Acoustic Characteristics of Japanese Vowels

Tatsuya Hirahara and Reiko Akahane-Yamada***

* NTT Communication Science Labs., NTT Corporation
**ATR Human Information Science Labs.
hirahara@idea.brl.ntt.co.jp, yamada@atr.co.jp

Abstract

The purpose of this study was to build a large database on Japanese vowels and to clarify their acoustic characteristics. Recordings were made of 256 males and 252 females aged from 6 to 76 years old, for both Kanto (Tokyo) and Kansai (Osaka) dialects, producing the vowels /i, e, a, o, u/ in isolation, in /h-V-da/ syllables and in /b-V-ta/ syllables within and without a carrier sentence. The five vowels uttered with different notes were also recorded. A total of 154,570 vowel tokens were collected and phoneme labels were given to each of them. Analysis of the formant data showed that there was little difference between the two dialects. Formant frequencies were the highest for children's vowels and those of boys and girls were comparable in distribution. Formant frequencies of men were the lowest. The mean F0 of men dropped at the age of 12 to 13 due to the voice breaking. Formant frequencies shifted up when vowels were uttered at higher musical notes. The rate of the shift due to F0 change, however, was smaller than that seen among age groups. Dynamic features of formants varied with phonemic environment but were not affected by mora structure and talkers age.

1. Introduction

The major acoustic correlates of vowel quality are formant frequencies that reflect vocal tract resonances. Formant frequencies of American English vowels have been well studied as represented by Peterson and Barney (1952) and Hillenbrand *et al.* (1995).

Research on the acoustics of Japanese vowels started in the 1930's. Obata (1933) applied Fourier transform to the analysis of sounds including vowels. Chiba and Kajiyama (1942) carried out a pioneering work on vowels. Honda (2002) wrote a history of the incunabula. In the late 1950's, Hattori *et al.* (1957) and Torii (1957) measured formant frequencies of Japanese vowels by reading sonagrams. Suzuki *et al.* (1963) analyzed 100 monosyllables for the articulation test uttered by two male professional announcers by a moment method. Kadokawa and Nakata (1964) analyzed isolated vowels uttered by five adult males by means of an Analysis-by-Synthesis method. Kasuya *et al.* (1968) investigated changes in F0 and formant frequencies with age and gender for over 100 talkers aged between 7 and 20 years old using sonagrams. In the 1970's, digital computers began to be commonly used in speech studies. Sato (1975) analyzed vowels uttered by 34 male and 29 female talkers to investigate acoustic correlates that determine male and female voice qualities. Keating and Huffman (1984) reported F1-F2 diagram of Japanese vowels in 30 words for seven young male talkers. Miwa (1991) analyzed formant frequencies of isolated vowels uttered by 10 male professional announcers using the A-b-S method. Konno *et al.* (1994) compared formant frequencies of voiced and whispered vowels of three male talkers. Ueda (1995) analyzed isolated vowels uttered by 6- to 22-year-old talkers of the Kumamoto dialect (83 males and 125 females) to investigate speech spectral normalization for a tactile display. Recently, Okuda *et al.* (2002) analyzed a large-scale Japanese speech database for automatic speech recognition studies. The database consists of 177,163 utterances of 3771 talkers recorded in 18 major prefectural capitals from Hokkaido to Kyusyu.

Among these past studies, there are no systematic investigations of the acoustic characteristics of Japanese vowels except for those by Kasuya *et al.* and Okuda *et al.* Further, there are few acoustic analysis studies that focussed on dynamic feature of Japanese vowels. Most Japanese vowels ever studied were uttered in isolation or uncontrolled phonetic environments. The purpose of the present study, therefore, is to clarify the acoustic characteristics of Japanese vowels based on a large-scale database in which the uttering conditions are systematically varied.

2. Method

2.1. Talkers

Talkers were 508 native speakers of Japanese between the ages of 6 and 76. They were categorized into four groups; AM (adult males), AF (adult females), AD (adolescents), and C (children), as shown in Table 1. Dialect classification was conducted based on a language questionnaire sheet. A talker was classified as a Kanto/Kansai dialect speaker if he/she had lived in Kanto/Kansai area from 5 to 18 years old, or had lived in Kanto/Kansai area more than a half of his/her life. Of the 508 talkers, 220 were Kanto dialect talkers and 233

Table 1: Number of talkers for each age group.

