Experimental Study on the Role of Prosodic Features in the Human Process of Spoken Word Perception

Keikichi Hirose, Nobuaki Minematsu and Mika Ito
Department of Electronic Engineering, Faculty of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

ABSTRACT

As a step toward the clarification and formulation of the human process of speech perception through prosodic features, perceptual experiments were conducted on the transmission of accent types as well as on the transmission of word meaning using synthetic speech of Japanese words of four morae. As for the transmission of word meaning, the role of prosodic features for word identification was found to be largest for type 1 accent. This is because the type 1 accent has a falling in the fundamental frequency contour at the beginning portion of the word and, therefore, the prosodic features can be utilized before the perception of whole segmental features. As for the transmission of accent types, several results were obtained implying the existence of the process for accent-type identification aside from the process for perceiving segmental features.

INTRODUCTION

Although the segmental features of speech acoustically play the dominant role in the human process of speech recognition, the prosodic features may also play an important role. This is because the prosodic features of speech are tightly related to the linguistic information of an utterance, such as the word meaning, the syntactic structure, and the focal condition. With the premise that the knowledge on the human process of speech perception should be incorporated in the machine systems for speech recognition to improve their current performance, we have been conducting several perceptual experiments to clarify and to formulate the process (Fujisaki, Hirose, Ohno and Minematsu 1990, Minematsu, Ohno, Hirose and Fujisaki 1992). These experiments, however, were restricted to the segmental features of speech and did not cover the prosodic features.

From the viewpoint above, as a first step toward the clarification and formulation of the process of speech recognition through prosodic features, we have conducted perceptual experiments on the transmission of accent types as well as on the transmission of word meaning using synthetic speech of 4-mora words of Japanese with their fundamental frequency contours manipulated. In this paper, each of both experiments is separately explained followed by the discussion on the results.

WORD ACCENT OF JAPANESE

For an n-mora word of Japanese, $(n+1)$ accent types are possible in the Tokyo dialect. These are denoted by "type i" accents (i=0 to n). Each of type i accents has a rapid downfall in the fundamental frequency contour respectively at the end of ith mora except for the case of i=0 without apparent downfall. When uttered in isolation, type n accent has a fundamental frequency contour similar to that of type 0 accent. For the current perceptual experiments, isolated utterances of 4-mora words were used with type 4 accent excluded. Utterances of a male speaker were recorded and digitized with 10 kHz sampling frequency and with 12 bit accuracy for the further process. Stimuli for the perceptual experiments were generated by the PARCOR analysis-synthesis method with manipulations on the fundamental frequencies. For the current experiments, the manipulations were conducted only on the fundamental frequency contour and no on other prosodic parameters, such as the syllable duration and the source power. This is because of the priority of the fundamental frequency contour in the acoustic manifestation of the prosodic features of Japanese speech.

ROLE OF ACCENT TYPES FOR THE IDENTIFICATION OF SPOKEN WORDS

Method of experiment

As shown in table 1, utterances of 12 nouns were selected for each of the accent types 0, 1, 2 and 3. Following three types of manipulation were conducted on the fundamental frequency contours during the process of PARCOR analysis-synthesis:

(Case 1) Keeping constant at 100 Hz,
(Case 2) alternating into other accent types,
(Case 3) with no modification.

Manipulation for case 2 was performed based on the model of fundamental frequency contour generation (Fujisaki and Hirose, 1984). Fundamental frequency contours of alternated accent types were generated by the model after shifting the onset and the offset of the accent command to their typical values. A band elimination of 0.5 kHz to 3.0 kHz was further performed for all of the synthetic speech samples so that the subjects of the perceptual experiment may perceive a sample as a whole. The syllable-based recognition using the segmental features is difficult for the band-eliminated speech stimuli. These stimuli were presented through headphones with 4 sec inter-stimulus interval to 10 male subjects of Japanese who were asked to reproduce the words orally. The experiments were conducted for the three cases shown above in the order of 2, 1 and 3.

