A CASE STUDY OF SPONTANEOUS SPEECH IN JAPANESE

Takayuki Arai
Sophia University, Tokyo, JAPAN

ABSTRACT

This paper investigates spontaneous speech in Japanese, particularly phonetic phenomena that do not normally occur in carefully pronounced formal speech. I discuss several pronunciation variations in a corpus, including reduction or deletion of both vowels and consonants.

In this study, I also analyzed a specific speech utterance, which has a combination of some of those pronunciation variations, in detail. It was often difficult to identify each segment in the utterance solely by listening to just the few segments themselves (micro-listening), even if listening to the entire phrase (macro-listening) sounded intelligible. I conducted a perceptual experiment using this utterance; the results showed that the same speech segment was perceived as two morae by micro-listening and as five morae by macro-listening.

Listeners use a combination of temporal and contextual cues to reconstruct a speaker's intentions, although the phenomena found in spontaneous speech show that phonetic segments may change their appearance.

1. INTRODUCTION

Nearly all that is known about spoken Japanese is derived from studies of either read text or citation-form speech. However, it is unclear how representative these speaking styles are of unscripted, spontaneous Japanese.

In spontaneous speech, there is more variation at both segmental and suprasegmental levels than is present in carefully spoken formal speech. For examples, analysis of spontaneous speech shows different results for the phenomenon of mora timing than have been found in previous literature based on careful speech, as described in [1]. At the segmental level, both carefully articulated and spontaneous speech exhibit coarticulation. However, spontaneous speech also exhibits many phonetic properties that are not found in careful speech.

The influence of speaking style is likely to result in a different perception of a speech segment depending on how much context is available to give the listener information about speaking rate and carefulness. Suppose a specific speech utterance sounds natural and is intelligible, when a listener processes the speech within its natural context. If the utterance contains many rapidly articulated segments, it might be difficult to identify the words entirely through "microscopic" listening to short snippets in isolation.

In the present study, I investigated the effect of micro and macro listening as well as several specific phonetic phenomena in a corpus. Section 2 discusses several phonetic phenomena observed in spontaneous speech of Japanese (OGI Multi-Language Telephone Speech Corpus). Section 3 describes a perceptual experiment using a specific speech utterance to investigate the effect of contextual information about speaking rate and style on the number of morae a signal is perceived as having.

2. PRONUNCIATION VARIATIONS

In this section, I discuss several common pronunciation variations which are especially characteristic of spontaneous speech. The Japanese language materials used in the present study form a subset of the OGI Multi-Language Telephone Speech Corpus of spontaneous, informal speech spoken over the telephone by native speakers, discussing a topic of their choosing for approximately 60 seconds [2]. Each monologue was carefully transcribed at the phonetic-segment and moraic levels by the author, a phonetically trained, native speaker of Japanese. Filled pauses, hesitations and other instances of significant interruption in the speech stream were also transcribed. The segmentation was performed using both the waveform and the spectrogram.

2.1. Vowels

1) Devoicing under /C[−voice] V[+high] C[−voice]/

is very common in Japanese [3], e.g. /i/ in /deshita/ 'COP-POL-PAST' or /-mashita/ 'Vinfl-POL-PAST'.

2) Devoicing under /C[−voice] V[+high] #/

is also common [3], e.g. /u/ in /desu/ 'COP-POL-NONPAST' or /-masu/ 'Vinfl-POL-NONPAST'.

Han [4] discussed words with strings of high vowels between voiceless consonants, e.g. /chishiki/ 'knowledge' and /tsukusu/ 'exhaust'. She claimed that devoiced and voiced vowels alternate in such words (with an effect of accent location), so that one would never find

3) Consecutive devoicing.

On the other hands, vowel deletion can lead to

4) Gemination

as a phonological process of Japanese, e.g. /sentakki/ for /sentakuki/ 'washing machine', /ongakkai/ for /ongakukai/ 'concert', and /sankakkei/ for /sankukei/ 'triangle'.