Group	Age Range	Number of talkers (boys/girls)	Number of tokens
AM (Adult Male)	19-77	153	46,600
AF (Adult Female)	19-77	151	45,898
AD (Adolescent)	12-18	83 (42 / 41)	25,282
C (Child)	5-11	121 (61 / 60)	36,790
TOTAL		508	154,570

were Kansai dialect talkers. The remaining 55 talkers were neither Kanto nor Kansai dialect talkers.

2.2. Recordings

Talkers read lists containing the five Japanese vowels in several different conditions. (1) Sustained vowels /i/, /e/, /a/, /o/ and /u/ uttered in isolation (5 vowels * 3 times). (2) The five vowels uttered in isolation with different musical notes (5 vowels * 8 notes * 2 times). (3) The five vowels in /h-V-da/ and in /b-V-ta/ syllables within and without a carrier sentence, such as "h-V-da, kore wa h-V-da desu" (35 sentences * 3 times). In the ideal case, utterances containing 305 target vowel tokens were recorded from one talker. Some utterances, however, were discarded because of inadequate pronunciation, noise interference, and/or input level overflow during the recordings.

The recordings were made in anechoic rooms at NTT CS Labs., located in Kanto, and ATR, located in Kansai, with a digital audio tape recorder (SONY, PCM7010) and a condenser microphone (B&K, 4003). Sampling frequency of the digital audio tape recorder was set at 44.1 kHz and amplitude resolution was 16 bits. To reduce low-frequency fluctuating component, a high-pass-filter (NF, DV04), whose cutoff frequency was set at 73 Hz, was placed in the audio line. The recorded PCM data were then transferred to a PC via a digital audio interface (DAT-Link +), and then down-sampled to 22.05 kHz by the digital interface.

2.3. Labels

For all the recorded speech data, phoneme labels were given manually. Labels were assigned to the following points: start and end of a target vowel, start of /h/ and /d/ in /h-V-da/ syllables or /b/ and /t/ in /b-V-ta/ syllables, start and end of a carrier sentence and the word containing the target vowel. Four well-trained labelers gave the labels to the data by looking at the waveform, power contour, voiced/unvoiced index, fundamental frequency contour, formant frequency contour on a spectrogram using a custom ESPS based interactive labeling editor and by listening to the sound.

2.4. Acoustic Analysis

For the acoustical analysis, speech data were down-sampled again to 11.025 kHz. Fundamental frequencies (F0s) were extracted every 1ms by an auto-correlation-based F0 tracker of the ESPS speech analysis package. Speech data was windowed with 30-ms hanning window, and analyzed with 14th-order LPC. F0 tracking errors such as F0 halving or doubling were corrected by reanalyzing the signal with an adjusted upper and lower limit on F0.

The first to the fourth formant frequencies (F1-F4) and their bandwidth (B1-B4) were extracted using LPC-based formant tracker of the ESPS. The 14-pole LPC spectra were calculated every 1ms over 25-ms hanning-windowed segments. When the 14-pole LPC spectra did not give stable formant frequency contours over the target vowel tokens or gave excessively wide formant bandwidths, the analysis was judged to have failed. In that case, data was reanalyzed with different orders of LPC analysis in a predetermined sequence, i.e. 14-, 12-, 11-, 16- and 18-pole. The reanalysis was terminated when satisfactory formant frequency contours were obtained. This decision was made based on an examination of extracted formant frequencies overlaid on a spectrogram, LPC or Fourier spectral slices, and general knowledge of acoustics phonetics on vowels.

The LPC analysis often gave false formant frequencies for high F0 data. When satisfactory formant frequency contours were not obtained with 11- to 18-pole LPC analysis, these vowel tokens were not included in the analysis.

