ACOUSTIC AND PERCEPTUAL PROPERTIES OF PHONEMES IN CONTINUOUS SPEECH AS A FUNCTION OF SPEAKING RATE

Hisao Kuwabara
Department of Electronics and Information Science
Teikyo University of Science & Technology
Uenohara, Kitatsuru-gun, Yamanashi 409-01, Japan
Tel. +81.554.63.4411, Fax. +81.554.63.4431, E-mail: kuwabara@ntu.ac.jp

BSTRACT

An investigation has been made for individual phonemes focusing mainly on their duration in continuous speech spoken in different rates: fast, normal, and slow. Fifteen short sentences uttered by four male speakers have been used as the speech material which comprises a total of 291 morae. Normal speaking rate (n-speech) is, on average, 150 milliseconds/mora (or 400 morae/minute) and the four speakers have been asked to read the sentences twice as fast as (f-speech) and 1/2 times as slow as (s-speech) the normal speed in reference to the n-speech.

Among consonants, the greatest influence has been found to occur on the syllabic nasal /N/ and the least on the voiceless stop /t/ in f-speech. For the s-speech, /N/ has also been found to be the greatest but the least is voiced stop /d/. The ratio of duration between consonant and vowel of a CV-syllable in the f-speech is kept almost the same as that in the n-speech while vowel lengthening becomes significantly large in the s-speech.

As it is expected, formant frequencies of vowels differ significantly between the three rates. Five vowels tend to be close together on the F1-F2 plane as the speaking rate becomes fast reflecting the neutralization of vowels. However, average difference of the third formant has been found to be very small.

1. INTRODUCTION

Recent progress in speech technology has made it possible to build elaborate systems that can process speech signals more precisely than ever before in many technological areas. Not to mention speech recognition and speech synthesis, there are a variety of applications in speech technology area. Keeping this technological background in mind, this study has been conducted as a basic research in order to provide an acoustic data for these speech technologies.

Japanese language basically consists of a series of consonant -vowel syllables (CV-syllables). Unlike English or other languages, each syllable corresponds exactly to one Japanese alphabet which is called "仮名." As it is well known, each syllable in a continuous speech does not carry enough phonetic information to be correctly identified by itself, but rather spread over adjacent phonemes due mainly to coarticulation effects. [1, 2] There are some attempt to recover these reduced ambiguous phonemes. [3, 4] These perceptual evidences must be attributed to such acoustic properties of each phoneme as shortening its duration, reduction of pitch and formant frequencies.

These acoustic properties should, of course, vary from speaker to speaker, from one speaking rate to another. Some studies have so far been made to establish prosodic rules for different speaking rates. [5]

This paper deals with the acoustic properties, focusing mainly on the duration and the formant frequencies, of individual phonemes in a continuous speech for different speaking rate.

2. SPEECH MATERIAL

Fifteen short sentences have been chosen as the speech material. Four male adult speakers who participated in this experiment were asked to read the sentences three times with different speaking rate: normal speech which is referred to as "n-speech" in this paper, fast rate (also referred to as "f-speech") and slow rate ("s-speech").

There is a rhythm when it comes to speak a Japanese sentence. The rhythm, which is sometimes called syllable-timed, is based on the mora which roughly corresponds to a Japanese letter or CV-syllable. The number of morae per minute defines the speaking rate. Generally, normal speaking rate (n-speech) falls into a speed from 300 to 400 morae per minute but it considerably differ from speaker to speaker, especially between the young and the old.

No special guidance and equipment have been used to control the speed in pronouncing the n-speech, f-speech and s-speech. For the f-speech, individual speakers were asked to pronounce the sentences twice as fast as the n-speech that they usually utter in daily conversation. For the s-speech, they were also asked to pronounce half as slow as the n-speech. For each speed, speech data were actually measured later on for speakers individually.

There are 291 morae in the fifteen sentences. Thus, a total of 3,492 (=291morae $\times$ 3rates $\times$ 4speakers) morae have been gathered to be analyzed.

