UTILIZATION OF MULTIPLE UNITS IN HUMAN AND MACHINE RECOGNITION OF CONTINUOUS SPEECH ---- PERCEPTUAL EVIDENCE AND A PROPOSAL FOR AN ASR SYSTEM

H. Fujisaki, H. Udagawa and N. Kanedera

Dept. of Electronic Eng., Faculty of Engineering
University of Tokyo, Bunkyo-ku, Tokyo, 113 JAPAN

The ultimate goal of automatic speech recognition (ASR) is obviously to replicate the human capability of speech processing by machine. Research of ASR will thus profit very much from investigations into the human processes of speech perception/comprehension. Few studies, however, seem to have been made along this line. The present paper summarizes a series of psycholinguistic experiments conducted to elucidate certain aspects of human speech perception, especially in relation to the units of processing. On the bases of these experiments, we propose a new system for continuous speech recognition utilizing multiple units.

AN EXPERIMENT ON HUMAN SPEECH PERCEPTION

Objective and Method

While it is desirable to design a psychological experiment that would directly disclose the size of the unit of human speech perception, the difficulty of the problem led us to adopt an indirect approach. We first designed an experiment which would show that certain segments are not processed as independent perceptual unit in human speech recognition. In the following experiment, we investigated perception of connected speech in the presence of deleted syllables to find out whether such deletions are always noticed by the listener[1,2]. If they are not noticed by the subject, one would be able to infer that the subject is not treating the deleted syllables as independent perceptual units, but is recognizing the input speech as a sequence of larger units. The fact that the deletion of a certain syllable is not noticed would indicate that it does not impair perception of a larger unit containing the deleted syllable.

The original speech material was one minute of speech recorded by a male speaker reading a Japanese text at a normal speech rate of approximately 7 morae/sec. The speech signal was low-pass filtered at 4.8 kHz, sampled at 10 kHz with 12 bit accuracy for processing by a digital computer. A total of 25 CV syllables was deleted on the basis of visual inspection of the speech waveform on an X-Y plotter. In order to avoid artifacts, only CV syllables, each starting with an unvoiced consonant and being followed by an unvoiced stop consonant, were selected for deletion. Figure 1 illustrates an example of syllable deletion. In order to examine the effect of context on the noticeability of the deletion, the following four types of test stimuli were prepared after deletion of the syllables.

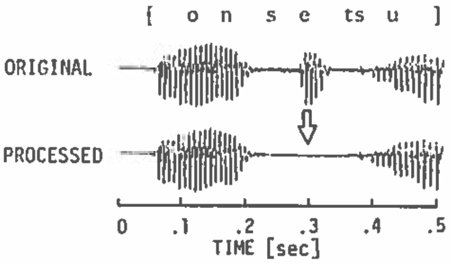

Fig. 1. An example of syllable deletion. The syllable [se] of the word "on'setsu" (meaning 'syllable') is deleted from the original signal.

(1) Segmented into lexical words and randomized.
(2) Segmented into prosodic words and randomized.
(3) Segmented at every pause and randomized.
(4) Without segmentation and randomization.

These stimuli were presented to each subject through a binaural headphone in four test sessions.

The subjects were three male adults with normal hearing. The subject's task was to count the total number of deleted syllables he could notice under each of the four test conditions. Each subject sat for the four test sessions at least five times.

Results and Interpretation

The results of the experiment is shown in Table 1 and the averaged results of the three subjects are shown in Fig. 2. The averaged probability of noticing the deleted syllables is approximately 70% under test condition (1), i.e., when the speech signal is segmented into lexical words and randomized, it drops only slightly under condition (2), but drops rather drastically below 40% under conditions (3) and (4), i.e., when the speech signal is either segmented at every pause or not segmented at all. The difference of results for condition (3) and for condition (4) is quite small.

Table 1. Probability(%) of detection of syllable deletion of each subject.

SUBJECT	LEXICAL WORD	PROSODIC WORD	CLAUSE	SENTENCE
A	77.3	77.3	40.8	40.8
B	69.3	54.7	34.4	32.8
C	69.3	64.0	40.8	37.6
AVERAGE	72.0	65.3	38.7	37.1

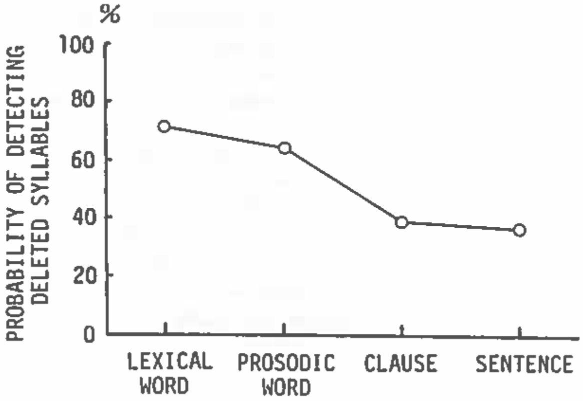

Fig. 2. Results of the perceptual experiment. Relation between size of given context and probability of detection of syllable deletion. Each circle expresses mean value of three subjects.