3. Results and Discussions

3.1. Vowels Uttered in Isolation

Figure 1 shows the mean F1 and F2 frequencies of isolated vowels for the four talker groups, along with

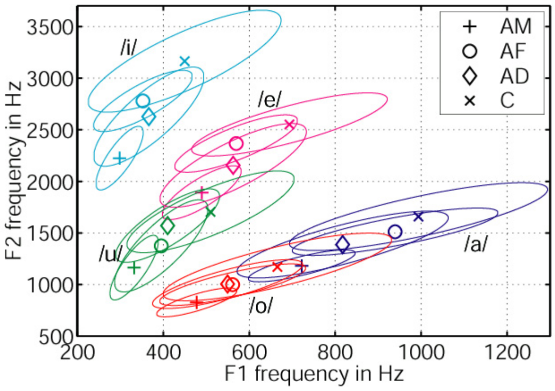

Figure 1. The mean F1 and F2 frequencies and 80% density ellipses of isolated vowels for the four talker groups. (N=152, 144, 82, and 119 for AM, AF, AD and C, respectively)

an elliptic fit to each vowel category. The vowels uttered by adult males (AM) are located in lower F1 and F2 regions, those of adult females (AF) and adolescents (AD) sit in similar regions, and those of children (C) are located in higher F1 and F2 regions. There is no distinct difference between the two talker dialect groups in the static features of the formant frequencies, possibly because the talkers read the sentence lists.

In Figure 2, mean F1 and F2 frequencies of isolated vowels uttered by adult male talkers of this study are compared with those of previous studies. The mean F1 and F2 frequencies for the five vowels bear close resemblance to previously reported data except for Okuda *et al*'s and Keating and Huffman's data. Vowels in these two data sets are found to lie inside the other vowel spaces. It is expected that co-articulation might have an affect on formant frequencies in these studies: Okuda *et al* analyzed the vowels in dialogues and read sentences, and Keating and Huffman analyzed short vowels in 30 words.

Figure 3 shows mean F0 frequency change along with age. While the mean F0 gradually declines as age increase for females, it drops rapidly for males at the age of 12 to 13 years due to the voice breaking. This age period is almost identical to that measured by Kasuya *et al.* in 1969.

3.2. Vowels uttered with different musical notes

Each talker uttered vowels with 8 different musical notes; i.e. talkers were asked to sing sol-fa in each vowel. Some talkers were able to sing with F0 range close to a full octave; many could not vary their F0s widely. Figure 4 shows the mean F1 and F2 frequencies of isolated vowels for the eight notes. The formant frequency shift due to individual talkers' F0 frequency shift was smaller than that seen among age groups. Note that the number of talkers is 356, i.e. almost 30% of the talkers were not used to plot the figure. The LPC analysis method is not versatile. The weakness of the LPC analysis is that it is hard to extract formant frequencies of vowels with very high F0 frequency. Many vowel tokens uttered at high notes by females and children, thus, remained unanalyzable.

3.3. Vowels uttered in syllables

As the vowel duration differs from token to token, the frequencies of F0, F1 and F2 were calculated at 11 equally spaced points for each vowel. Namely, frequencies were obtained at every 10% of the vowel duration including the starting point and end point.

Figure 5 shows the dynamic feature of mean F1 and F2 frequencies for short isolated vowels uttered in isolation, in /h-V-da/ syllables and in /b-V-ta/ syllables. F1 and F2 frequencies are almost fixed for isolated vowels, while they shift both in syllables. F1

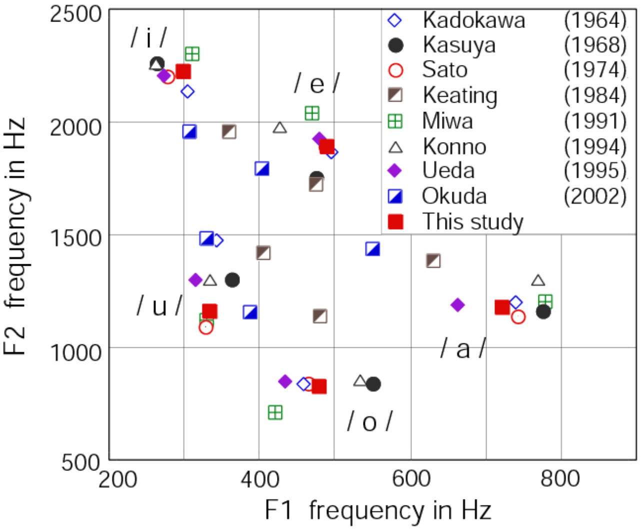

Figure 2. Comparison of the mean F1 and F2 frequencies of isolated vowels uttered by adult male talkers. For the present study, data by AM group are plotted (N=152).