3. MEASUREMENT OF DURATION

Measurements of duration for individual vowels and consonants have been made to investigate how the duration is affected by the speaking rate. There are no clear-cut standard positions to define the beginning and the end of each phoneme in a continuous speech except, for example, for plosive consonants where a silent interval usually precedes. Position of each CV-syllable has been identified first and then consonant- and vowel-parts have been

separated for the measurement of length.
Three criteria have been set in order to define the length of phonemes.

- The beginning of such consonants as fricatives, plosives, and affricates is easily determined by inspecting the waveform.
- When it is inseparable between consonant (or vowel) and vowel, transitional part is defined first, and the distinction should be made at the center of the transition.
- Distinction is made between syllabic nasal /N/ and nasal consonant /n/ or /m/ based on hearing.

UNIX workstations have been used in defining the positions of individual phonemes by inspecting both speech waveforms and spectrograms sometimes with a help of hearing. A total of approximately 7,000 distinctions have been made so far.

3.1. Speaking Rate

Duration data have been pooled for speakers individually and a statistic analysis has been made first, and then these statistic data have been analyzed over all speakers. Silent intervals, though undoubtedly contribute to the speaking rate, are discarded from the analysis this time except those for /Q/.

Figure 1 represents average duration per mora, or approximately one CV-syllable duration, for the three speaking rates. Slow speech tends to be uttered slower than expected, that is, a mora is greater than twice as long as the normal speech (220.7% in reference to the n-speech being n-speech length a 100%).

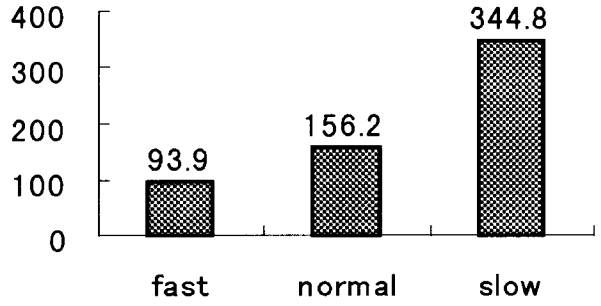

Fig.1 Average duration in millisecond of a CV-syllable for three different speaking rate.

3.2. Consonant versus Vowel

Table 1 depicts the ratios at which an average CV-syllable is divided into consonant- and vowel-parts. Graphical illustration for this is shown in **Figure 2**. For n-speech, the duration of consonant is approximately 1/3 of the total syllable length.

Table 1 Ratio of duration for consonant and vowel parts in a CV-syllable.

	slow	normal	fast
Consonant	24.3%	34.8%	35.7%
Vowel	75.7%	65.2%	64.3%

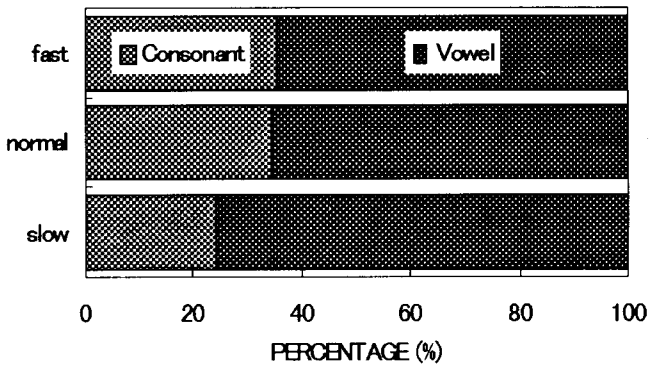

Fig. 2 Consonant versus vowel ratio in a CV-syllable.

An interesting fact is that almost the same ratio is kept for the f-speech as the n-speech while, in s-speech, vowel part becomes significantly large. This means that, when we utter f-speech, consonants and vowels tend to be shortened with approximately the same rate as that for the n-speech but this does not retain any more in the s-speech.

Figure 3 represents the average duration of consonant and vowel for three different speaking rates. **Figure 4** illustrates the ratio of vowel and consonant in reference to the n-speech being the duration for n-speech a 100%.

