Perceptual Properties of Syllables Isolated from Continuous Speech for Different Speaking Rate

Hisao Kuwabara

Teikyo University of Science & Technology
Uenohara, Kitatsuru-gun, Yamanashi, 409-0193, Japan
kuwabara@ntu.ac.jp

Abstract: An investigation has been made on the perceptual nature of CV-syllables taken out from a running speech and their acoustic characteristics. Fifteen short Japanese sentences uttered by four male speakers with three different speaking rates, fast, normal, and slow, have been used. Syllable identification for speech segments taken out from a running speech has been made in three different ways: 1) one-syllable segmentation, 2) two-syllable segmentation, and 3) three-syllable segmentation. In the one-syllable segmentation, individual syllables have been taken out from their phonetic environments and presented to listeners for their identification. In the two- and three-syllable segmentations, every two and three successive syllables have taken out and the listeners have been asked to identify them as a whole. In the one-syllable segmentation experiments, the average syllable identifications for the fast, normal, and slow speech are 35%, 59%, and 86%, respectively. The result reveals that individual syllables for the fast and normal speech do not have enough phonetic information to be correctly identified, but for the slow speech it retains fairly well. Phonetic information for a syllable is not sufficiently preserved in the two syllable segment (two-syllable segmentation experiment) especially for the fast speech. However, the middle syllable in the three-syllable segmentation has been found to carry enough phonetic information to be correctly identified even for the fast speech. A relation between the perceptual results and the acoustic properties has been discussed.

1 Introduction

Japanese language, both in written and spoken, basically consists of a series of consonant-vowel syllables that we refer to as CV-syllables in this paper. Unlike English or other languages, each syllable corresponds exactly to one Japanese alphabet called “Kana” and this is also a basic unit of pronunciation. In early days of speech technology research, this language structure seemed to be beneficial for developing a continuous speech recognition scheme of unlimited vocabulary for Japanese. It was soon realized, however, that the problem was not as easy task to solve as it was thought first. As it is well known, a syllable in continuous speech does not carry enough phonetic information to be correctly identified by itself, but rather spread over adjacent phonemes due mainly to the so-called coarticulation effects

There are some attempts to recover these reduced ambiguous phonemes [1], [2]. These perceptual evidences must be attributed to such acoustic properties of each phoneme as shortening its duration, reduction of pitch and formant frequencies and so on. These acoustic properties should, of course, vary from speaker to speaker, from one speaking rate to another.

This paper deals mainly with perceptual properties of individual syllables segmented from continuous speech with different speaking rate to find out how and to what extent does the speaking rate affect to the phonetic information of individual syllables. A possible link has been discussed between the perceptual results and the acoustic properties of vowels and consonants that were already reported before [3].

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×3rates×4speakers) morae have been gathered to be analyzed.

3 Experimental Procedure

This experiment was designed to investigate on how each CV-syllable or vowel would be perceived by ordinary listeners when it was isolated from its phonetic environments. Among the four speaker's utterance, one speaker's speech was used this time since the objective of this experiment was to compare the difference of syllable identification between three speaking rates. The remaining three speaker's speech was left for further investigation to examine speaker's difference.

3.1 Syllable Segmentation Scheme

Each CV-syllable or vowel must be electrically cut off from the stream of running speech. It is obvious that the level or magnitude of coarticulation effect will certainly depend on the rate of speaking. The faster the speaking rate, the larger the coarticulation effect. The purpose of this experiment is to examine on how the speaking rate affects to the coarticulation in terms of phoneme identification of individual syllables. Three types of segmentations, 1) one-syllable segment, 2) two-syllable segment, 3) three-syllable segment, have been performed in this experiment.

There is no clear-cut definition to divide syllables in a running speech unless a silent interval exists between two successive syllables. It is difficult to draw a line to separate syllables if the speech wave continues. So the syllable boundary is defined by audition with a help of speech wave and spectrogram on a computer.

One-Syllable Segment. A total of 291 syllables have been separated individually from the 15 test sentences for each speaking rate to present to listeners. There are 873 (291x3) test materials as a whole.

Two-Syllable Segment. Every two successive syllables have been taken out from the continuous speech to present to listeners. Each speech piece, therefore, consists of two successive syllables. The syllable boundaries, i.e. the starting and end points of each speech piece are the same as those in the one-syllable segment.

Three-Syllable Segment. Every successive three syllables have been segmented from speech. Again, the syllable boundaries of each speech piece are the same as in the one-syllable segment.

