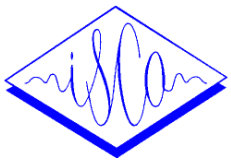

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

ROLE OF PROSODIC FEATURES IN THE HUMAN PROCESS OF SPEECH PERCEPTION

Nobuaki Minematsu and *Keikichi Hirose*
mine@gavo.t.u-tokyo.ac.jp hirose@gavo.t.u-tokyo.ac.jp

Dept. of Electronic Engineering, University of Tokyo
7–3–1 Hongo, Bunkyo–ku, Tokyo, 113 Japan

ABSTRACT

Prosodic features of speech are tightly related to the linguistic information of an utterance, such as the lexical meaning of a word and the syntactic structure of a sentence. Although we have conducted a series of perceptual experiments and constructed a model for the human process of speech recognition, only the segmental features have been taken into account. In order to quantitatively examine the role of prosodic features, two perceptual experiments were performed, where speech stimuli were synthesized by manipulating accent and phrase components of the fundamental frequency contour. Word stimuli were used in the first experiment to clarify the effects of word accent types on word perception. It was found that the words with type 1 accent were perceived differently from those with the other types. In the other experiment, sentence stimuli were used to examine the effects of phrase components on the perception of higher-order linguistic information. The results indicated that the phrase component, even if its command value was small, could work as a cue for detecting the syntactic structure in a sentence.

1 INTRODUCTION

It is obvious that the human process of spoken language recognition is not a simple bottom-up or left-to-right process. When recognizing the lexical items in spoken sentences, a listener is supposed to utilize various sources of acoustic and linguistic information. In order to clarify and model the process, a series of psycholinguistic experiments have been conducted on the following themes[1]–[3].

(1) The size of the unit of speech perception used in various sizes of context.
(2) The effect of differences in the perceptual unit size on the matching process in word recognition.
(3) The relation of familiarity with a word to the amount of necessary information for its correct recognition.
(4) The identification of words in a semantically common, semantically uncommon, anomalous or ungrammatical context.

In these experiments, stimuli were prepared after some manipulations of acoustic parameters or linguistic attributes in speech. However, these included no experiments with the direct control of prosodic features although they are tightly related to the linguistic information of an utterance, such as the word meanings, the syntactic structures and the focal conditions. From this point of view, two perceptual experiments were designed, where only fundamental frequency (henceforth, F_0) contours of stimuli were controlled directly, though the prosodic features are also related to glottal source power, phoneme length and pause length as well as F_0 contours. This is because of the priority of the F_0 contour in the acoustic manifestations of prosodic features of Japanese speech. In the first experiment, word stimuli were synthesized after the modification of their accent types to investigate the effects of accent types on word perception. In the other experiment, sentence stimuli were synthesized and presented to subjects after the manipulation of a phrase command for each phrase and an accent command for each word in a sentence, in order to examine the effects of phrase components on the perception of syntactic structures, and the influence of word accent types in a sentence. In the following sections, each of the two experiments is described in detail.

2 ROLE OF ACCENT TYPES ON SPOKEN WORD PERCEPTION

2.1 Objectives

Each lexical item has its own accent type. This fact lets us easily assume that the accent type is a kind of information that facilitates the identification of words, though a particular word cannot be defined only with the information on its accent type. In the case of 4-mora words in the Tokyo dialect, there are only 4 possible accent types when uttered in isolation, though 16 combinations are considerable for the total number of moraic high-low patterns in F_0. As the first step to investigate the effects of prosodic features on speech perception, the differences among the effects of 4 possible accent types were examined.

2.2 Word Accent in Japanese

Word accents in Japanese can be categorized into some abstract patterns. In these patterns, F_0 for each mora is denoted by "high" or "low". This means that there are 2^n patterns considerable for n-mora words. The number of actual accent types is, however, much smaller than 2^n. In fact, only $n+1$ accent types for n-mora words are possible in the Tokyo dialect, which are denoted by "type i" accents (i=0∼n) as shown in **Fig. 1** for the case of n=4. Type i accent has a rapid downfall

in the F_0 contour at the end of the i-th mora, except for the case of i=0 which does not have an apparent downfall. When uttered in isolation, type n accent has the same F_0 contour as type 0. Therefore, in the current experiment, some 4-mora words were selected excluding the words with type 4 accent.