The two figures clearly show again that, for f-speech, consonant and vowel are approximately the same ratio (66.4% and 63.3%) while, for s-speech, vowel part is significantly lengthened (154.4% Vs 254.3%).

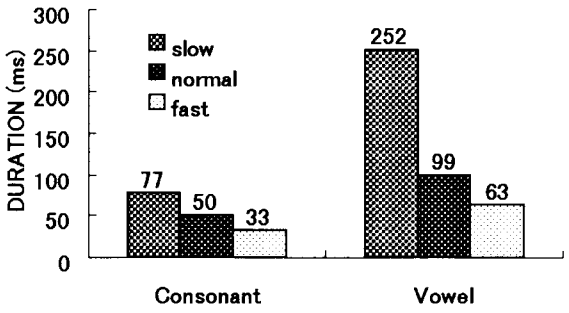

Fig.3 Average duration of consonant and vowel.

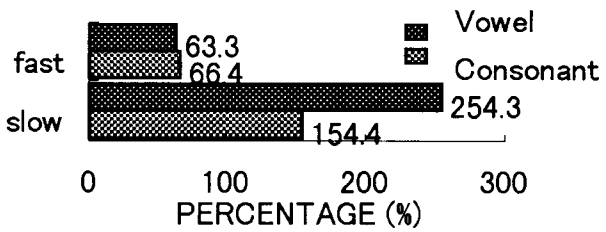

Fig.4 Duration ratio of consonant and vowel in reference to the normal speaking rate.

3.3. Individual Consonants

Duration for individual consonants that have appeared in the speech samples used in this experiment has been analyzed. If we look closely at individual consonants' duration, extremely large ones in every speaking rate can be found for the syllabic nasal /N/ for which the duration reaches as high as 348 milliseconds in the s-speech.

Figure 5 stands for the percentage of individual consonants' duration as compared with the n-speech. Syllabic nasal /N/, which reaches as high as almost 300% lengthening compared to the n-speech, has been eliminated from this figure.

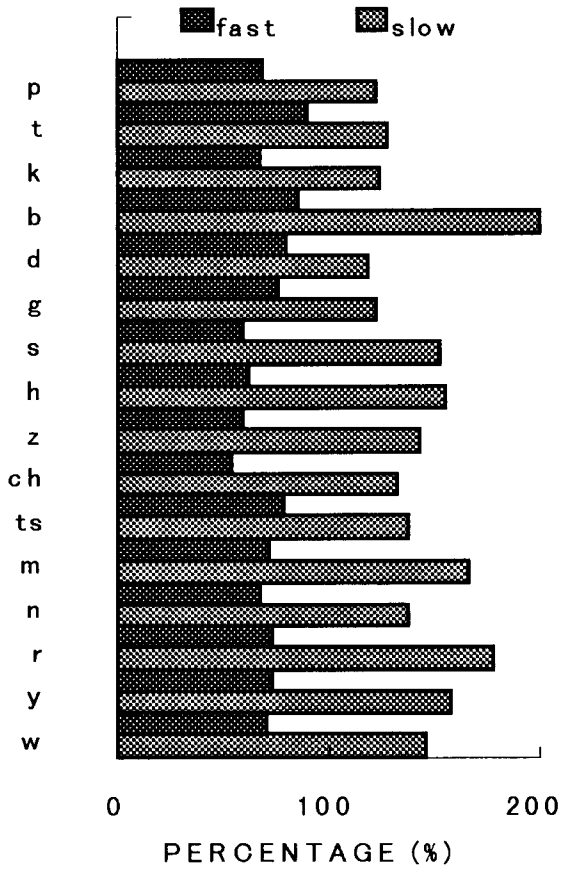

Fig.5 Duration ratio of individual consonants in reference to the normal speaking rate.

An interesting fact is that the voiced plosive /b/ shows a significantly large duration change for the s-speech (about 200%) while other plosives remain relatively small. Liquid consonant /r/, quite different in pronunciation from English retroflex consonant /r/, has also been found to be the second greatest change (about 178%).

