ACOUSTICAL LETTER

Analysis of spontaneous Japanese in a multi-language telephone-speech corpus

Takayuki Arai[1,*], Natasha Warner[2] and Steven Greenberg[3]

[1]*Department of Electrical and Electronics Engineering, Sophia University, 7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan*
[2]*Department of Linguistics, University of Arizona, PO Box 210028, Tucson, AZ 85721–0028, USA*
[3]*Silicon Speech, 46 Oxford Drive, Santa Venetia, CA 94903, USA; Centre for Applied Hearing Research, Technical University of Denmark, Kgs. Lyngby, DK-2800, Denmark*

(*Received 2 May 2006, Accepted for publication 24 July 2006*)

Keywords: Phonetic analysis, Spontaneous Japanese, Speech corpora, Frequency of occurrence, Duration
PACS number: 43.70.Fq [doi:10.1250/ast.28.46]

1. Introduction

We phonetically analyzed a spontaneous Japanese component of the Oregon Graduate Institute Multi-Language Telephone Speech (OGI-TS) Corpus [1], which is widely used for a variety of speech technology applications such as automatic language identification. There is increasing interest in quantitative analysis of spontaneous speech for Japanese [2] as well as other languages. Phonetic descriptions have traditionally focused on either read text or citation-form (i.e., formal, carefully pronounced) speech. However, the phonetic properties of careful speech often differ from spontaneous speech [3–7]. Because speech technology is becoming increasingly focused on real-world applications, it is important to quantitatively analyze spontaneous material. Therefore, we discuss several phonetic phenomena characteristic of spontaneous Japanese, including segmental reduction and deletion, as well as the frequency of occurrence and duration of vowels and consonants in Japanese.

2. Analysis

We analyzed a subset of spontaneous, informal Japanese spoken over the telephone by native speakers from the OGI-TS Corpus [1]. This corpus contains 90 calls. Each call was uttered by a unique adult speaker. From the speech of each person, one-minute of free speech, discussing a topic of their choosing for approximately 60 seconds, was chosen. This one-minute of free speech was divided into the utterances before and after the tone, which informed the caller that the remaining time was ten seconds. For the analysis, we used the first 50 seconds (the longest utterance of spontaneous speech for each speaker in this corpus) for each of 30 speakers with Tokyo dialect. The speech data were sampled at 8 kHz with a 13-bit resolution.

Each monologue was carefully transcribed at the phonetic-segment and moraic levels by the first 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.

*e-mail: arai@sophia.ac.jp

2.1. Vowels

Although there was some speaker variation, we observed many phenomena that were in common among multiple speakers.

2.1.1. Devoicing

Devoicing in

/C[−voice] V[+high] C[−voice]/ (1)

sequences is very common in Japanese [8], e.g., /i/ in /deshita/ ‘was’ or /-mashita/ ‘past tense verb ending.’ (The plus ‘+’ and minus ‘−’ annotation reflects the presence or absence of a distinctive feature [9]. The feature [voice] refers to the presence or absence of voicing, which is the acoustic manifestation of vibration of the vocal folds that is generally applied to the entire duration of a segment. The feature [high] refers to the height of the tongue blade during articulation of a vowel and is also associated with the frequency of the first formant. A high vowel generally has a first formant frequency below 450 Hz. Instances of a high vowel in Japanese would be /i/ and /u/.) Devoicing in

/C[−voice] V[+high] #/ (2)

sequences is also common [7], e.g. /u/ in /desu/ ‘is’ or /-masu/ ‘non-past polite verb ending.’

In addition to observing these expected phenomena in spontaneous speech, we also often observed the following cases of devoicing under non-typical environments in the corpus [4]. The first case of non-typical devoicing is in

/C[−voice] V[+high] C[+voice]/ (3)

sequences, e.g. /u/ in /gogatsu no .../ ‘... of May,’ /i/ in /hanashimasu/ ‘speak (non-past polite verb ending),’ the first /u/ in /kuru/ ‘come,’ /suru/ ‘do,’ /shumi/ ‘hobby,’ and /sugoku/ ‘terribly.’ The second case is in

/C[+voice] V[+high] C[−voice]/ (4)

sequences, e.g. /i/ in /jitensha/ ‘bicycle.’ The third case is in

/C[+voice] V[+high] C[+voice]/ (5)

sequences, e.g. /i/ in /hajime/ ‘beginning.’

Table 1 shows the frequency of occurrence for devoicing

Table 1 Devoicing rate of vowels.

		Devoiced	Not devoiced	Devoicing rate
High vowel	Devoicing context	279	10	96.5%
	Non-devoicing context	157	1,021	13.3%
Non-high vowel	(Non-devoicing context)	43	3,856	1.1%

Table 2 Frequency of occurrence and duration (in ms) for short vowels. "*N*" denotes the frequency of occurrence. "20%," "50%," and "80%" denote percentiles.