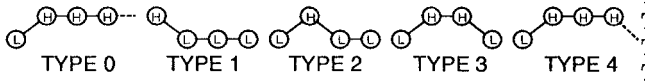

Fig. 1 All the accent types for 4-mora Japanese words.

2.3 Speech Material

As shown in **Tab. 1**, 12 utterances of 4-mora nouns were selected for each of accent types 0~3. These materials were uttered by an adult male speaker of the Tokyo dialect at the rate of around 7 mora/sec. After recorded into a DAT, the materials were re-sampled at 10 kHz and with 12 bit accuracy. The following three types of manipulation were conducted on the F_0 contours during the process of PARCOR analysis-synthesis.

CASE 1 Keeping F_0 constant at 100 Hz.
CASE 2 Transforming the F_0 contour into other accent types.
CASE 3 Without any modifications.

Manipulation for CASE 2 was performed based on the model of F_0 contour generation[4] which will be merely called "F_0 model" in the rest of this paper. F_0 contours for CASE 2 were generated by the F_0 model after shifting the onset and the offset timings of the accent commands to their typical values. A band elimination of 0.5 kHz to 3.0 kHz was further performed for all the synthetic words to make it difficult to perceive a word only by the syllable-based matching process.

2.4 Method and Procedure

The stimuli after the manipulation above were presented through headphones at 4 sec inter-stimulus intervals to 10 male subjects, who were asked to reproduce the words orally. The experiments were conducted in the order of CASE 2, CASE 1 and CASE 3.

2.5 Results and Discussions

After the experiments, two types of word recognition rates were obtained. One is calculated separately for each case and each original accent type. The other is calculated separately for each case and each resulting accent type after the CASE 2 transformation. **Figs. 2** and **3** respectively show these rates. In **Fig. 3**, the bar for type i and CASE j indicates the recognition rate for the words which would be modified into type i accent under the CASE 2 transformation and were actually presented according to CASE j. In both figures, CASE 2' represents the recognition rate by accent type, where the response from a subject was judged if it was reproduced with the same accent type as that of the presented word.

Tab. 1 4-mora words used for the CASE 2 transformation.

Type 0→1	/raion/	/akabou/	/niNjiN/	/naiyou/
Type 0→2	/shingou/	/omatsuri//yokujitsu//amerika/		
Type 0→3	/hiroshima//aimai/		/orugaN/	/raihiN/
Type 1→0	/nekutai/	/koumori/	/wakuchin/	/randamu/
Type 1→2	/naitaa/	/monoraru//amazon/		/uNsei/
Type 1→3	/kamakiri/	/uNmei/	/ookami/	/eNbuN/
Type 2→0	/imomushi/	/norimaki//omusubi/		/yononaka/
Type 2→1	/mimizuku/	/katakori//onigiri/		/nodoame/
Type 2→3	/aomori/	/toraburu//murasaki/		/origami/
Type 3→0	/kamisori/	/tamanegi//nokogiri/		/machigai/
Type 3→1	/kaminari/	/nissuu/	/neNryou/	/noNbiri/
Type 3→2	/kaminoke/	/nakigoe/	/noumiso/	/teNkizu/

The recognition rate of each accent type has a similar value for CASE 3. As clearly shown in **Fig. 2**, the largest drop due to accent type transformation is observed for the words with type 1 accent for CASEs 1 and 2. In **Fig. 3**, the largest drop is observed when the words with non-type 1 accent is modified to those with type 1 accent. These results imply the greater role of prosodic features for the perception of words with type 1 accent. The recognition rate by accent types (CASE 2') has the largest score for the samples with type 1 accent as shown in **Fig. 3**. This result also supports the findings above. The reason of the findings is considered to be the fact that the type 1 accent has a falling in the F_0 contour at the beginning of the word and, therefore, the prosodic features can be utilized before the perception of the segmental features of the word is completed and can also facilitate the accessing process to the mental lexicon by restricting the searching space in the LTM. If we assume the importance of prosodic features is affected only by the location of the downfall in the F_0 contour, the role should be decreased gradually for types 2, 3 and 0. No result, however, was found in this experiment to support this hypothesis. This implies the existence of the perceptual mechanism which makes it possible to recognize a word only with the partial acoustic features and linguistic information of the stimuli.