Fricative consonants, either voiceless or voiced, show the largest duration change from f-speech to s-speech. Among voiced plosives, duration shortening seems to occur with almost the same rate, but seems different for the unvoiced plosives. Among the unvoiced plosives, /k/ exhibits the least shortening for the f-speech.

4. FORMANT ANALYSIS OF VOWELS

It is obvious that the formant frequencies of a vowel in continuous speech are quite different from the same vowel uttered in isolation due largely to the coarticulation effect. Also, formant frequencies of the same vowel in the same read sentence significantly differ depending on the phonetic context it is in. Formant frequencies of vowels have been analyzed and a comparison has been made between the three speaking rates.

4.1. Formant Measurement

Since the speech samples are digitized at 16kHz with an accuracy of 16bits, formant analysis has been performed within 8kHz frequency range based on the LPC method. Pole frequencies, corresponding each to the first, second, and third formant frequency, have been obtained for each vowel in the test sentences.

Generally, there are no clear-cut acoustic cues that define the whole part of a vowel in a continuous speech. Thus, the center point of a vowel or the point of the closest approach to the target has been determined first by inspecting both speech wave and spectrogram displayed on a computer screen. Based on this point, the whole part of the vowel has been defined taking as long steady portion as possible.

Table 2 represents first three formant frequencies of five vowels averaged over the four speakers. There are 201 vowels in the fifteen sentences: /i/ 31, /e/ 31, /a/ 60, /o/ 54, and /u/ 25. As a total, 2412 vowels (210x3x4) have been analyzed to find three formant frequencies.

All formant frequencies seem to have a clear trend that they either increase or decrease from s-speech to f-speech with a few exception. The trend is particularly true for vowel /a/. Coarticulation effect will certainly play an important role for this formant change for which qualitative discussion will be given later.

Table 2 Average formant frequencies of five vowels for three different speaking rate.

		slow	normal	fast
	F1	280	305	339
i	F2	2182	2114	2031
	F3	3099	3007	2876
	F1	465	455	447
e	F2	1952	1918	1814
	F3	2646	2637	2610
	F1	702	665	631
a	F2	1302	1340	1387
	F3	2688	2641	2619
	F1	451	452	450
o	F2	869	968	1093
	F3	2682	2641	2607
	F1	320	344	361
u	F2	1400	1402	1385
	F3	2412	2463	2480

Figure 6 illustrates the conventional F1-F2 and F1-F3 diagram of five vowels for three different speaking rates. For every vowel group, each point represents formant values averaged over 4 speakers for every speaking rate. The outermost and the innermost points for each vowel