These results indicate that human listeners pay more attention to syllabic units in a word context, but pay much less attention when the context is as large as a clause or a sentence. In other words, the unit of speech perception is more likely to be syllable-sized when the available context is of the size of a word, but the unit is more likely to be word-sized when the context is as large as a clause or a sentence. Although further experimental studies are necessary, the result of the present experiment suggests that the unit of human speech perception is not unique, but is rather multiple.

FURTHER EXPERIMENTS

Although the above-mentioned experiment revealed the multiplicity of perceptual units, we still need to know the actual size of the units as well as the exact conditions at which one type of units is predominantly used. In this section, we describe some of further experiments being carried out or planned to investigate more deeply into the human processes of speech perception.

Size of Perceptual Units

Granting that the unit in perception of connected speech is larger than a syllable, we need to know whether it is a morpheme, a lexical word, or a prosodic word. The following experiment was designed to answer this question.

Since it has become clear that deletion of a syllable is more easily noticed at the initial position of a perceptual unit than elsewhere, the following three types of stimuli were prepared.

(1) Stimuli in which syllable deletions occur only at the morpheme-initial position which is not the word-initial position.

(2) Stimuli in which the same number of syllable deletions occur only at the word-initial position which is not the initial position of a prosodic word.

(3) Stimuli in which the same number of syllable deletions occur only at the initial position of a prosodic word.

The experimental procedure is the same as in the experiment described in the previous section. If there is no significant difference in the detection rate of syllable deletion among the three types of stimuli, we may infer that the perceptual unit in this case is most likely a morpheme. If the detection rate for the type (1) stimuli is significantly lower than for the type (2) stimuli, but the latter show no significant difference from the type (3), then we may infer that the perceptual unit is a lexical word. In the same vein, if the detection rate is significantly higher only for the type (3) stimuli, we may infer that the perceptual unit is a prosodic word or a still larger unit. Our preliminary results suggest that the latter case is most likely, although we still need more experimental data to confirm it.

Effect of Syntactic Roles on Detectability

Assuming that the unit in perception of connected speech is a prosodic word, one can naturally ask whether all the prosodic words in a sentence receive the same degree of attention and thus show approximately equal detection rate of deleted syllables, or they show different detection rate depending on the difference in their syntactic roles. This question can be answered by investigating the dependency/independence of the detection rate on the syntactic role of the prosodic word containing a deleted syllable. Preliminary results indicate that there are significant differences in the detection rate depending on the syntactic role.

Size of Context on Syllable Recognition

While it is true that most of the evidences and discussions in the foregoing sections are in favor of the use of units larger than the syllable, there are also cases where one has to rely on syllable recognition[3]. If, for example, we are to deal with a very large vocabulary, or even with an unlimited vocabulary, the system will occasionally have to recognize (or transcribe) unknown words syllable by syllable, just as a human listener will do when presented with an unknown word.

In order to design a recognition system whose performance is comparable to that of a human listener, it is thus necessary to know human perception of syllables in connected speech. It has been shown that a human listener needs a context of one syllable each immediately before and after the target syllable in order to be able to recognize with high accuracy the target syllable in connected utterances of one speaker[4]. Likewise, syllable recognition by machines will have to take into account the influences of the context of similar span.

OUTLINE OF AN ASR SYSTEM USING MULTIPLE UNITS

From the evidences and discussion in the foregoing sections, we have proposed a new system for continuous speech recognition based on template matching of multiplicity of liguistic units (idioms, prosodic words, and syllables)[2]. The system operates in the following four steps:

1) Extract acoustic parameters of input speech. (formant frequencies, fundamental frequency, band-limited power, etc.)

2) Detect syllable nuclei, prosodic word boundaries, and clause/sentence boundaries.

3) Detect and recognize frequently used idioms and prosodic words in the continuous speech signal by using their templates. For the portions of input speech where the template matching fails, syllables are detected and recognized by using context-dependent syllable templates.

4) Construct a lattice of (prosodic) word candidates based on the results of the preceding step. Syntactic and semantic coherence is evaluated for all combinations of candidates.

If real-time processing is not required, the system performance would be still more improved. When ambiguity remains, it can also be checked for global coherence to reduce the candidates and to obtain the most probable output. Global coherence is also utilized to re-examine and revise the results of recognition already obtained for a prior input. This is only possible when real-time processing is not required.

REFERENCES

[1] Fujisaki, H., K. Hirose, H. Udagawa, N. Kanedera and Y. Sato, "Considerations on units for continuous speech recognition based on human process of speech perception," Rec. Spring Meeting, Acoust. Soc. Japan, 3-1-14 (1985).

[2] Fujisaki, H., K. Hirose, H. Udagawa and N Kanedera, "A New Approach to Continuous Speech Recognition Based on Considerations on Human Speech Perception," Proc. 1986 IEEE Int. Conf. ASSP (1986).

[3] Fujisaki, H., K. Hirose, H. Udagawa, T. Inoue, T. Ohmori and Y. Sato, "Analysis of variability in the acoustic-phonetic characteristics of syllables for automatic recognition of connected speech" Trans. of the Committee on Speech Research, Acoust. Soc. Japan, S84-69 (1984).

[4] Kuwahara, H., H. Sakai, "Perception of vowels and CV syllables segmented from connected speech," J. Acoust. Soc. Japan 28, 225 (1972).