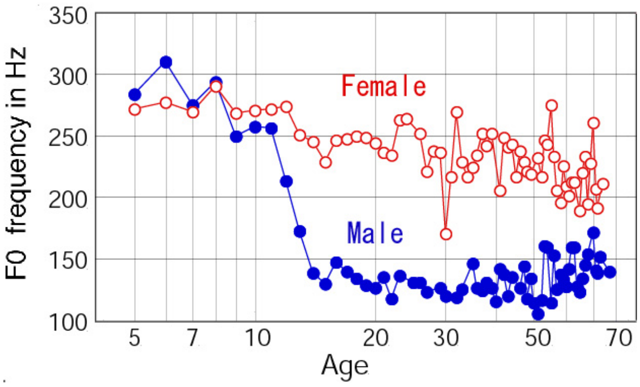

Figure 3. The mean F0 frequency change along with age groups (N=152 for male, and N=144 for female).

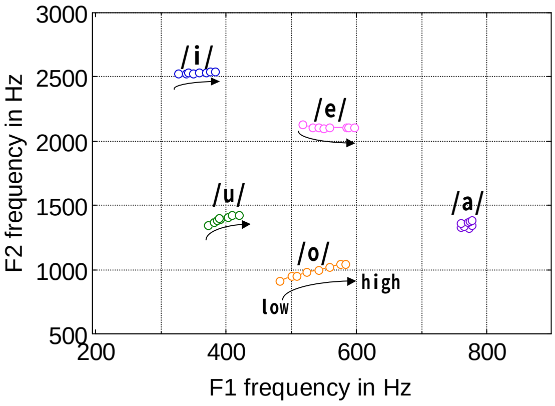

Figure 4. The mean F1 and F2 frequencies isolated vowels uttered in eight notes (N=356, from all taker groups).

frequencies shift down almost an octave in /h-V-da/ syllables and 2/3 octave in /b-V-ta/ syllables. In /h-V-ta/, F2 frequencies of /i/ and /e/ shift down a few hundred Hertz, those of /a/ remain constant and those of /o/ and /u/ shift up about 500 Hz. In /b-V-ta/ syllables, F2 frequencies of /i/ and /e/ remain constant and those of /a/, /o/ and /u/ shift upward a few hundred Hertz.

These dynamic features of F1 and F2 frequencies did not show substantial difference among different mora structures (short vowel vs. long vowel vs. short vowel followed by double consonant), speaking styles (in isolation vs. in carrier sentence), and age and dialect of the talkers.

4. Conclusions

The present study clarified some aspects of the acoustic characteristics of Japanese vowels based on a large-scale database, in which talker-related attributes, phonetic environments, and speaking styles were systematically balanced. F1 and F2 frequencies as well as F0 of isolated vowels in the present study bear close resemblance to previously reported data, suggesting that the present data can be regarded as a standard for the acoustic characteristics of Japanese vowels. It was newly found that the formant frequency shift due to individual talkers' F0 frequency shift was smaller than that seen among age groups, implying that the shift of the formant frequency is affected more by the length of the vocal tract than by tension of the vocal folds. It was also shown that the phonetic environment significantly affects the dynamic feature of formant frequencies.

Acknowledgment

Authors are grateful for Rieko Kubo and Naoko Ohtsuki at ATR for their technical assistance.

5. References

[1] Peterson, G.E and Barney, H. L. (1952), "Control Methods used in a study of the vowels," J. Acoust. Soc. Am. 24, 175-184.

[2] Hillenbrand, J., Getty, L.A., Clark, M.J., and Wheeler, K. (1995), "Acoustica characteristics of American English vowels," J. Acoust. Soc. Am. 97, 3099-3111.

[3] Obata, J. (1933), Jikken-Onkyogaku (Experimental Acoustics) (Iwanami Shoten, Tokyo, 1933). (in Japanese)

[4] Chiba, T. and Kajiyama, M. (1942), The vowel: Its Nature and Structure, (Tokyo-Kaiseikan, Tokyo, 1942).