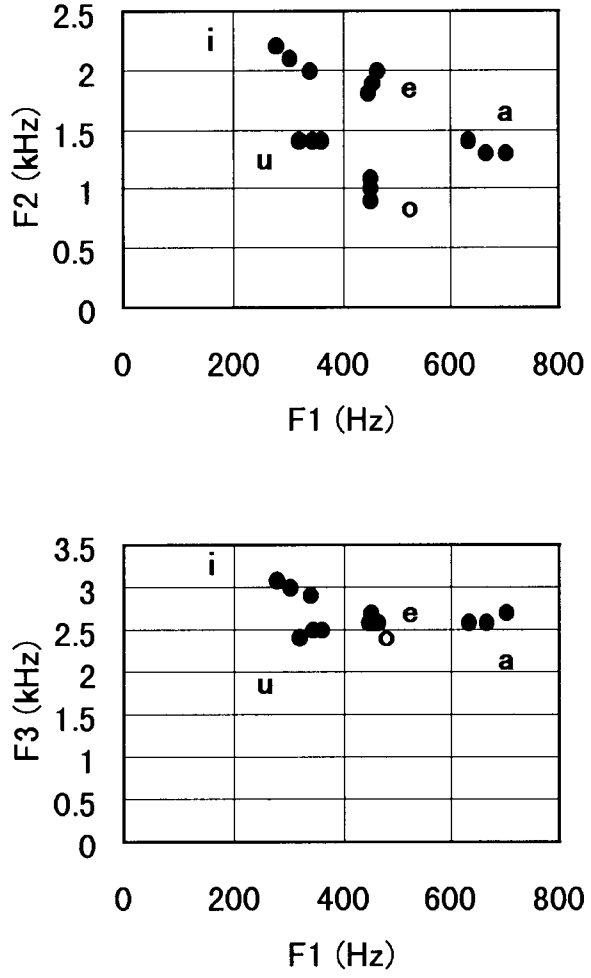

Fig.6 Average first, second and third formant frequencies of five vowels for different speaking rate.

category represent frequencies for s-speech and f-speech, respectively. Those for n-speech lie in-between.

It is clear from the figure, especially from the upper one (F1-F2 diagram), that all vowels tend to move to the center position in the diagram as the speaking rate becomes fast. This trend reflects the coarticulation effect and all vowels go the neutral vowel region as the rate goes fast.

4.2. Maximum Frequency Difference

Some differences can be seen in the vowel movement if we look more closely at the figure. The greatest vowel movement can be observed for /a/ and /i/. Vowel /e/ has the least movement. Vowels /u/ and /o/ are somewhere in between. Among three formants, the greatest change can be observed for the second formant and the third formant shows little change.

Table 3 represents the maximum frequency difference between speaking rates for each formant. Generally, the third formant frequency seems to be affected very little by the speaking rate except for vowel /i/, and the greatest

Table 3 Average maximum difference of formant frequencies between three speaking rates.

	F1 (Hz)	F2 (Hz)	F3 (Hz)
I	59	151	223
e	18	138	36
a	71	85	69
o	2	224	75
u	41	15	68
Average	38	123	94

influence is seen on the second formant especially on vowel /o/. Vowel /o/ has particularly different formant-difference pattern from other vowel. It has almost no frequency change for F1 but the largest for F2. Vowel /u/ receives the least overall influence through the speaking rate among five vowels.

5. CONCLUSIONS

formant frequencies of five Japanese vowels in continuous speech have been measured for different speaking rate. A statistical analysis of duration has also been made for consonant- and vowel-parts. Four adult male speakers read 15 short Japanese sentences with three different speaking rates; normal speed (about 156 ms/mora), fast speed (about 94 ms/mora) and slow speed (about 345 ms/mora).

Vowels receive greater influence than consonants by speaking rate. Among consonants, the greatest influence has been found to occur on the syllabic nasal /N/ and the least on the voiceless stop /t/. The ratio of duration between consonant and vowel of a CV-syllable in the fast speech has been found to be almost the same as that for the normal speech. However, this ratio changes a great deal in the slow speech in which duration of vowel-part becomes extremely large.

As it is expected, formant frequencies of every vowel differ as the speaking rate changes reflecting the co-articulation effect. The greatest change has been found to occur in the second formant while the third formant is the least change. Among the five vowels, the greatest change appear in /i/ and /a/ while the least in /u/. Almost no change in the first formant has been found for vowel /o/.

REFERENCES

1. Fujimura, O., and Ochiai, K "Vowel identification and phonetic contexts," J. Acoust. Soc. Am., Vo.35, 1889(A), 1963
2. Kuwabara, H. "Perception of CV-syllables isolated from Japanese connected speech," LANGUAGE AND SPEECH, Vol.25, 175-183, 1982
3. Lindblom, B.E.F., and Studdert-Kennedy, M. "On the role of formant transitions in vowel recognition," J. Acoust. Soc. Am., Vol.42, 830-843 1967
4. Kuwabara, H. "An approach to normalization of coarticulation effects for vowels in connected speech," J. Acoust. Soc. Am., Vol.77, 686-694, 1985
5. Miyatake,M., and Sagisaka,Y., "Prosodic characteristics and their control in Japanese speech with various speaking styles," IEICE Trans., Vol.J73-D-II, 1929-1935, 1990