGG signal.





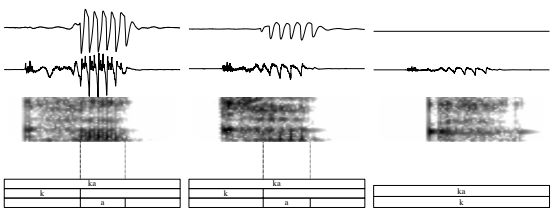
T represents the fundamental period, A is the opening phase of the vocal folds, and B their contact. The CQ and the OP are calculated as follows:

$$(1)\ CQ = \frac{B}{T} \times 100\% \qquad (2)\ OP = \frac{A}{T} \times 100\%$$

2.4. Classification of the data

The 1040 target vowels obtained in the experiment were classified as voiced, partially devoiced, or fully devoiced. This categorical classification is in accordance with a gestural approach to Japanese vowel devoicing, since an "increase in overlap among gestures in fluent speech is a general gradient process that can produce apparent (perceived) discrete alternations" [9, p. 39]. Voiced vowels were identified by a perturbation of the EGG signal (since vocal cord vibration occurs), a visible perturbation in the sound wave of the oscillogram, and stable first and second formants. For partially devoiced vowels, a smaller displacement of the amplitude of the EGG signal and of the oscillogram signal was observed. For fully devoiced vowels, there was a flat EGG signal (since no vocal fold vibration occurs). The oscillogram of a fully devoiced vowel corresponds to the articulation of its preceding sound. For both partially and fully devoiced vowels, the steady first and second formants observed in fully voiced vowels are not present. Figure 3 shows, from top to bottom, the EGG signal, oscillogram, spectrogram, and the annotations of the first mora of the word ***ka****karu* (take).

Figure 3: Comparison between voiced, partially devoiced and devoiced vowels.



The distribution of the experimental data according to the classification proposed in this section for vowel voicing quality is illustrated in Table 1.

Table 1: Occurrences of vowel devoicing for vowel quality.

	a	e	i	o	u
Voiced	188 (90%)	194 (93%)	107 (51%)	188 (90%)	56% (116)
Partially devoiced	18 (9%)	14 (7%)	4 (2%)	16 (8%)	4 (2%)
Devoiced	2 (1%)	0 (0%)	97 (47%)	4 (2%)	88 (42%)
Overall	208 (100%)	208 (100%)	208 (100%)	208 (100%)	208 (100%)

Table 1 shows that the high vowels [i] and [u] present higher rates for devoicing when compared to [a], [e], [o].

3. RESULTS

3.1. Duration values

The script Duration multiple [10] was used to obtain duration values for the vowels segmented through the TextGrid. Since it was not possible to segment fully devoiced vowels, the analysis of the duration values considered only fully voiced and partially voiced vowels. Furthermore, the item *kega* 'injury' (subject 2) was discarded as the quality of the signal obtained for it was not satisfactory. Overall, 848 items were evaluated in this analysis.

In order to analyze the relationship between the duration of the target vowel with the voicing quality and vowel type, the linear mixed effect model [11] illustrated in (2) was adopted. R [12] was employed in the analysis so the model is presented in R syntax. This model has vowel duration as the dependent variable. Voicing quality (two levels: voiced versus partially devoiced) and vowel type (five levels: a, e, i, o, and u) were considered as independent variables. The next and previous sounds of the target vowel were also added as independent variables in the model. The effects of these two predictors will not be discussed in this paper. Speaker and item were considered as random effects in the model.

(2) Duration~Voicing*Vowel+Previous+Next + (1|Speaker) + (1|Item)

A likelihood-ratio test was applied to the model. Considering a significance level of α=5%, the test showed that the predictor Voicing is statistically significant ($\chi^2(1)=100.33$, $p < 0.001$) and that the duration of partially devoiced vowels is approximately 10.18 m.s. (standard error: 1.49) shorter than the duration of voiced vowels. The predictor Vowel was also significant ($\chi^2(4)=61.42$, $p < 0.001$). The slope estimates for high vowels (-10.52 for [i] and -16.52 for [u]) have greater magnitude than the estimates for non-high vowels (-2.66 for [e] and -1.88 for [o]), with the intercept corresponding to the duration of [a]. This result indicates that high vowels are significantly shorter than non-high vowels. However, the interaction between voicing quality and vowel type did not show significance ($\chi^2(4)=2.08$, $p=0.72$). These results show that there are no statistically significant differences between the duration of devoiced vowels.[1]

