PERCEPTUAL AND ACOUSTIC PROPERTIES OF PHONEMES IN CONTINUOUS SPEECH FOR DIFFERENT SPEAKING RATE

Hisao Kuwabara

Teikyo University of Science & Technology
Uenohara, Kitatsuru-gun, Yamanashi, 409-0193, Japan

ABSTRACT

Investigations have been made on the perceptual and acoustic properties of individual phonemes in continuous speech for different speaking rate. Fifteen short sentences spoken by four male speakers have been used as the test material. Each speaker has been asked to pronounce the sentences with three different rates: normal, first and slow.

For perceptual experiment, individual CV-syllables have been taken out from their contexts and presented to listeners in isolation to be identified. The results reveal that individual syllables in continuous speech do not have enough phonetic information to be correctly identified especially for the fast speech. The average identification of syllables for the fast speech is 35% and even vowels are identified less than 60%. Slow speech shows highest identification among the three rates; 86% for the syllables, 87% for the consonants and 91% for the vowels.

Duration of consonants and vowels are both affected by the speaking rate and the latter has been found greater in change. An important finding is that the duration ratio between consonant and vowel of a CV-syllable in the fast speech is kept almost the same as that in the normal speech. Vowel lengthening in the slow speech becomes significantly large. Formant frequencies of individual vowels have largely shifted toward the neutral region in the conventional F1-F2 plane as the rate becomes fast and, at the same time, distribution of vowels in each category becomes large.

1. INTRODUCTION

Speech technologies have been developed a great deal recently and speech signals can be processed more precisely than ever before in such areas as speech recognition and synthesis. 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 "Kana "

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. Our recent studies [5, 6] show that the speaking rate affects very much on the acoustic values of phonemes in continuous speech, such as duration and formant frequencies, which partly included again in this paper.

This paper deals mainly with the perceptual properties of individual CV-syllables and vowels when they are taken out of their phonetic environment. Comparisons are also being made between the perceptual results and the acoustic values of individual syllables.

2. SPEECH MATERIAL

The speech material that has been used in this experiment is the same as in the literature 5 and 6, which consists of fifteen short sentences uttered by four male adult speaks. As it is mentioned in the literature, they were asked to read the sentences three times with different speaking rate: normal speed 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.

3. PERCEPTUAL EXPERIMENT

This experiment is designed to investigate on how each CV-syllable or vowel will be perceived by human listeners when it is isolated from its phonetic environment. Thus, each CV-syllable or vowel must be electrically cut off from the stream of speech. It is obvious that the coarticulation effect will certainly

depend on the rate of speaking. The purpose of this experiment is to find out how large the effect will be in terms of phoneme identification for every speaking rate.

3.1. Test Material

There are 291 morae in the fifteen sentences. Among the four speakers' utterance, one speaker's speech is used this time since the main purpose of this experiment is to compare the perceptual difference of individual syllables between speaking rate. The remaining three speakers speech is left for further investigation to find difference among speakers.

There is no clear-cut definition to divide syllables in a running speech unless a silent interval exists between tow successive syllables. It is almost impossible to draw a line to separate syllables if the speech wave of two successive syllables continues without silence. So, we have decided to separate these syllables by audition with the help of a speech-wave editor on a computer. The beginning and the end of a syllable are defined as follows. At first, a piece of speech wave which roughly corresponds to three syllables with the syllable in question in the middle is selected on the computer screen. Then, the starting point of the speech wave is cut off step-by-step until the entire preceding syllable is not audible. This defines the beginning of the syllable in question. Next, the start-point of the syllable is fixed and similarly, the end point is cut off step-by-step until the successive syllable becomes not audible, which defines the end of the middle syllable. It is somewhat complicated and time consuming. In this way, a total of 873 (291x3) test materials have been obtained for perceptual experiment.

3.2. Hearing Test

For each speaking rate, the 291 speech material are randomized and presented to listeners over a loudspeaker in a sound-proof chamber. Six listeners, including the author, participated in the hearing test and they were asked to identify each piece of speech sound as one of the Japanese syllables or vowels. Listeners' response data were grouped in the following three ways.

1. Syllable identification: whether the entire CV-syllable was correctly identified or not.
2. Consonant identification: whether the consonant part was correctly identified or nor.
3. Vowel identification: whether the vowel part was correctly identified or not.