Type 0 to Type 1:	"raion"	"akabou"	"ninjin"	"naiyou"
Type 0 to Type 2:	"shingou"	"omatsuri"	"yokujitsu"	"amerika"
Type 0 to Type 3:	"hiroshima"	"aimai"	"orugan"	"raihin"
Type 1 to Type 0:	"nekutai"	"koumori"	"wakuchin"	"randamu"
Type 1 to Type 2:	"naitaa"	"monoraru"	"amazon"	"unsei"
Type 1 to Type 3:	"kamakiri"	"unmei"	"ookami"	"enbun"
Type 2 to Type 0:	"imomushi"	"norimaki"	"omusubi"	"yononaka"
Type 2 to Type 1:	"mimizuku"	"katakori"	"onigiri"	"nodoame"
Type 2 to Type 3:	"aomori"	"toraburu"	"murasaki"	"origami"
Type 3 to Type 0:	"kamisori"	"tamanegi"	"nokogiri"	"machigai"
Type 3 to Type 1:	"kaminari"	"nissuu"	"nenryou"	"nonbiri"
Type 3 to Type 2:	"kaminoke"	"nakigoe"	"noumiso"	"tenkizu"

Table 1. *Four-mora words used for the experiment.*

Experimental results and considerations

Figures 1 and 2 show the word recognition rate separately for the three cases. The recognition rates of Fig. 1 are calculated for each original accent type, while those of Fig. 2 are calculated for each accent type after the case 2 alternation. Case 2' indicates the rates of accent-type recognition. In both figures, the recognition rate of each accent type has a similar value for case 3 samples without modification in fundamental fre-

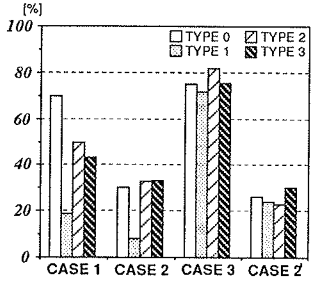

Figure 1. *Word recognition rates summarized separately for each original accent type.*

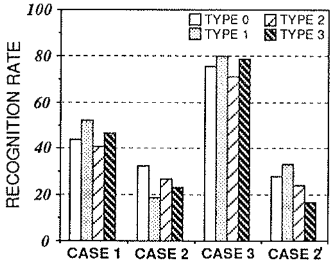

Figure 2. *Word recognition rates summarized separately for each accent type after alternation.*

quency contours. As clearly indicated in Fig. 1, the largest drop in the recognition rate due to the accent type alternation is observed for the words with type 1 accent. In Fig. 2, the largest drop is observed when non–type 1 accent is altered to type 1 accent. These results imply the greater role of prosodic features for the perception of words with type 1 accent. The recognition rate of accent types has the largest score for the samples with type 1 accent as shown in Fig. 2. This result may support the above findings on the role of type 1 accent for the word identification process. The reason of the findings is considered to be due to the fact that type 1 accent has a falling in the fundamental frequency contour at the beginning portion of the word and, therefore, the prosodic features can be utilized before the perception of whole segmental features. Supposing the importance of the prosodic features for the word recognition is affected only by the location of the downfall in the fundamental frequency contour, the role should decrease in the order of types 2, 3 and 0. No result, however, was found supporting this hypothesis. This implies the existence of the perceptual mechanism which makes possible to recognize the word only with the acoustic features of the first half of the utterance.

PERCEPTION OF WORD ACCENT

Method of experiment

Utterances of 4–mora words of type 3 accent were selected for the experiment on the transmission of accent types. Three Japanese words "aozora (blue sky)," "genshiro (atomic reactor)," "korigori (have had enough and never do it again)" were selected together with three nonsense words "imeyuro," "nemeira," "ominere." For each of these words, synthetic speech samples were prepared with accent types 0, 1, 2, 3 and 6 artificial accent types not found in the Tokyo dialect. As shown in Fig. 3, the same rising/falling pattern in fundamental frequency contour was realized at the boundary of 3rd and 4th morae for each sample. The following two types of gating techniques were then applied to the samples to produce stimuli for the perceptual experiment:

(Case A) retaining the initial portion of x msec and replacing the rest by silence,

(Case B) adding to case A, replacing the portion of first and second morae by silence. These replacements are shown schematically in Fig. 4. Case B was planned to investigate the effect of initial part of an utterance for the accent–type identification.