Segment	*N*	Mean	20%	50%	80%
/a/	1,855	82.3	53.0	73.3	103.0
/e/	848	85.7	48.8	71.0	113.4
/i/	1,022	67.5	41.9	59.0	86.2
/o/	1,196	75.4	47.0	66.0	94.3
/u/	447	56.8	33.0	48.1	71.2
Total	5,368	76.4	46.7	66.1	97.1

Table 3 Frequency of occurrence and duration (in ms) for long vowels. "*N*" denotes the frequency of occurrence. "20%," "50%," and "80%" denote percentiles.

Segment	*N*	Mean	20%	50%	80%
/a:/	31	122.4	94.0	110.0	136.7
/e:/	44	120.5	93.0	109.9	145.2
/i:/	58	123.4	90.0	122.0	148.8
/o:/	188	116.4	82.0	110.0	145.0
/u:/	86	110.3	67.5	101.0	137.8
Total	407	117.0	83.0	110.0	145.0

of vowels observed in this corpus [6]. In this table, "devoicing context" corresponds to the cases (1) and (2), and "non-devoicing context" corresponds to the cases (3), (4) and (5) above.

Non-high vowel devoicing is far more common in this corpus than would be anticipated on the basis of the published literature [8]. The first case of non-high vowel devoicing is in

$$/C[-\text{voice}]\ V[-\text{high}]\ C[-\text{voice}]/ \quad (6)$$

sequences, e.g. first /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.' The second case is in

$$/C[-\text{voice}]\ V[-\text{high}]\ C[+\text{voice}]/ \quad (7)$$

sequences, e.g. the first /o/ in /sono/ and /sore/. The third case is in

$$/C[+\text{voice}]\ V[-\text{high}]\ C[+\text{voice}]/ \quad (8)$$

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

2.1.2. Frequency of occurrence and duration

Tables 2 and 3 show the frequency of occurrence and duration for vowels [6]. The ratio between the average durations of long and short vowel is 1.53 (117.0 ms/76.4 ms). This ratio is much less than reported in the previous literature for careful speech, such as when speakers read a list of words one at a time while sitting in a sound booth (e.g., [10]). Careful speech is usually characterized by careful pronunciation of most consonants and many vowels. Words tend to be articulated more distinctly and are generally longer in duration. This is particularly the case for vowels and consonant codas (which may be highly reduced or even deleted). In Japanese, one of the main differences between careful and spontaneous speech may be in terms of the proportion of vowel devoicing and deletion.

2.2. Consonants

For consonants, we also observed many phenomena that are common among multiple speakers, although there was some speaker variation.

2.2.1. Variation in pronunciation

The common variations in pronunciation of consonants in Japanese are: voicing during the glottal fricative /h/, nasalization of vowels before nasals, approximated voiced stops, retroflex stops and approximants for flap /r/, and other forms of consonant reduction. In spontaneous Japanese speech, we have observed many of these pronunciation patterns as well as some others [4].

The consonant may sometimes be 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 by the listener, especially in the presence of other cues, such as lengthening of surrounding segments. As an intermediate step short of deletion for the consonants /n/ and /d/, these segments sometimes become flaps, e.g. the first /n/ in /kanojo/ 'she,' and /yonen/ 'four years'; and /d/ in /keredo/, /kedo/ 'but,' and /cho:do/ 'just.'

In the corpus, other variations of pronunciation for consonants were observed, as well [4]. One example is voiced glottal fricatives merging with adjacent segments, e.g. /sukoshi hanashite .../ 'speak a little bit ...' becomes [sɯkoʃɑnɑʃite]. There is also the word /tenisu/ 'tennis,' which becomes [tẽːsɯ]. Deletion of the nasal and following /i/ and nasalization and lengthening of the /e/ are related processes.

The articulation of stop consonants in rapid speech is often imprecise, with approximation rather than full oral closure. This phenomenon is called spirantization and is common in Japanese as well as in English [6]. The voiced stops /b, d, g/ are often approximated in the corpus, e.g. /g/

Table 4 Frequency of occurrence and duration (in ms) for short consonants. "*N*" denotes the frequency of occurrence. "20%," "50%," and "80%" denote percentiles. "/ /" is used for a phoneme, while "[ ]" is used for an allophone.