Other variations are as follows:

5) Elisions, and
6) Glide formation.

These typical phonetic phenomena are occasionally observed in spontaneous speech. In addition to it, devoicing under non-typical environments is often observed in the corpus.

2.1.1. Devoicing under non-typical environments. Between voiceless consonants such as 1) and 2) are not the only environments observed in the corpus:

1a) /C[−voice] V[+high] C[+voice]/,

e.g. /u/ in /gogatsu no .../ '... of May', /i/ in /hanashimasu/ 'speak-POL-NONPAST', the first /u/ in /kuru/ 'come', /suru/ 'do', /shumi/ 'hobby', and /sugoku/ 'terribly'.

1b) /C[+voice] V[+high] C[−voice]/,

e.g. /i/ in /jitensha/ 'bicycle'.

1c) /C[+voice] V[+high] C[+voice]/,

e.g. /i/ in /hajime/ 'beginning'.

1d) /C[−voice] V[−high] C[−voice]/

Some cases of non-high vowel devoicing have been noted in the literature on careful speech, particularly when a mora is repeated, as the first /o/ in /kokoro/ 'heart' [3]. However, non-low vowel devoicing is much more widespread in this corpus than the previous literature has indicated. Examples in the corpus are: /o/ in /totemo/ 'very', /e/ in /heta/ 'clumsy', the first /o/ in /koko/ 'here', the first /a/ in /kakaru/ 'take', and /e/ in /omotteta/ 'thought' (Fig. 1).

Such devoicing may occur more easily when the vowel is low-pitch accented in careful speech. Some of the examples are, however, the cases where devoicing is occurred with high-pitch accented vowels, e.g. the first /o/ in /sótoni/ 'outside', the second /e/ in /yameteshimau/ 'stop', and the second /a/ in /mata ténisu/ 'again, tennis', where I indicated the last high-pitched mora with an acute accent mark over the vowel.

1e) /C[−voice] V[−high] C[+voice]/,

e.g. the first /o/ in /sono/ and /sore/.

1f) /C[+voice] V[−high] C[+voice]/,

e.g. the first /e/ in /madedete(iku)/ '(go) out to'.

2.1.2. Consecutive devoicing. I observed some of consecutive devoiced syllables in the corpus. The following examples are the combination of 3) and 1a): [wɑtɑkɯ̥ʃi̥no] 'my' (two in a row), [jo:ɸɯ̥kɯ̥mo] 'clothes also', and [ki̥te̥kɯ̥ɾeɾ ɯ] 'come (for someone)' (three in a row).

2.1.3. Gemination. Vowel deletion leading to 4) gemination can apply in more words and more environments as a type of fast speech variation than it does as a phonological process. Examples in the corpus are as follows: /kakukoto/ 'writing' as /kakkoto/, /hatarakukoto/ 'working' as /hatarakkoto/, /kikukoto/ 'hearing' as /kikkoto/, /rokugatsu/ 'June' as /rokgatsu/, and /kankaku ga/ 'sensation-SUBJ' as /kankakga/.

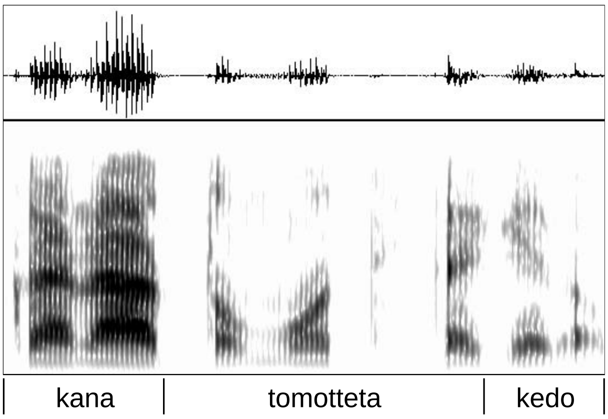

Figure 1. Waveform and its spectrogram of a speech utterance /... kana tomotteta kedo/ 'thougt it may be ... but'.