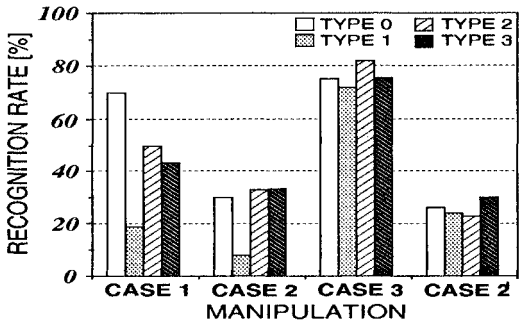

Fig. 2 Word recognition rates summarized separately for each original accent type.

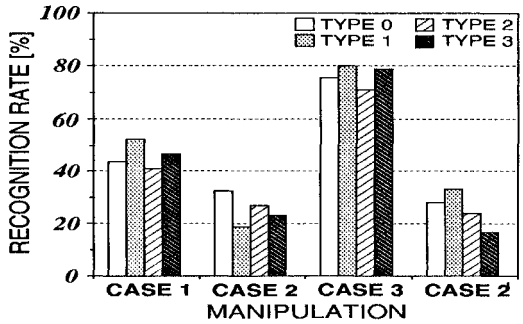

Fig. 3 Word recognition rates summarized separately for each accent type after transformation.

3 ROLE OF PHRASE COMPONENTS ON SENTENCE SPEECH PERCEPTION

3.1 Objectives

In the F_0 model, an F_0 contour is represented by the sum of two types of components, viz., accent and phrase components. In the experiment in section **2**, only the accent components were controlled, with the phrase components unchanged. This control is valid only when our experiments are restricted to the clarification of the human process of word perception. In the usual cases of human communications, however, humans use spoken language in the form of sentences and, therefore, the role of phrase components must be carefully examined. After these considerations, the following experiment was designed, where sentence stimuli were synthesized using an LMA (Log Magnitude Approximation) filter[5]. Both phrase and accent components in a sentence were modified to investigate the role of phrase components on detecting syntactic structures and the effect of accent components on recognizing words in a sentence.

3.2 Speech Material

16 utterances of 11-word target sentences and 27 utterances of dummy sentences were prepared. The syntactic structure, shown in **Fig. 4**, was only adopted for the target sentences, in order to avoid the effects of the differences of difficulty in syntactic analysis among the stimuli. In the upper raw, it is shown in Japanese word order and is converted into English word order in the lower raw. Bold face letters represent major syntactic elements. This structure includes three phrases, all of which are underlined in **Fig. 4**. In the target sentences, 4 or 5-mora words with 4 accent types were placed at the locations O_1 and O_2. Eight words were selected for each accent type and for each location. Target and dummy sentences were uttered by an adult male speaker who was asked to speak each sentence in one breath, in order to avoid the effects of the prosodic features other than F_0 such as pause and duration. As a result, the speech rate of the stimuli for this experiment was approximately 9 mora/sec, which was faster than that for the experiment in section **2**. After recorded into a DAT, the speech stimuli were re-sampled at 10 kHz and with 16 bit accuracy.

3.3 Acoustic Manipulation of the Stimuli

All the stimuli were synthesized using the LMA filter which can well approximate any form of logarithmic spectrum. With this digital filter, the characteristics of vocal tract is closely approximated and an analysis-resynthesis speech of high-quality can be obtained. During the synthesis process, accent and phrase components were manipulated. Unlike the experiment in section **2**, both the amplitudes of accent commands and the magnitudes of phrase commands were varied from 0.0 to a fixed value (0.3 in this experiment).

It is possible to produce stimuli with no phrase component by changing the magnitudes of all the phrase commands into 0. The accent components, however, tend to be larger at the beginning of a phrase and smaller at the end. It follows that a sentence speech synthesized only with the ordinary accent components and without phrase component still sounds to have non-zero phrase components. These results led us to the following methods for the acoustic manipulation.

(1) Following to the binary accent description, accent components are added to the corresponding segments. The onset and the offset timings of the accent command are set to 40 msec earlier than the onsets of the vowels corresponding to the rise and the fall of the F_0 contour respectively, as shown in **Fig. 5**.