Segment	*N*	Mean	20%	50%	80%
/p/	32	74.1	40.3	68.0	106.2
/t/	640	76.5	56.0	76.4	96.3
/k/	816	83.1	61.0	80.0	104.9
/b/	115	55.9	36.0	52.9	73.4
/d/	401	44.6	27.0	43.0	61.0
/g/	240	40.3	26.7	37.0	51.0
[s]	429	105.5	69.0	92.6	133.4
[ʃ]	320	97.1	72.0	92.0	120.0
[h]	146	69.4	46.9	66.0	84.5
[ɸ]	32	86.2	63.2	84.0	101.0
[ts]	77	108.7	90.0	108.3	130.0
[tʃ]	105	102.7	78.9	100.8	122.1
[(d)z]	51	61.8	40.4	62.0	75.1
[(d)ʒ]	106	69.1	48.0	67.0	92.0
/r/	495	29.3	20.4	27.2	37.0
/w/	197	47.8	29.3	45.0	61.6
/y/	221	46.5	29.0	41.7	60.3
/m/	521	61.2	46.0	61.4	75.6
/n/	635	52.8	35.1	49.3	68.0
/N/	321	77.3	55.0	72.5	96.0
Total	5891	67.2	37.0	63.0	92.0

in /daigaku/ 'university,' /b/ in /obasan/ 'aunt,' and /d/ in /... kata desu/ 'is ... person.' According to a statistical analysis of allophones of /g/ in this corpus, [g] that has a clear burst was 20.4%, [ɣ] that does not have a clear burst but only frication was 72.2%, and velar nasal [ŋ] was 7.4% out of the entire frequency of occurrence of phoneme /g/.

/r/ in Japanese is generally a flap; however, in spontaneous speech, the variations are widely spread. According to another statistical analysis of allophones of /r/ in the intervocalic context in this corpus, the frequency of the flap [ɾ] was 82.3%, the weak flap was 6.4%, while there was no evidence of /r/ in 4.7% of the instances. In addition, 1.7% of /r/ was the flap [ɾ] preceded by a consonant. Furthermore, 4.9% out of /r/ was stop [ɖ] (one in intervocalic context: 2.8%; and one preceded by silence: 2.1%). In the corpus, /r/ is also sometimes pronounced as [l], e.g. the /r/ in /kara/ 'from' and /abura/ 'oil.'

2.2.2. Frequency of occurrence and duration

Tables 4 and 5 show the frequency of occurrence and duration for consonants. In these tables, each phoneme is divided into the several main ways it can be pronounced (its allophones) except that all allophones of the syllabic nasal are grouped together as /N/. The distinction between short and long consonants is made based on orthographic representation (the typical form of the word as pronounced in careful speech). The ratio between the average durations of long and short consonants is 1.85 (142.4 ms/76.9 ms). As with vowels, this ratio is much less than the one reported in the literature for careful speech of Japanese [10].

Table 5 Frequency of occurrence and duration (in ms) for long consonants. "*N*" denotes the frequency of occurrence. "20%," "50%," and "80%" denote percentiles. "/ /" is used for a phoneme, while "[ ]" is used for an allophone.

Segment	*N*	Mean	20%	50%	80%
/pp/	9	139.6	118.0	140.0	156.5
/tt/	127	148.7	115.0	142.0	174.0
/kk/	38	172.2	123.9	153.0	211.3
[ss]	1	114.0	—	114.0	—
[ʃʃ]	9	165.9	130.0	144.0	177.0
/mm/	5	92.2	75.0	86.8	115.0
/nn/	64	113.8	78.2	102.0	136.0
Total	253	142.4	96.0	135.0	174.0

3. Summary

In this paper, we discussed pronunciation variations such as reduction or deletion, and frequencies of occurrence and duration of both vowels and consonants in a spontaneous speech corpus of Japanese. We need to continue to investigate spontaneous speech for further understanding of natural speech.

References

[1] http://www.ldc.upenn.edu/Catalog/LDC94S17.html

[2] H. Kikuchi, K. Maekawa, Y. Igarashi, K. Yoneyama and M. Fujimoto, "Phonetic labeling of the 'Corpus of Spontaneous Japanese'," *J. Phonet. Soc. Jpn.*, **7**(3), pp. 16–26 (2003).

[3] T. Arai, "A case study of spontaneous speech in Japanese," *Proc. Int. Congr. Phonetic Sciences*, pp. 615–618 (1999).

[4] T. Arai and N. Warner, "Word level timing in spontaneous Japanese speech, *Proc. Int. Congr. Phonetic Sciences*, pp. 1055–1058 (1999).

[5] N. Warner and T. Arai, "The role of the mora in the timing of spontaneous Japanese speech," *J. Acoust. Soc. Am.*, **109**, 1144–1156 (2001).

[6] T. Arai, N. Warner and S. Greenberg, "Analysis of spontaneous Japanese in OGI-Multi-Language Telephone-Speech Corpus," *Proc. Spring Meet. Acoust. Soc. Jpn.*, Vol. 1, pp. 361–362 (2001).

[7] S. Greenberg, H. Carvey, L. Hitchcock and S. Chang, "Temporal properties of spontaneous speech: A syllable centric perspective," *J. Phonet.*, **31**, 465–485 (2003).

[8] T. J. Vance, *An Introduction to Japanese Phonology* (State University of New York Press, Albany, 1987).

[9] R. Jakobson, G. Fant and M. Halle, *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates* (MIT Press, Cambridge, Mass., 1952).

[10] N. Warner and T. Arai, "Japanese mora-timing: A review," *Phonetica*, **58**, 1–25 (2001).