[5] Honda, K. (2002) "Evolution of vowel production studies and observation techniques," Acoust. Sci. & Tech. 23, 189-194

[6] Hattori, S., Yamamoto, K., Obashi, Y. and Fujimura, Y. (1957), "Vowels of Japanese," Bulletin of the Kobayasi Institute of Physical Research 7, 69-79. (in Japanese)

[7] Torii, N. (1957) " Some consideration on Sonagrams of Japanese," Progress Report of NTT-ECL, No.579, 1-65. (in Japanese)

[8] Suzuki, J., Kadokawa, Y. and Nakata, K. (1963) "Formant-Frequency Extraction by the Method of Moment Calculations," J.Acoust.Soc.Am. 35, 1345-1353.

[9] Kadokawa, Y. and Nakata, K. (1964) "Extraction of Formant Frequencies by Analysis-by-Synthesis Techniques," J. Acoust. Soc. Jpn. 20, 1-13. (in Japanese)

[10] Kasuya, H., Suzuki, H. and Kido, K. (1968), "Changes in Pitch and first Three Formant Frequencies of Five Japanese Vowels with Age and Sex of Speakers," J. Acoust. Soc. Jpn. 24, 355-364. (in Japanese)

[11] Sato, H. (1974), " Acoustic Cues of Female Voice Quality," Trans. IECE 57-A, 23-30. (in Japanese)

[12] Keating, P.A. and Huffman, M.K. (1984), "Vowel Variation in Japanese," Phonetica 41, 191-207.

[13] Konno, H., Toyama, J., Shimbo, M. and Murata, K. (1994), A study on the formant frequency and phonemic quality of Japanese whispered vowels," J. Acoust. Soc. Jpn. 50, 623-630. (in Japanese)

[14] Miwa, J. (1991) Speech Signal Processing on a PC, (Shokodo, 1991, Tokyo). (in Japanese)

[15] Ueda, Y. and Watanabe, A. (1995), "Speech spectral normalization and its vowel vector representations for tactile display," J. Acoust. Soc. Jpn. 51, 519-528. (in Japanese)

[16] Okuda, K., Matsui, T., Naito, M., Sagisaka, Y. and Nakamura, S. (2002), "Creation and evaluation of a large-scale Japanese speech database," J. Acoust. Soc. Jpn. 58, 569-578. (in Japanese)

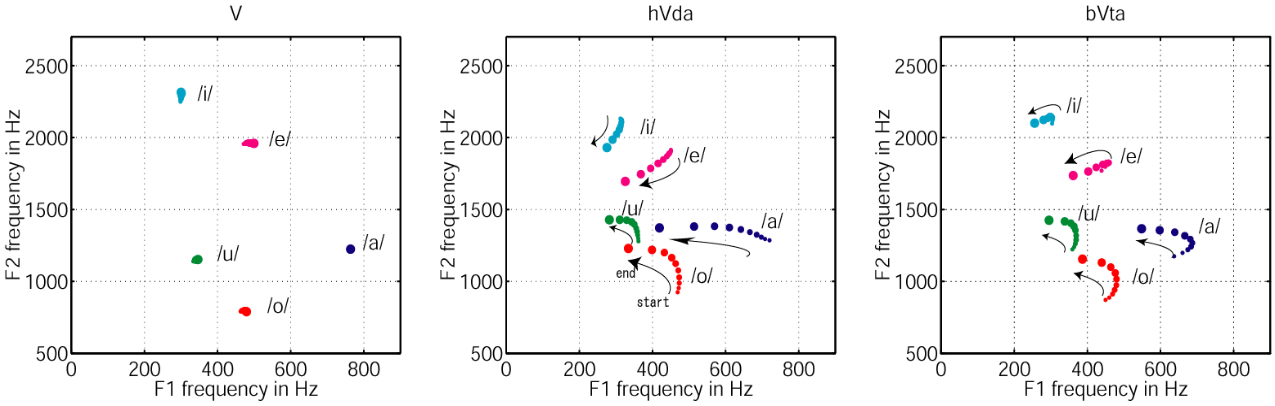

Figure 5. F1 and F2 formant frequencies for every 10% of the duration of short vowels in isolation (left), in /h-V-da/ syllables (middle), and in /b-V-ta/ syllables (right). Data by AM talkers are plotted (N=141).