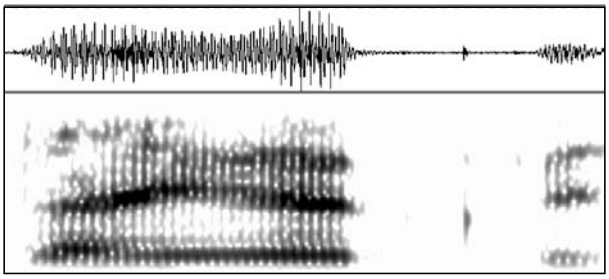

Figure 2. Waveform and its spectrogram of a speech utterance /daigaku/ 'university'. The segment /g/ is approximated.

Whether /u/ between two velar obstruents elides in spontaneous speech or not is a matter of free variation, and there is also an intermediate degree of reduction, i.e.,

[kɯk] → [kɯ̥ k] → [kː̚k], or
[kɯg] → [kɯ̥ g] → [kː̚g].

This pronunciation variation may depend on the speaking rate.

2.1.4. Elisions. The elision of consecutive vowels are also common in the corpus, e.g. /to omotte/ surfacing as [tomotː̚te]. In this example two consecutive /o/'s merged and became one vowel /o/. The resulting vowel /o/ is not necessarily two morae long in its duration, although it sounds like two /o/'s perceptually (Fig. 1).

2.1.5. Glide formation. Glide formation is also observed in the corpus, e.g. /dake attara/ 'if there is only' becomes [dɑkjɑtː̚tɑɾ ɑ], /bungaku o/ 'literature-OBJ' becomes [bɯŋ gɑkwo], and /... tteiu/ 'called ...' becomes [tː̚tijɯː].

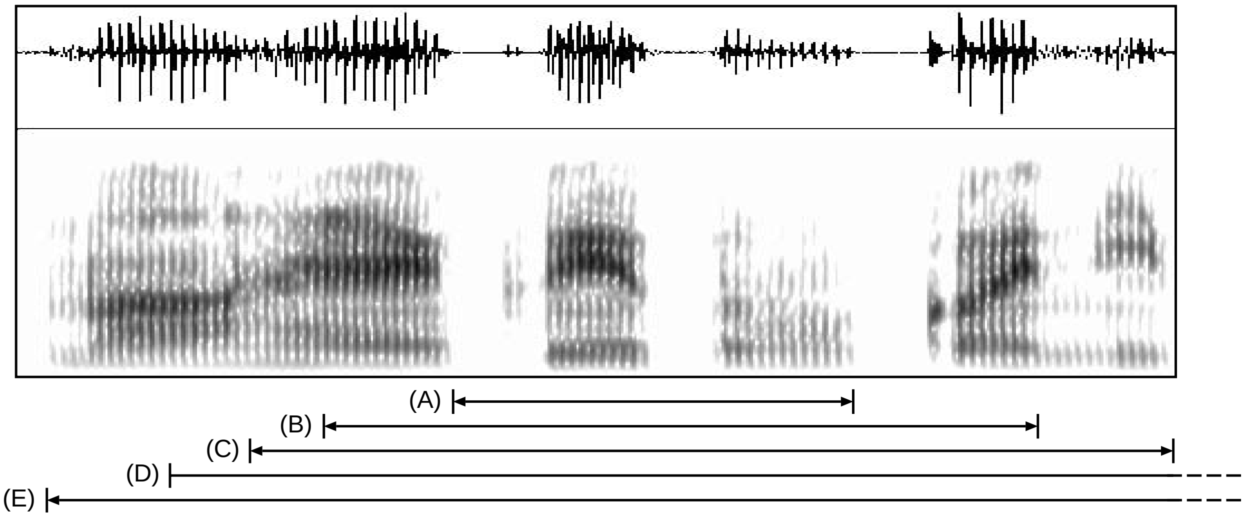

Figure 3. Waveform and its spectrogram of a speech utterance used for the perceptual experiment. The truncated speech intervals (A)-(F) were presented to each subject where the interval (F) is the whole phrase of this utterance.