(2) A phrase component is placed for each phrase. The onset timing of the phrase command is set to 50 msec earlier than the onset of the first vowel in the phrase.

The command values of the accent and phrase components were varied from 0.0 to 0.3 at intervals of 0.1. All the accent components and all the phrase components in a sentence were respectively set to have the same command values as shown in **Fig. 6**. The logarithmic mean value of F_0 was kept at 110 Hz in all the modifications. The top panel shows the F_0 extracted from the original material by * marks and the F_0 contour used for the synthesis. In the lower panel, the phrase and accent commands for the synthesized F_0 contour are shown.

$$\underline{\text{O+V+}\boldsymbol{S}} + \underline{\text{S+V+}\boldsymbol{O_1}} + \underline{\text{O+V+}\boldsymbol{O_2}} + \text{Adv+}\boldsymbol{V}$$

$$\underline{\boldsymbol{S}\text{ who V+O}} + \text{Adv+}\boldsymbol{V} + \underline{\boldsymbol{O_2}\text{ who V+O}} + \underline{\boldsymbol{O_1}\text{ which S+V}}$$

Fig. 4 Syntactic structure of the target sentences.

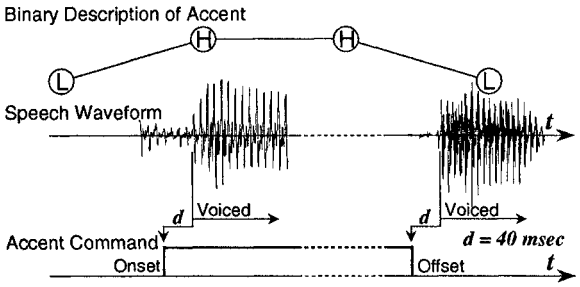

Fig. 5 Acoustic manipulation for an accent command.

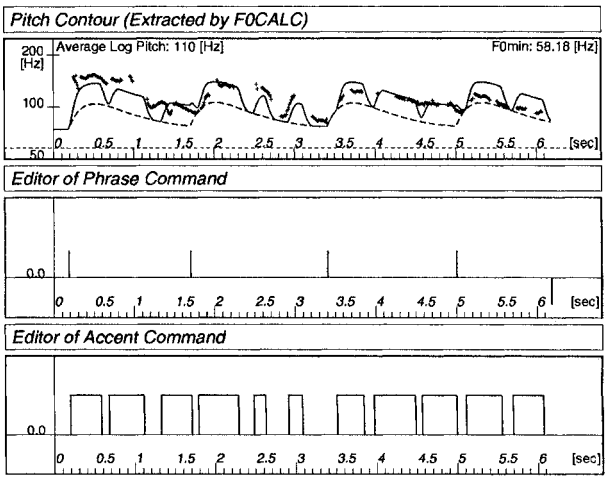

Fig. 6 The F_0 from the original material and the F_0 contour for the synthesis with two types of commands.

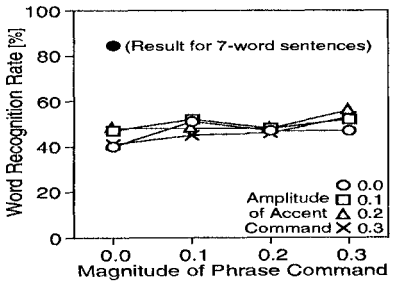

Fig. 7 Word recognition rate for each accent and phrase component.

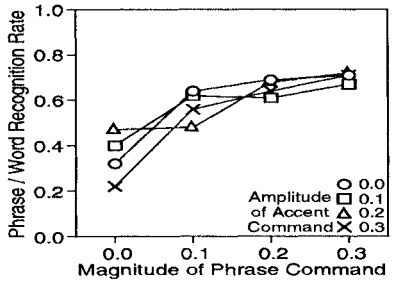

Fig. 8 Ratio of phrase recognition rate to word recognition rate.

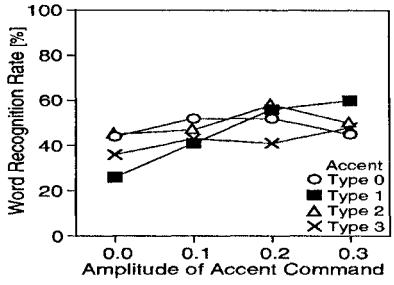

Fig. 9 Word recognition rate for each accent type.