3.2 Syllable Identification

For each segment, speech pieces were randomized and presented to listeners over a loudspeaker in a sound-proof room. Six listeners, including the authors, participated in the hearing test. They were asked to identify each speech sound as one of Japanese syllables. For two-syllable or three-syllable segments, they were required to identify each of the two or three syllables in a speech piece simultaneously.

Listeners response data were pooled and assembled in the following three ways.

- Syllable identification: whether the entire CV-syllable is correctly identified or not.
- Consonant identification: whether the consonant part is correctly identified regardless of the vowel part.
- Vowel identification: whether the vowel part is correctly identified or not regardless of consonant part

4 Experimental Results

Results of perceptual experiment were pooled and arranged for each of one-, two-, or three-syllable segment separately. For two-syllable segment, there are two syllables in each speech piece, and three syllables in three-syllable segment.

4.1 One-Syllable Segment

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.

Table 1 Percentage of average identification for syllables, consonant and vowel parts

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

Figure 1 shows a graphic illustration of the average identification of CV-syllables together with its consonant part C and vowel part V isolated from their phonetic environment for every speaking rate. Identification score seems to increase almost linearly as the rate becomes slow. No drastic 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. It seems that a syllable identification can almost be determined by its consonant-part identification.

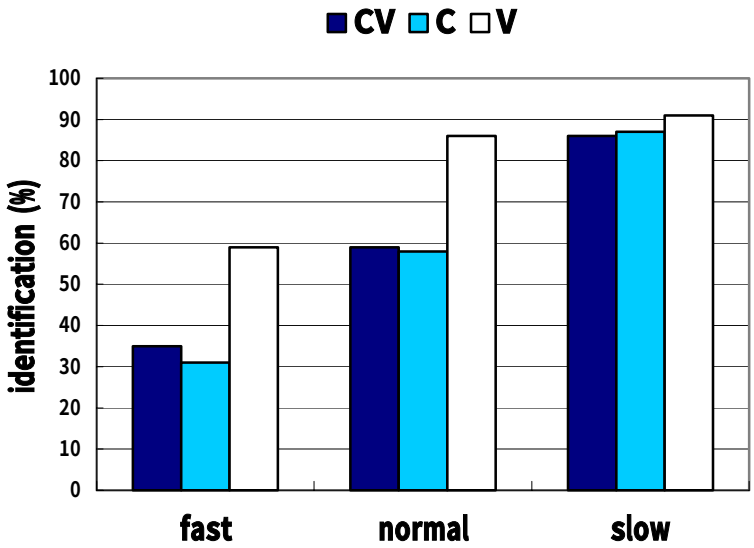

Fig. 1 Identification of CV-syllables for 1-syllable segment. Consonant (C) and vowel parts (V) are displayed.

Very high identification score can be found in the s-speech samples. This might be a speaker dependent phenomenon and further investigation will be needed. To look at the result more closely with every consonant category,

Figure 2 depicts the result compiled according to every manner of articulation. As a whole, semi-vowel has the highest scores for all speaking rates. In the f-speech, the least identification scores can be found in the voiced plosives and the voiceless fricatives. Almost all voiceless fricatives are perceived as an affricate indicating an artifact of cutting off the very beginning of the noise release for each fricative samples. Voiced plosives, 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 s-speech, around 80% or higher can be observed.

