Notes on So-called Inter-speaker Difference in Spontaneous Speech: The Case of Japanese Voiced Obstruent

Kikuo Maekawa
Dept. Corpus Studies, National Institute for Japanese Language and Linguistics, Tokyo
kikuo@ninjal.ac.jp

Abstract
In spontaneous speech, seemingly ‘inter-speaker’ variation is not thoroughly speaker-dependent. It is often the cases that, in addition to the genuine speaker-dependent factors, contextual factors like speaking rate and linguistic environment also have their influence on the variation. In this paper, the variations in the vocal-tract closure articulation of Japanese voiced obstruent consonants are reanalyzed in view of the separation between speaker-dependent and contextual factors. The consonants in the Corpus of Spontaneous Japanese were analyzed by means of logistic regression using the realization of vocal-tract closure as the binary dependent variable and the TACA (time allotted for consonant articulation), a recently proposed articulatory parameter, as the independent variable. Thresholds of TACA for the realization of vocal-tract closure were computed for all speakers. The results revealed clear inverse correlation between the threshold and the mean realization rate of closure. To explain the perturbation in the correlation pattern, influence of skewed distribution of phonemes and pauses were evaluated. The results showed the expected patterns in /z/ and /b/, but not in /d/ and /g/.

Index Terms: spontaneous speech, inter-speaker difference, Corpus of Spontaneous Japanese (CSJ), voiced obstruent

1. Introduction
Most study of speech productions confront with inter-speaker difference of some sort [1]. Known sources of inter-speaker variability include anatomical difference of vocal tract [2-4], differences in motor control and speech planning [1, 5-8], and, language-external contextual factors like speaking rate [9,10]. It is therefore necessary to control the factors in the experiments that deal with the inter-speaker variability. On the other hand, there is a wide consensus among linguists that it is in so-called spontaneous speech that speech variability including the inter-speaker variability can be observed most typically [11, 12]. Here is the situation where the needs for experimental control and spontaneity are incompatible.

It is probably impossible to solve the incompatibility problem completely, but at the same time, there must be better ways of investigating inter-speaker difference in spontaneous speech than what we do today, i.e. comparison of simple means computed for each speaker. In the rest of this paper, it will be shown that in the case of vocal-tract closure articulation, it is possible, at least to some extent, to dissect the variation into genuine inter-speaker difference in the closure articulation and the contextual variation. Such separation is possible given there is a simple, preferably single, phonetic parameter behind the variation in question, and, the relationship between the parameter and articulation is understood clearly enough. As example of such variation with considerable inter-speaker difference, the weakening of closure articulation in Japanese voiced obstruent (affricates and stops) was analyzed using spontaneous corpus.

2. Method and data
2.1. Background
Japanese voiced obstruent shows variation with respect to its manner of articulation. Realization of /z/ in Standard Japanese alternates between affricate [dz] and fricative [z]. Similarly, voiced stops alternate between stops and fricatives, namely, [b] alternates between [b] and [β], /d/ between [d] and [ð], and /g/ between [g] and [ɣ] (for [ŋ], see the text about Fig.2). These alternations are all concerned with the weakening of vocal-tract closure articulation.

Literatures of the Japanese phonetics often describe these variations as allophonic variation, viz., stop (affricate) in word-initial position, fricative elsewhere [13,14], following, probably, the description in Shiro Hattori’s classical work [15]. Recently, however, it turned out by the statistical analysis of spontaneous speech, that the weakening could better be interpreted as coarticulatory variation [16, 17]. In this new account, the key parameter of the alternation is the quantity called TACA, or time allotted for consonantal articulation.

2.2. TACA and closure rate
TACA is a simple function of the durations of the obstruent in question and its preceding segments. If the obstruent is immediately preceded by a geminate /Q/ or a moraic nasal /N/, the duration of the preceding segment is included in the TACA; it is because in /Q/ and /N/ the articulation gestures are fully underspecified and are completely determined by the assimilation to the following segments (i.e., the obstruent in question). As a result, the sequence /Qx/ or /Nx/ (where x stands for an obstruent) functions virtually as a single long consonant with respect to the closure articulation. Similarly, the duration of immediately preceding pause is included in TACA, because speaker can utilize the time of the pause for the realization of closure. In the computation convention of TACA proposed in [16], the maximum increment of TACA due to preceding pause is limited to the duration of the target obstruent, in order to avoid serious side-effect of a very long pause (Fig.1). The present study follows this convention.

Figure 1: Schematic example of the computation of TACA. Arrows show the durations of TACA. The symbol /x/ stands for voiced obstruents (/z, b, d, g/).
The data set used in the current and preceding studies is a subset of the CSJ-Core consisting of 162 spontaneous monologue samples (either academic presentation talk or so-called simulated public speaking) of the Core part of the Corpus of Spontaneous Japanese (CSJ, hereafter) [18]. The CSJ-Core is annotated with respect to segmental and prosodic features using the X-JToBI annotation scheme [19].