3.4 Method and Procedure

Since both accent commands and phrase commands were varied at 4 steps, there were 4×4=16 combinations in the value assignment to the two types of commands. These assignment were performed for the 16 target sentences. On the other hand, 7-word dummy sentences were modified to have the constant F_0 of 110 Hz. The other dummy sentences were manipulated by taking the procedures for the target sentences and for the 7-word sentences randomly. After the manipulation, 16 target and 27 dummy sentences were synthesized and presented in a session to 8 male subjects through a loud speaker, where the inter-stimulus interval had the same duration as the preceding stimulus. The target sentences had the command values different from one another in a session. And the way of assignment to each target sentence differed among the subjects. Two days after a session, another session was performed to the same subject using the different way of assignment. Subjects were asked to reproduce the whole sentence orally or, if not possible, as many words in the stimulus as they could. During the session, the subjects were required to do an extra task, where they had to trace a mark by a mouse pointer moving slowly on a CRT of a PC.

3.5 Results and Discussions

After the experiments, word recognition rates were obtained for each command value pair of accent and phrase components, as shown in **Fig. 7**. The response was also considered as correct when it was a synonym of the stimulus. In this figure, the result for 7-word sentences is also marked, showing much higher rate than that for 11-word sentences. This is considered to be due to the fact that the number of possible entries for the STM is around 7, i.e., the so-called "7 chunks". While every constituent word can be stored in the STM in the case of 7-word sentences, the STM is supposed to be flushed over by the consecutive words in the case of 11-word sentences. This difference between the two conditions is thought to produce a large gap in recognition rates. Although the stimuli with no accent component showed a great drop in recognition rate in section **2**, no drastic change was observed among any combinations of accent and phrase components. This is because segmental features of the stimuli were not distorted unlike the experiment in section **2** and, therefore, the effect of accent components on word recognition was not shown clearly. **Fig. 8** shows the ratio of phrase recognition rate to word recognition rate, where a lower value in the vertical axis indicates the percentage of correctly recognized words being not included in correctly recognized phrase is lower, and *vice versa*. In this figure, the ratio for every accent command amplitude increases with the increase of the magnitude of the phrase command. This fact indicates that the words in a sentence with phrase components tend to be recognized after the grouping in prosodic phrases by the phrase components. This implies that the phrase component can be used as a cue for detecting the syntactic structures in a sentence. It should be noted also that there is a steep rise in the rate between the two cases of the magnitudes 0.0 and 0.1. **Fig. 9** shows the recognition rate for the words at the locations O_1 and O_2 for each accent type. The largest increase of word recognition rate with the increase of the accent command value is found for the words with type 1 accent. The findings obtained in section **2** were also supported in the sentence level experiment.

4 CONCLUSIONS

In this paper, two perceptual experiments were described to investigate the effects of accent types on word perception and those of phrase components on detecting syntactic structures in a sentence. The results showed the greater influence of type 1 accent on word perception than of any other types, and the effectiveness of phrase components on perceiving words under the phrase structure. It should be noted that, in the latter experiment, the phrase component was shown to affect the grouping process of words in a prosodic phrase, even if its command value was very small (0.1 in the experiment).

REFERENCES

[1] H.Fujisaki, K.Hirose and H.Udagawa, "*A Study on Units of Processing in the Perception of Continuous Speech*," IEICE Technical Report, SP86-53, pp.15-22, 1986.

[2] H.Fujisaki, K.Hirose, S.Ohno and N.Minematsu, "*Influence of Context and Knowledge on the Perception of Continuous Speech*," Proc. ICSLP90, vol.1, pp.215-218, 1990.

[3] N.Minematsu, S.Ohno, K.Hirose, and H.Fujisaki, "*The Influence of Semantic and Syntactic Information on Spoken Sentence Recognition*," Proc. ICSLP92, vol.1, pp.153-156, 1992.

[4] H.Fujisaki and K.Hirose, "*Analysis of voice fundamental frequency contours for declarative sentences of Japanese*," J.Acoust. Soc. Jpn., vol.5, No.4, pp.233-242, 1984.

[5] S.Imai, "*Log Magnitude Approximation (LMA) Filter*," Trans. IEICE, vol.J63-A, No.12, pp.886-893, 1980.