The word stimuli were presented through headphones to the subjects in several sessions for each of 6 words. Each session includes 10 stimuli, viz., one for each of 10 accent types. The gating duration x for stimuli in the first session was set equal to the total duration of the first to the third morae and was increased in 5 msec steps. The experiment was conducted firstly for the stimuli of case A with an inter–stimulus interval of 6 sec, and was then conducted for those of case B with an inter–stimulus interval of 2 sec. The subjects were asked to reproduce the accent types by schematically drawing them on a sheet of paper specially prepared. Based on their answers, necessary duration x was decided for the identification of the accent types.

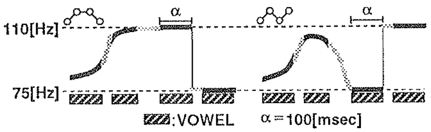

Figure 3. *Method for manipulating the fundamental frequency contours.*

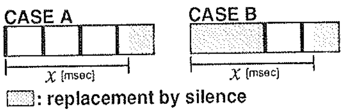

Figure 4. *Schematic illustration for the two cases of silence replacement.*

Experimental results and considerations

The results are shown in Figs. 5 and 6 separately for 3 Japanese words and 3 nonsense words. The ordinate denotes the length of 4th morae in the gating period. As for the case B, the accent types are identified with the shorter length for the words with rising fundamental frequency contour at the boundary of 3rd and 4th morae than those with falling fundamental frequency contour. This result indicates the rising is perceived faster than the falling. As for the case A, the length for accent type identification is shorter for the Japanese words than for the nonsense words. If we compare the results of the two cases for the words of accent type 3, the downfall in the fundamental frequency contour is shown to be perceived for shorter gating period in case A than in case B. This fact is found both for the Japanese words and the nonsense words. Assuming that human may have no inner dictionary for nonsense words, the above fact implies that an accent dictionary of known accent types exists aside from the ordinal word dictionary with information on part of speech, meaning and others, and that a process exists to perceive input spoken words as if they are accompanied by one of the known accent types. A pointer may exist from each item of the accent dictionary to each item of the inner word dictionary.

The above hypothesis was also supported by another preliminary experiment where the spoken word stimuli with original and alternated accent types were presented to the subjects in isolation and continuously. A larger drop in the rate of word identification due to the continuous presentation was observed for the stimuli with alternated accent types than for those with original accent types. Incorrect pointing to the inner word dictionary may occur in the case of alternated accent types and may largely degrade the performance of word identification based on the segmental features typically in the case of continuous presentation.

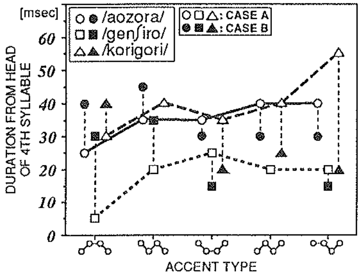

Figure 5. *Duration of 4th mora necessary for the identification of accent types for Japanese words "aozora," "genshiro," and "korigori."*

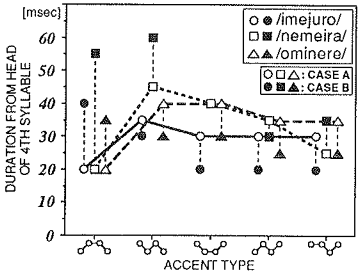

Figure 6. *Duration of 4th mora necessary for the identification of accent types for nonsense words "imeyuro," "nemeira," and "ominere."*

CONCLUSION

Perceptual experiments were conducted on the role of prosodic features for the identification of spoken words. It was found that the role is largest for type 1 accent. Several results were also obtained implying the existence of the process for accent-type identification. The above experiments, however, were restricted to word level information. Further experiments are necessary to examine the role of prosodic features in the process of perceiving higher-order linguistic information, such as syntactic structures.

REFERENCES

H. Fujisaki & K. Hirose (1984), "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Jpn.*, Vol.5, No.4, pp.233-242.

H. Fujisaki, K. Hirose, Sumio Ohno & N. Minematsu (1990), "Influence of context and knowledge on the perception of continuous speech," *Proc. Int'l Conf. on Spoken Language Processing, Kobe, 18-22 November 1990*, 10.9, Vol.1, pp.417-420

N. Minematsu, S. Ohno, K. Hirose and H. Fujisaki (1992), "The influence of semantic and syntactic information on spoken sentence recognition," *Proc. Int'l Conf. on Spoken Language Processing, Banff, 12-16 October 1992*, Tu.fPM.4.5, Vol.1, pp.153-156