3.2. Articulatory analysis

The articulatory analysis, expressed through the CQ and OP measures of the EGG signal, considered only non-high vowels. From the initial 624 tokens, full vowel devoicing was observed for 6 vowels and therefore 618 tokens were left from the initial analysis. The item *kega* 'injury' (subject 2) was also excluded from the articulatory analysis. Therefore, the final analysis for the articulatory measures considered 617 tokens.

3.2.1. The contact quotient

In order to evaluate the effects of vowel type and voicing quality in the CQ, the linear mixed effect model expressed in (3) was employed. The model described in (3) has the CQ as the dependent variable, and the interaction between voicing quality and vowel type as independent variables. The previous and next sounds were also added as fixed effects but their effects are not the focus of this paper.

(3) *CQ*~Voicing*Vowel+Previous+Next + (1|Speaker) + (1|Item)

Considering a significance level of α=5%, the results of the likelihood-ratio test for the CQ showed that voicing quality affected CQ ($\chi^2(1)=32.19$, $p < 0.001$). The CQ for partially voiced vowels was lowered by approximately 2.72% (standard error: 0.81) when compared to voiced vowels. Vowel quality also affected the CQ ($\chi^2(1)=20.01$, $p < 0.001$). The CQ was raised approximately 0.74% (standard error: 0.36) for the vowel [e] and approximately 1.47% (standard error: 0.36) for the vowel [o] in comparison to vowel [a]. The interaction between voicing quality and vowel type did not show statistical significance ($\chi^2(2)=5.6$, $p > 0.5$). These results show that there are no statistically significant differences between the CQ of partially devoiced vowels.

3.1.2. The opening phase

In order to analyze the relationship between the OP of the target vowel with voicing quality and vowel type, the linear mixed effect model illustrated in (4) was adopted. Voicing quality and vowel type were considered as independent variables. The next and previous sounds of the target vowel were also added as independent variables in the model (whose effects will not be discussed in this paper).

(4) OP~Voicing*Vowel+Previous+Next+(1|Speaker) + (1|Item)

A likelihood-ratio test was applied to the model. The results showed that the voicing quality affected OP ($\chi^2(1)=13.9033$, $p < 0.001$), such that when the vowel was partially devoiced its OP was shortened approximately by 3.47% (standard error: 1.22). Neither the predictor vowel ($\chi^2(2)=0.29$, $p > 0.5$) nor its interaction with voicing showed statistical significance (χ^2 $(2)=1.85$, $p > 0.5$). This indicates that the OP does not vary depending on vowel type, regardless of voicing quality.

4. DISCUSSION

This paper aimed to evaluate the timing and glottal gesture configurations for Japanese vowel devoicing. The duration of both voiced and devoiced vowels and their interaction with vowel quality were analyzed. The results showed that devoiced vowels are shorter than voiced vowels. This is in accordance with previous research [2, 3].

This paper took previous analyses further by considering the duration of devoiced high and non-high vowels and employing an EGG in the analysis. It is known that voiced high vowels are shorter than non-high vowels [6], and this tendency was observed in the data collected in this study. However, no significant difference was observed for the duration of partially devoiced high and non-high vowels. This suggests that Japanese devoiced vowels are merging towards a targetless vowel [13]. The schwa in English is targetless in the sense that it is not specified for its vocal tract dimension. More specifically, tongue gestures unfolding through time cannot characterize the schwa in English. Similarly, vowel duration in Japanese does not differentiate partially devoiced vowels and this is further evidence for its targetlessness. This is because gestures and timing are closely related. When there is a time reduction, one can argue that the articulators have less time to reach their targets and perform a full, targeted gesture. Glottal gesture configurations, expressed through the CQ and the OP were also evaluated. The results of the EGG experiment showed that the CQ and the OP are lower for partially devoiced vowels when compared to fully voiced vowels. Intensity can be controlled by the medial compression of the vocal folds [14]. In this case, intensity lowering can correspond to a lower CQ, because there is less contact of the vocal folds, and a lower OP, because the opening occurs at a higher speed. The CP and OP significantly lower values for partially devoiced vowels could be related to a decrease in intensity. This decrease in intensity is corroborated by the literature [2].