The same data as analyzed in literatures [16, 17] was reanalyzed in the present study. They all belong to the monologue samples (either academic presentation talk or so-called simulated public speaking) of the Core part of the Corpus of Spontaneous Japanese (CSJ, hereafter) [18]. The CSJ-Core is annotated with respect to segmental and prosodic features using the X-JToBI annotation scheme [19].

The analysis conducted in the present study consists in the estimation of each speaker’s threshold of closure articulation, i.e., the TACA value with which the probability of successful vocal-tract closure becomes 50%. The working hypothesis is that, for a given consonant, the threshold is nearly constant for a given speaker within a given speaking style, hence is a good indicator of the speaker’s individuality with respect to the articulation of the consonant. Adopting this hypothesis, it becomes possible to make comparison between the threshold value and the observed success rate of the closure articulation for a speaker. Our expectation is that there is an inverse relationship between the threshold and observed closure rate; because the smaller the threshold value is, the larger the number of samples whose TACA values exceed the threshold value, and vice versa.

3. Analysis and discussion

3.1. Threshold of closure articulation

The analysis conducted in the present study consists in the estimation of each speaker’s threshold of closure articulation, i.e., the TACA value with which the probability of successful vocal-tract closure becomes 50%. The working hypothesis is that, for a given consonant, the threshold is nearly constant for a given speaker within a given speaking style, hence is a good indicator of the speaker’s individuality with respect to the articulation of the consonant. Adopting this hypothesis, it becomes possible to make comparison between the threshold value and the observed success rate of the closure articulation for a speaker. Our expectation is that there is an inverse relationship between the threshold and observed closure rate; because the smaller the threshold value is, the larger the number of samples whose TACA values exceed the threshold value, and vice versa.

The threshold (abbreviated as TC, hereafter) of a given consonant for a given speaker was computed in the following way. Firstly, logistic regression analysis was conducted using the success (realization as a stop or affricate) or failure (as a fricative) of vocal tract closure as the binary dependent variable, and the observed TACA values as the independent variable. The TACA values were all log-transformed prior to the analysis for better normality (the base of logarithm was 10). The analysis was conducted using the whole samples of a given consonant spoken by a given speaker in the CSJ-Core, by use of the glm function of the R language ver. 2.15.2.

Fig. 3 shows a typical example of analysis. Distribution of the binary dependent variable is shown by circles on top or bottom of ordinate (where the probability of closure is either 1.0 or 0.0). The smooth S-shaped line stretching along the abscissa is the probability of success in closure articulation as estimated by the logistic regression analysis. The value in abscissa where the probability becomes 0.5 is 1.72, and this is the TC value for this particular speaker (See the arrow in Fig. 3). In this example, the samples are distributed nearly evenly in the range 1.0-2.5, but this is not always the case. In some speakers’ data, most of the samples distribute above, or below, the threshold value. Hence the mean TACA of all samples cannot be a good estimator of TC.

The number of samples used in the analysis differed from speaker to speaker and from consonant to consonant, as shown in Table 2. The first row shows the total number of samples in the current data set, the second to the fourth rows show the minimum, maximum, and mean numbers of samples used in a regression analyses. Given the result of regression analysis, the threshold of closure articulation could be obtained by the following formula: \(-1.0 * \text{Intercept} / \text{Slope}\).
Table 2: Number of samples used in regression analysis

<table>
<thead>
<tr>
<th></th>
<th>/z/</th>
<th>/b/</th>
<th>/d/</th>
<th>/g/</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>14,603</td>
<td>9,272</td>
<td>34,289</td>
<td>18,893</td>
</tr>
<tr>
<td>MIN</td>
<td>31</td>
<td>20</td>
<td>69</td>
<td>37</td>
</tr>
<tr>
<td>MAX</td>
<td>324</td>
<td>273</td>
<td>1074</td>
<td>476</td>
</tr>
<tr>
<td>MEAN</td>
<td>118.7</td>
<td>75.4</td>
<td>278.8</td>
<td>152.9</td>
</tr>
</tbody>
</table>

There were a few cases where the slope value (i.e. the weight of log-transformed TACA value in the regression formula) did not show statistical significance at the level of 0.05. This happened mostly when the number of sample is relatively small or the speaker’s mean closure realization rate shows extreme values (higher than 0.9 or lower than 0.2), or both. These speakers were removed from the following analysis. The number of removed speakers were 6, 12, 7, and, 5 respectively for /z/, /b/, /d/, and /g/. Fig. 4 compares the densities of probability distributions of the threshold of closure articulation for /z/, /b/, /d/, and /g/, which were estimated using the density function of R. The four distributions retain the mutual order of the consonants as appeared in Fig. 2 above.