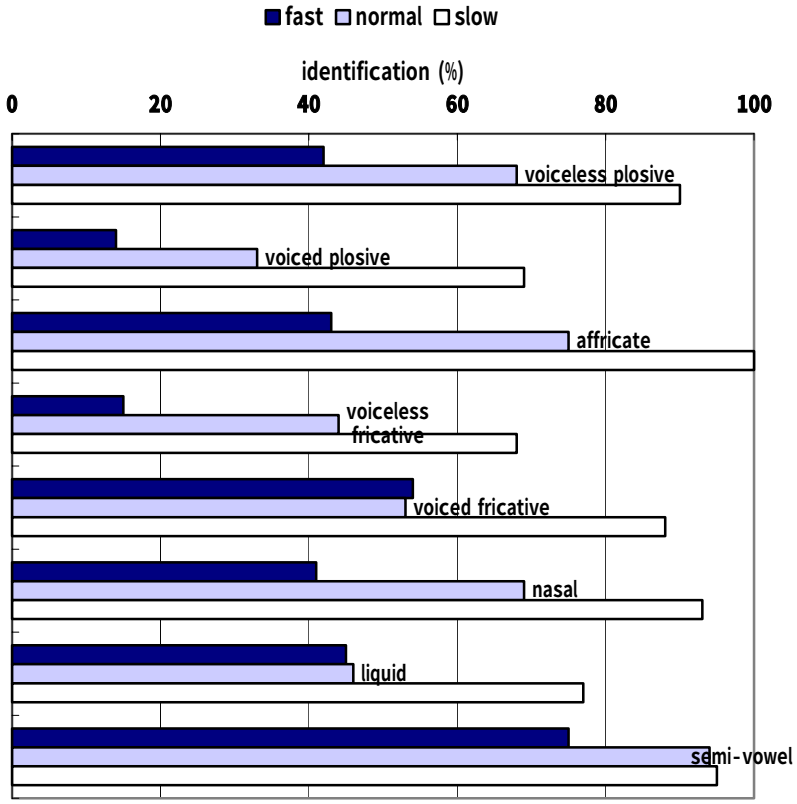

Fig.2 Average identification for each consonant category for three different speaking rate.

4.2 Two-Syllable Segment

For 2-syllable segment, listeners were asked to identify two syllables simultaneously for each stimulus. Identification scores for the first and second syllables were calculated separately. **Figure 3** (1) represents the result for the first syllable, and (2) for the second (last) syllable. As it is expected, the second (or last) syllable shows higher identification scores as a whole. It is interesting to note that the syllable identification scores for the first syllable do not vary very much across the three speaking rate. In the syllable identification, an average of 54% has been obtained for the f-speech in the first position of 2-syllable stimuli, which is quite low, and 73% in the second position, still low yet. For n- and s-speech, however, nearly 90% identification has been achieved when the syllables are in the second position.

If we look at the results more closely, we find that, in the first syllable identification, f-speech is significantly lower in all CV, C, and V categories than the other two speaking rates. In the second syllable position, however, scores for those f-speech increase to a large extent. This is understandable because a large amount of phonetic information is kept for syllables in the second position. It is quite interesting to note that, in C and V identification, correct identification goes up as the rate becomes slow despite the fact that every truncated-signal begins with the same position. It is quite natural that correct identification increases dramatically for all speaking rates when the syllables in question are in the second position. This is because, in the second syllable position, the phonetic information is kept greater than the syllable

in the first position.

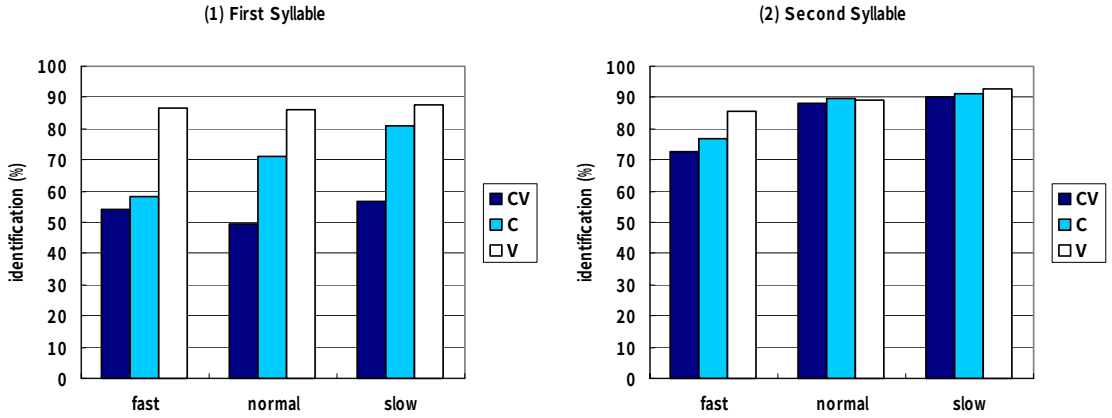

Fig. 3 Average syllable identification for 2-syllable segment speech, (1) for the first syllable and (2) for the second syllable

4.3 Three-Syllable Segment

Listeners were also asked to make a simultaneous syllable identification for each 3-syllabel stimulus and the responses were arranged separately according the syllable position. **Table 2** represents the percentage of syllable identification.