Furthermore, different CQ values for vowels [a], [e], and [o] were observed. However, these differences were not found for partially devoiced vowels. It is argued that this is also evidence for the targetlessness of Japanese devoiced vowels. Similarly to the schwa in English, which cannot be characterized for its tongue gesture unfolding through time, Japanese devoiced vowels also cannot be characterized for their glottal gestures as they unfold through time.

In summary, this paper showed that Japanese vowel devoicing is a phenomenon that involves gradual temporal reduction. This temporal reduction is accompanied by the reduction of vocal fold gestures expressed by the CQ and the OP. The outcome of this reduction is a targetless vowel.

5. REFERENCES

[1] McCawley, J. D. 1968. *The Phonological Component of a Grammar of Japanese*. The Hague: Mounton.

[2] Kondo, M. 2005. Syllable structure and its acoustic effects on vowels in devoicing environments. In: Weijer, van de J. Nanjo, K. Nishihara, T. (eds.). *Voicing in Japanese.* Berlin; New York: Mouton de Gruyter. 229-245.

[3] Maekawa, K. 1993. Boin no museika (Vowel devoicing). In: Miyaji, S. et al. (eds.). *Kouza no nihongo to nihongo no kyouiku, Nihongo no onsei – on'in*. (Japanese course and Japanese education – Japanese phonetics/phonology). Tokyo: Meiji Shoin. 135-153.
[4] Maekawa, K. Kikuchi, H. 2005. Corpus-based analysis of vowel devoicing in spontaneous Japanese: an interim report. In: Weijer, van de J. Nanjo, K. Nishihara, T. (eds.). *Voicing in Japanese.* Berlin; New York: Mouton de Gruyter. 205-228.
[5] Labrune, L. 2012. *The Phonology of Japanese*. Oxford: Oxford University Press.
[6] Beckman, M. 1996. When is a syllable not a syllable? In: Otake, T. Cutler, A. (eds.). *Phonological Structure and Language Processing*. Berlin; New York: Mouton de Gruyter. 95-124.
[7] Vieira, M. 1997. *Automated measures of dysphonias and the phonatory effects of asymmetries in the posterior larynx*. Doctoral Thesis. University of Edinburgh, Edinburgh.
[8] Boersma, P. & Weenink, D. 2014. *Praat*: doing phonetics by computer. Computer program. Version 5.3.80. Available at: <http://www.praat.org/>.
[9] Browman, C. Goldstein, L. 1992. Articulatory Phonology: An overview. In: *Phonetica*, 49(3-4). 155-180.
[10] Arantes, P. 2008. *Duration Multiple.* Praat script. Available at: <https://code.google.com/p/praat-tools/>.
[11] Bates, D. M. Maechler, M. Bolker, B. 2012. *lme4: Linear mixed-effects models using S4 classes*. R package version0.999999-0.
[12] R Core Team 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Available at: <http://www.R project.org/>.
[13] Browman, L. Goldstein, C. 1990. "Targetless" schwa: an articulatory analysis. In: *Haskins Laboratories Status*. Report on Speech Research. SR-101/102. 194-219.
[14] Fant, G. 1981. The source filter concept in voice production. In: *Speech, Music and Hearing - Quarterly Progress and Status Report*. 22. 21-37.

[1] Professor Yuki Hirose and Kato Tsuneaki from the University of Tokyo suggested that this result may indicate a different trend. It is intended to evaluate other interpretations of the statistical results in further studies.

The authors gratefully acknowledge support from the Ministry of Education, Culture, Sports, Science and Technology of Japan, The Federal University of Minas Gerais, CNPq grant 30.65.95/2011-7, and FAPEMIG-CAPES PACCSS II 015/2013.
From August 2015 Marco Fonseca will be associated with the University of Illinois at Urbana-Champaign.