Fig. 5 shows the scatter plots of the estimated thresholds of closure articulation and the observed mean rate of vocal-tract closure. As expected, in each consonant, there is negative correlation between the TC and the mean closure rate. All correlations were statistically significant at the level of \( p < 0.0001 \).

3.2. Skewed distribution of phonemes and pauses

It is fundamentally important, from a point of view of the present study, to point out that the correlations observed in Fig. 5 are not perfect. See the coefficients of correlation (r) shown in each panel. In the case of /b/, for example, the mean closure rates of speakers whose TC values are in the range 1.60-1.65 distribute widely from 0.39 to 0.77. Similarly, in the case of /g/, the mean closure rates of speakers whose TC values are in the range 1.70-1.75 distribute widely from 0.12 to 0.76. It might be the case that, in addition to the normal measurement error that follows normal distribution, the dispersion was caused by factors other than the TC.

A candidate of such factor is the distributional properties of phonemes /N/ and /Q/ and the pauses that immediately precede the obstruent in question. It is clear from the definition of TACA given above that the increase in frequencies of these segments can raise the mean realization rates of vocal-tract closure of a speaker, via the increase in the mean TACA value of the speaker. And, in spontaneous speech, the occurrences of these segments differ considerably from talk to talk depending on, among other things, the topics of talks and speaking styles.

If, for an example, the topic of a talk is concerned both with /nihoNzin/ (Japanese) and /roNgo/ (Confucian Analects), frequent occurrence of these keywords in which obstruents are in a context favoring the realization of vocal-tract closure, will contribute to the increase in the mean closure rate of /z/ and /g/.

Part of the perturbation of data along the ordinate of Fig. 5 could be the result of such skewed distribution of /N/, /Q/, and preceding pauses in each talk.
To examine this hypothesis, the ratio of samples that were directly preceded by either /N/, /Q/, or a pause to the total number of samples was computed for each speaker. Fig. 6 plots the same data as in Fig. 5, but the top 15 speakers with respect to the ratio of /N/, /Q/, and pause were plotted using the symbol of “X” and the bottom 15 speakers were plotted using the symbol “%”.

In the cases of /z/, the top 15 speakers were all located above the bottom 15 speakers. This is exactly the relationship as predicted by the hypothesis stated above. It seems to be the case that considerable amount of the vertical perturbation of data can be explained by the distribution of /N/, /Q/, and pause in front of the obstruent in question.

The same relationship can be seen in the case of /b/ as well. But in the cases of /d/ and /g/, the expected relationship is hardly recognizable.

Lastly, coefficients of determination were computed using the data of speakers who were not included in neither the top nor the bottom 15 speakers. Table 3 compares these coefficients (shown as Fig. 6) with the coefficients computed using the data of all speakers (shown as Fig. 5). There is considerable increase in the coefficients in /z/ and /b/, a few in /d/, and none in /g/.

### Table 3: Comparison of coefficients of determination ($R^2$) in figures 5 and 6.

<table>
<thead>
<tr>
<th></th>
<th>/z/</th>
<th>/b/</th>
<th>/d/</th>
<th>/g/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig. 5</td>
<td>0.747</td>
<td>0.536</td>
<td>0.537</td>
<td>0.488</td>
</tr>
<tr>
<td>Fig. 6</td>
<td>0.808</td>
<td>0.761</td>
<td>0.578</td>
<td>0.438</td>
</tr>
</tbody>
</table>

### 4. Concluding remarks

The present study revealed that inter-speaker differences in the mean rate of vocal-tract closure of voiced obstruents in spontaneous Japanese did not necessarily reflect the inherent characteristic of each speaker’s speech behavior. It seems to be the case, rather, that the observed differences were the conjoint result of the difference in the threshold of closure articulation (which is a genuine speaker-specific factor) and the skewed distribution of specific phonemes and pause that directly influence the TACA value (which is a property of phonetic and/or lexical contexts rather than the property of speakers). The influence of the latter factor was observed clearly in /z/ and /b/, but not in /d/ and /g/. Part of the reason for the phoneme-dependent difference may consist in the peculiar behavior of /g/ as mentioned in section 2.2 above. But the full understanding of the difference including both /g/ and /d/ is an important next step of the inquiry.

Another research question of prime importance both for phonetics per se and socio-phonetics, which is not treated in the present study, is the influence of speaking style on the TC. In the present study, talks of different speaking registers, academic presentation talk and simulated public speaking, namely, were analyzed altogether. But there is possibility that speakers change their TC depending on speech registers.

A preliminary comparison of six speakers who provided both academic talk and public speaking in the CSJ-Core revealed the tendency for four speakers to have smaller TC in their academic talks than in the public speaking. Needless to say, we need to annotate more data than we have presently to arrive at any more reliable result in this line of research. This should be the theme of a separate paper.

### 5. Acknowledgements

This work is supported by the NINJAL collaborative research project “Foundations of Corpus Japanese Linguistics.”
6. References