For the single vowels in the sentences, consonant perception was not counted and the syllable identification was counted only if it was perceived as a single vowel. If a stimulus /a/ was identified as /pa/, for example, the syllable identification is regarded as incorrect but the vowel identification is correct. The syllable identification will be correct only when it is perceived as /a/.

3.3. The Results

	fast	normal	slow
syllable	35	59	86
consonant	31	58	87
vowel	59	86	91

Table 1 *Percentage of average identification score for CV-syllables, consonant and vowel parts.*

3.3.1. Average Syllable Identification

Table 1 shows the average percent of identification in the three categories. Identification of individual syllables is very poor, as low as 35%, for the f-speech. Even the vowel part has been identified less than 60%. It has increased to 59% for the n-speech indicating that the individual syllables in the n-speech do not have enough phonetic information to be correctly identified . However, it goes up to as high as 86% for the s-speech and almost perfect identification has been achieved.
Figure 1 shows a graphic illustration of the average

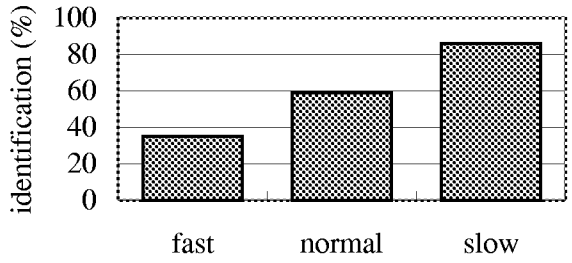

Figure 1 *Average identification for CV-syllables taken out of their phonetic environments for each speaking rate.*

identification of CV-syllables isolated from their phonetic environment for every speaking rate. Identification score seems to increase almost linearly as the rate becomes slow. No drastic

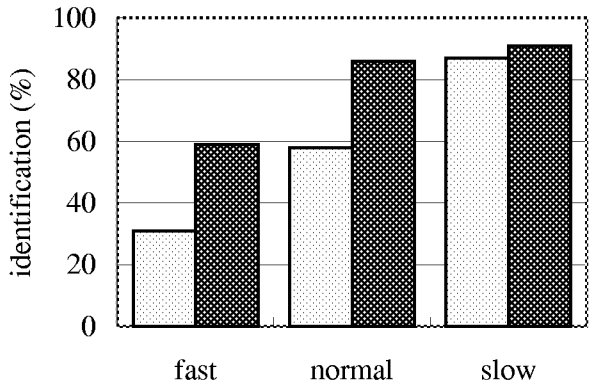

Figure 2 *Average identification scores for consonant part (left bar) and vowel part (right bar) in each CV-syllable.*

jump in identification can be observed from fast to normal or from normal to slow, but essentially, little coarticulation effect is taking place in the slow speech. This might be a speaker dependent phenomenon and further investigation will be needed

3.3.2. Average Consonant and Vowel Identification

.**Figure 2** represents the average identification scores for consonant and vowel parts for every speaking rate. Relatively high scores for vowel part for every speaking rate can be seen from the figure but still less than 60% for the fast speech. Consonant, on the other hand, is extremely low for the fast speech and also for the normal speech while it goes almost as high as the vowel part for the slow speech. In other words, slow speech has almost no co-articulation effect in terms of phonetic information to be identified by humans for at least this speaker. An interesting fact is that the fast and normal speech have a common feature that the identification differences between consonant and vowel parts are exactly the same (28%) though their absolute values of individual parts are different.

Vowel part is identified more than 80% for both normal and slow speech while it is less that 60% for the fast speech. This contrasts to the consonant score which increases almost linearly from fast to slow, 27% increase from fast to normal and 29% increase from normal to slow. Another interesting fact is that the scores of vowels for the fast speech and consonant for the normal speech is almost the same. No immediate implication and explanation are given here toward this fact.