2.2. Consonants

The following pronunciation variations are common for consonants in any styles of speech:

7) Voicing during the glottal fricative /h/
8) Nasalized vowels before nasals
9) Approximated voiced stops
10) Other consonantal reduction
11) Retroflex stops and approximants for flap /r/

In spontaneous speech, I observed many of the pronunciation variations above, as well as many others.

2.2.1. Consonatal reduction. As an extreme example of 10), the consonant is sometimes completely deleted, e.g. /z/ in /tsuzukete/ ‘continuously’, the first /r/ in /korekara/ ‘from now on’, /r/ in /irutoki/ ‘when (I) am there ...’. These reductions may go completely unnoticed, especially in the presence of other cues, such as elongation of surrounding sounds.

As an intermediate step short of deletion for the consonants /n/ and /d/, they sometimes become flaps, e.g. the first /n/’s in /kanojo/ ‘she’, and /yonen/ ‘four years’; and /d/’s in /keredo/, /kedo/ ‘but’, and /cho:do/ ‘just’ (see also /n/ in /kana/ of Fig. 1).

2.2.2. Voiced glottal fricatives. As an example of 7) in the corpus, voiced glottal fricatives merge with their adjacent segments, e.g. /sukoshi hanashite .../ ‘speak a little bit ...’ becomes [sɯ̥koʃɑnɑʃite].

2.2.3. Nasalized vowels before nasals. An example of 8) in the corpus is that the word /tenisu/ ‘tennis’ becomes [tẽːsɯ̥]. Deletion of the nasal and following /i/ and nasalization and lengthening of the /e/ are related processes.

2.2.4. Approximated voiced stops. The articulation of stop consonants in rapid speech is often inexact, with approximation rather than full oral closure. The voiced stops /b, d, g/ are often approximated in the corpus, e.g. /g/ in /daigaku/ ‘university’, /b/ in /obasan/ ‘aunt’, and /d/ in /... kata desu/ ‘is ... person’. This is also a very common phenomenon in English [5].

In a more extreme version of this process, consonants may elide completely, i.e.,

$$[b], [d], [g] \rightarrow [\beta], [\breve{ð}], [ɣ] \rightarrow \emptyset$$

Fig. 2 shows an example of /daigaku/. In this case, there is no acoustical evidence of [g], but native speakers perceive it as having a /g/.

2.2.5. Retroflex stops and approximants for flap /r/. Retroflex stops [ɖ] is common in Japanese for flap /r/. In the corpus, /r/ is also sometimes pronounced as [l], e.g. /r/’s in /kara/ ‘from’ and /abura/ ‘oil’.

3. MICRO AND MACRO LISTENING

In spontaneous speech, listening to a small portion in isolation (micro listening) often gives us very different perceptual impressions from listening to it as a part of long portion (macro listening). It is a result of segmental and suprasegmental effects, such as pronunciation variations and context. As described in Section 2, a segment becomes different from the prototypical form as a result of pronunciation variation. At the same time, it has been reported that phonetic perception depends on speaking rate [6]. By micro listening it is difficult to identify the intended form. Macro listening, however, gives more contextual information about speaking rate and style, and underlying forms and/or acoustically missing segments can be restored perceptually.

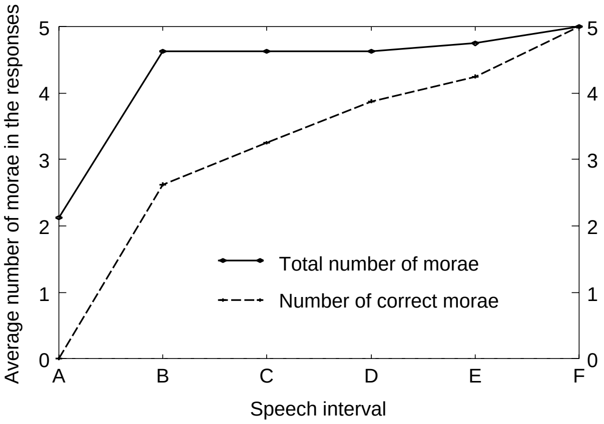

Figure 4. Average number of morae in eight subjects' written responses which corresponded to the portion of the utterance in interval (A). The horizontal axis represents the speech intervals they listened to. The dashed line is the average number of correct mora corresponded to its portion.