Table 2 Percentage of average syllable identification for 3-syllable segment stimuli.

	fast	normal	slow
1st syllable	68	77	78
2nd syllable	91	97	98
3rd syllable	84	94	96

It is quite obvious from the result that the second syllables for all speaking rate show the highest score. **Figure 4** illustrates the results in the same way as in Fig. 3. The figure clearly shows that the second syllable has got the highest identification score for all speaking rate. This indicates that the phonetic information of a syllable is not lost as long as it is in the middle position of a three-syllable segment. It is interesting to note that, in the first syllable identification, n-speech and s-speech have almost the same

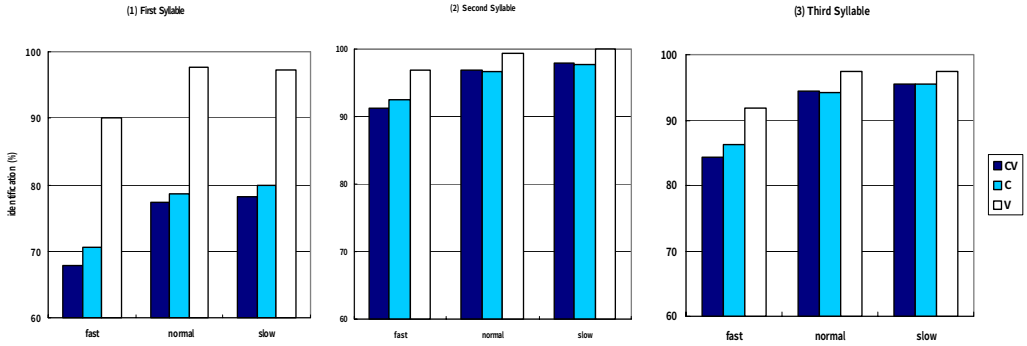

Fig. 4 Average syllable identification for 3-syllable segment speech, (1) for the first syllable, (2) the second

pattern while f-speech has slightly different pattern. Even for the f-speech, syllable identification has reached higher than 90% when the syllable is in the middle position.

5 Discussion and Conclusions

Let's re-arrange the perceptual data so that we can explicitly find the influence of segmentation-length on the syllable identification. **Figure 5** represents the re-arranged data to show how the length – the number of syllables – could effective to increase the correct identification. This is the result for the f-speech which is more visible of the influence than the other two speaking rate. Though individual syllables (1-syllable segment) have only 35% identification in average, it goes as high as 91% when it is in the middle position of 3-syllable segment. This implies that phonetic information of a syllable is spread over adjacent syllables. It is interesting to observe that the identification for first syllable gradually increases as the segment length becomes long despite the fact that it starts from the same position across the three segments.

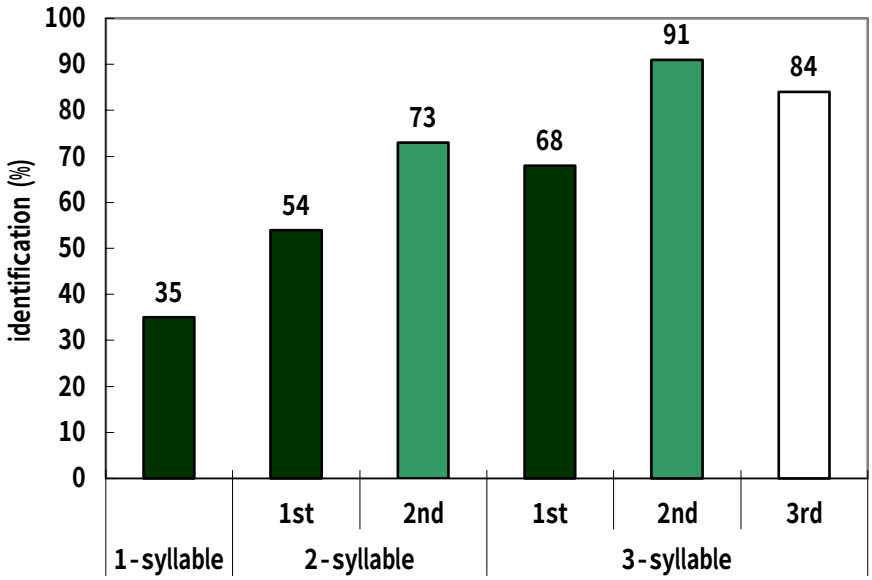

Fig. 5 Average syllable identification for the f-speech rearranged to see the influence of segment length.

Some considerations have been made to relate these results with acoustic features that have been reported before. Formant frequencies, very much reduced for f-speech vowels, are responsible to the results to a certain extent. Duration ratio between consonant and vowel parts in a CV-syllable has some link to the results.

References

[1] Lindblom, B.E.F. and Studdert-Kennedy, M., "On the role of formant transitions in vowel recognition," J. Acoust. Soc. Amer., Vol.42, 1967, pp.830-843

[2] Kuwabara, H. "An approach to normalization of co-articulation effects for vowels in connected speech," J. Acoust. Soc. Amer., Vol.77, 1985, pp.686-694

[3] Kuwabara, H. "Acoustic properties of phonemes in continuous speech for different speaking rate," Proc. ICSLP, 1996, pp.2435-2438