3.3.3. Average Identification for Consonant Category

Let's look more closely at the results with every consonant category. **Figure 3** stands for the result which is compiled according to every manner of articulation. They are voiceless plosives, voiced plosives, affricates, voiceless fricatives, voiced fricatives, nasals, liquids, and semi-vowels. As a whole, semi-vowel has the highest score for all speaking rate. If we focus our attention on the fast speech, it can be observed that the least identification scores are voiced plosives and voiceless fricatives. Almost all voiceless fricative samples are perceived as an affricate which indicates that there is an artifact of cutting off the very beginning of the noise release for each fricative samples. Voiced plosive samples, on the other hand, are found to be confused within the same consonant category and /d/ samples are very often perceived as /r/ sound. Generally, identification goes up as the speaking rate goes down and, for slow speech, around 80% or higher scores are seen to be achieved. An interesting fact is that the identification of the voiced fricatives and liquid consonants for the fast and normal speech show almost the same score. No relevant explanation for this can found yet.

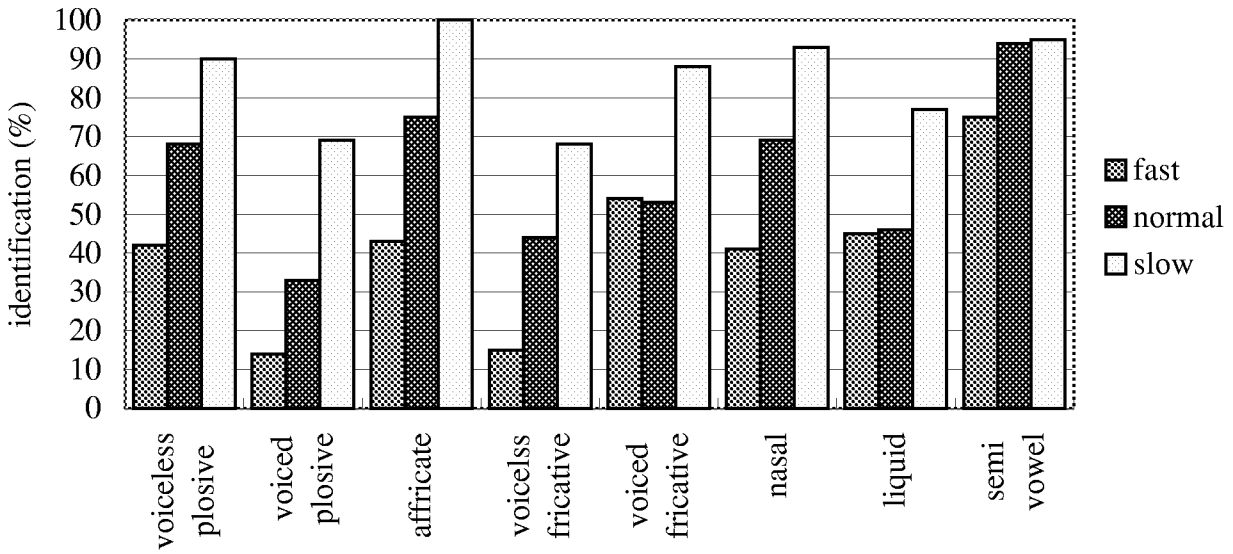

Figure 3 *Consonant identification for eight different manner of articulation.*

For voiceless/voiced plosives and affricates, a significant increase in identification is observed either from fast to normal or from normal to slow. This probably reflects the fact that duration of consonants increases as the speaking rate decreases. For each speaking rate, voiced plosives show the lowest score of identification while semi-vowels exhibit the highest score with an exception of affricates for the slow speech (100%).

3.3.4. Identification for Individual Consonants

Let's look the response data more closely for each consonant. **Figure 4** represents the result for 19 individual consonants that have appeared in the test material. For each consonant, the top bar stands for the result for the fast speech, second bar for the normal speech and the bottom bar for the slow speech.