Fig. 3 shows an example which gives us very different perceptual impressions depending on how much of the signal is heard. The short speech interval (A), which has a duration of about 300 ms, when heard by itself sounds like two morae, e.g. /kebo/, although /kebo/ is a meaningless, unparsable string. This is natural when considering the average duration of a consonant-vowel (CV) mora in this corpus, 138 ms.

As one listens to longer portions of the signal (speech intervals B through F), the whole phrase becomes intelligible and sounds like a very natural speech utterance of /hayaku ieba hokani/ 'in short, other...'. In fact, the intended phrase corresponding to the speech interval (A) is /kuiebaho/ with five morae — much more than one can hear by micro listening. It seems that this is mainly caused by elision (5), glide formation (6), and voicing of the glottal fricative (7).

I conducted a perceptual experiment using this utterance in order to investigate the effect of speaking rate context on number of syllables (morae) perceived, and to confirm that naive listeners share the perception of this utterance described above. Twelve subjects participated in this experiment and were asked to respond with what they heard for each speech interval from (A) through (F). The same subjects heard all the different speech intervals, and they heard them in order from shortest to longest. The number of morae in their written responses which corresponded to the portion of the utterance in interval (A) was measured.

Fig. 4 shows the average number of morae in the written responses which corresponded to the portion of the utterance in interval (A). The average was taken over eight subjects whose mora-level recognition of the utterance in interval (A) was 100% when listening to the speech interval (F). (The average number of correct mora corresponded to its portion is shown as the dashed line in this figure.) As shown in Fig. 4, the graphs jump from the interval (A) to (B) dramatically. Thus, the average number of morae in responses increased from two to five as the speech interval heard was lengthened.

4. SUMMARY

I discussed several pronunciation variations in the OGI Multi-Language Telephone Speech Corpus. I also described a perceptual experiment using a specific speech utterance to investigate the effect of contextual information about speaking rate and style on the number of morae a signal is perceived as having. The phenomena found in spontaneous speech show that phonetic segments may change their appearance (even to the point of deletion). However, listeners use a combination of temporal and contextual cues to reconstruct a speaker's intentions.

ACKNOWLEDGMENTS

I would like to express my appreciation to Steven Greenberg, John Ohala, Natasha Warner, Masahiko Komatsu, Makiko Aoyagi, and Noboru Kanedera for their comments on this paper. I also gratefully acknowledge the assistance of the listeners from Sophia University who participated in the perceptual experiment in this study.

REFERENCES

[1] Arai, T. and Warner, N. 1999. Word Level Timing in Spontaneous Japanese Speech. In *Proc. Int'l Cong. on Phonetic Sciences*.
[2] Muthusamy, Y.K., Cole, R.A. and Oshika, B.T. 1992. The OGI Multi-Language Telephone Speech Corpus. In *Proc. Int'l Conf. on Spoken Language Processing.*
[3] Vance, T.J. 1987. *An Introduction to Japanese Phonology*. State University of New York Press, Albany.
[4] Han, M.S. 1962. Unvoicing of Vowels in Japanese. *Onsei no Kenkyū*, 10, 81–100.
[5] Greenberg, S. 1997. The Switchboard Transcription Project, Technical Report, 1996. *Johns Hopkins CSLP Workshop on Innovative Techniques for Large Vocabulary Continuous Speech Recognition,* Baltimore, MD.
[6] Miller, J.L. and Dexter, E.R. 1988. Effects of Speaking Rate and Lexical Status on Phonetic Perception. *Journal of Experimental Psych.: Human perception and performance*, 14, 369–78.