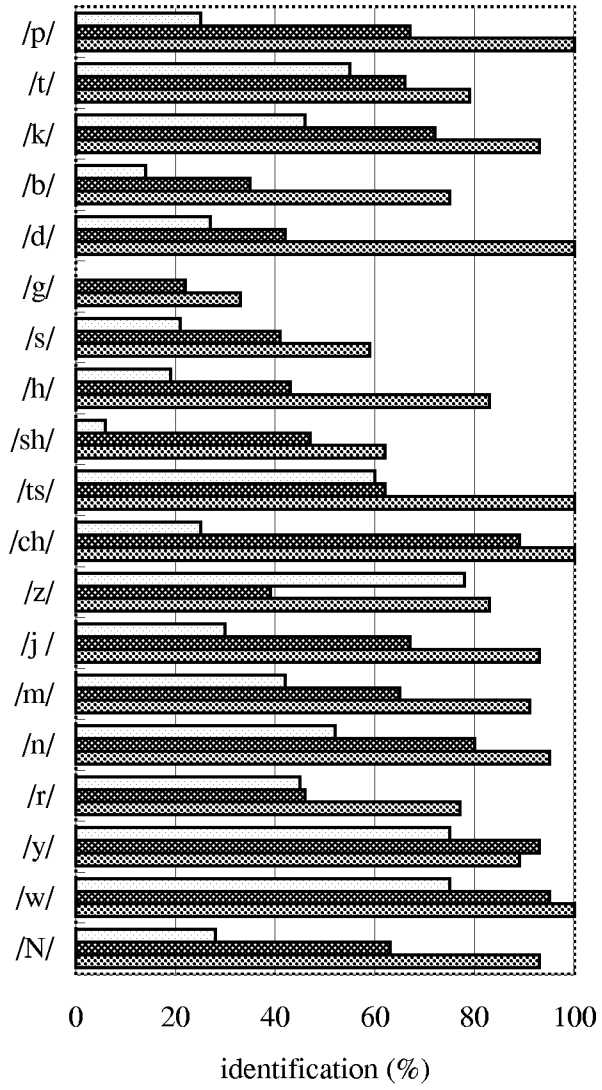

Figure 4 *Consonant part identification for individual consonants for the fast (top bar), normal (middle bar) and slow speech (bottom bar).*

It clearly shows that the lowest score is observed for the voiced plosive /g/ for all speaking rate. /g/ is very often perceived as a single vowel and one of another plosives. On the other hand, the highest score for all speaking rate is obtained for semi-vowel /w/. Semi-vowel /y/ shows almost as high scores as /w/ sound but a little lower for the slow speech. Voiceless fricative /sh/ has the least percentage for the fast speech. Almost all /sh/ samples were perceived as /ch/ especially in the fast speech because of artifact. A strange phenomenon is that a voiced fricative /z/ changes its score between fast and normal, i.e. score for the fast speech exceed significantly that for the normal speech. One reason for this is that the way of speaking of this speaker. This speaker often pronounces /s/ sound as /z/ sound in the fast speech but not in the normal and slow speech.

There is an interesting contrast if we compare scores between bilabial and alveolar consonants, such as /p/ vs /t/, /b/ vs /d/ and /m/ vs /n/. It is observed that the alveolar consonants are always higher than the bilabial.

4. CONCLUSION

A perceptual experiment has been performed to investigate how individual CV-syllables, which are the basic constituent of the Japanese spoken language, carry phonetic information and how they differ when the speaking rate changes. Individual CV-syllables are taken out from their phonetic environments for three different speaking rates, fast, normal and slow and presented to listeners for identification.. It has been found that individual syllables carry very little phonetic information for the fast speech (35%), fairly low for normal speech (59%) and high for the slow speech (86%). The CV-syllable identification is determined almost entirely by consonant perception.

Identification for consonant and vowel parts in a CV-syllables are collected separately. It has been found that the consonant identification increases with almost the same ratio as the speaking rate downs. However, the identification of the vowel part shows quite different pattern. It significantly increases from fast to normal but little from normal to slow. For slow speech, consonant perception, and hence syllable perception , shows very high identification score. This experiment is done only for one speaker, and the results will include those which may reflect speaker's individuality. Further investigations will be needed to offset this sort of speakers variability.

5. REFERENCES

1. Fujimura, O., and Ochiai, K "Vowel identification and phonetic contexts," J. Acoust. Soc. Am., Vo.35, 1889 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 co-articulation effects for vowels in connected speech," J. Acoust. Soc. Am., Vol.77, 686-694, 1985

5. Kuwabara, H. "Acoustic properties of phonemes in continuous speech for different speaking rate," Proc. of ICSLP, 2435-2438 1966

6. Kuwabara, H. "Acoustic and perceptual properties of phonemes in continuous speech as a function of speaking rate," Proc. of EUROSPEECH, 1003-1006 1997