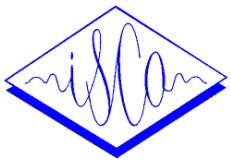

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

INFLUENCE OF CONTEXT AND KNOWLEDGE ON THE PERCEPTION OF CONTINUOUS SPEECH

Hiroya Fujisaki, Keikichi Hirose, Sumio Ohno and Nobuaki Minematsu

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

ABSTRACT

While it is generally assumed that human speech perception starts with the identification of the smallest units, i.e., phones, followed by lexical access, the great variability found in the acoustic characteristics of continuous speech on the one hand and the apparent ease of human listeners in coping with the variability on the other call for re-examination of the conventional view. This paper describes a few experiments conducted to examine the influence of context and knowledge on the units of recognition as well as that of familiarity on the ease of lexical access. A tentative model is then presented for the human processes of spoken language perception.

1. INTRODUCTION

It seems to be the prevailing view of the human process of speech perception that the process starts with identification of the smallest units, i.e., phones, followed by lexical access. While great variability is observed in the acoustic manifestations of smaller units such as phones or syllables in continuous speech [1], the apparent ease of human listeners in coping with the variability suggests that the human process is far more flexible, and calls for a need for the re-examination of the conventional view. The present paper describes a few experiments conducted to elucidate some of the characteristics of the human processes. The first experiment investigates the influence of context on the size of units of speech perception, the second experiment looks into the influence of knowledge on the size of units and rate of correct recognition of these units, and the third experiment investigates the effect of familiarity of lexical items on the process of lexical access. Finally, a tentative model is presented for the human processes of spoken language perception that is consistent with the findings of those and other experiments.

2. INFLUENCE OF CONTEXT ON THE SIZE OF UNITS OF SPEECH PERCEPTION

2.1 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 continuous speech in the presence of syllables replaced by silent intervals to find out whether such replacements are always noticed by the listener [2]. If they are not noticed by the subject, one would be able to infer that the subject is not treating the replaced syllables as independent perceptual units, but is recognizing the input speech as a sequence of larger units. The fact that such a total lack of acoustic manifestation of a certain syllable is not noticed would indicate that it does not impair perception of a larger unit containing the 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 replaced by silence on the basis of visual inspection of the speech waveform. 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 replacement. Figure 1 illustrates an example of syllable replacement. In order to examine the effect of context on the noticeability of the replacement, the following four types of test stimuli were prepared.

(1) Segmented into lexical words and randomized.
(2) Segmented into prosodic words and randomized.
(3) Segmented at every pause and randomized.

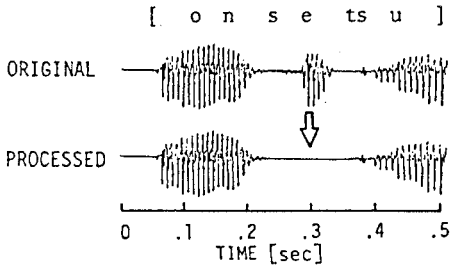

Fig. 1. An example of syllable replacement by a silent interval. The syllable [se] *of the word "*onsetsu*" (meaning 'syllable') is replaced by silence.*

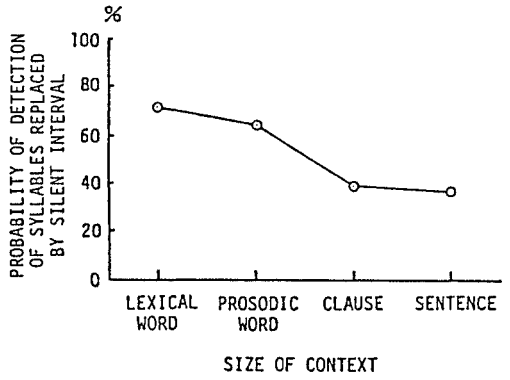

Fig. 2. Relation between the size of given context and the probability of detection of syllables replaced by a silent interval. Each circle represents the averaged result of three subjects.

(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 lacking syllables he could notice under each of the four test conditions. Each subject sat for the four test sessions at least five times.

2.2 Results and Interpretation

The averaged results of the three subjects are shown in Fig. 2. The average probability of noticing the replaced 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.

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. The results suggest that the unit of human speech perception is not unique but is rather multiple.

3. INFLUENCE OF KNOWLEDGE ON THE SIZE OF UNITS AND RATE OF CORRECT RECOGNITION

3.1 Objective and Method

The results of the foregoing experiment suggest that lexical access in continuous speech dose not necessarily require phonetic recognition, and thus can tolerate certain lack of information at the acoustic-phonetic level when a phrase-sized context is available. The amount of acoustic-phonetic information necessary for correct lexical access, however, is expected to vary depending on the listener's knowledge [3]. The following experiment was designed to investigate the influence of available knowledge on the size of units of lexical access as well as on the accuracy of recognition.

In order to control the amount of knowledge available to the listener, the following three types of sentences were prepared.

Sentence Type 1: Well-known proverbs. In order to maintain the uniformity, each selected proverb consists of two phrases, each of which in turn consists of two prosodic words, as shown by the following example.
"hetana kangae yasumuni nitari."
(Poor thinking is just like resting.)

Sentence Type 2: Well-known proverbs in which one prosodic word was replaced by another prosodic word, so that the sentence is still meaningful but is no more a proverb. The proverbs used for constructing these sentences had the same structure as those of Type 1, but were different.

Sentence Type 3: Meaningful sentences whose structures are the same as those in Types 1 and 2, but are not proverbs.

Five sentences were selected for each type. They were read by a male speaker. Each sentence was read in one breath group.

In order to investigate the effect of context on the recognizability of units of different size, the following three types of test stimuli were prepared using all three types of recorded sentences.

Stimuli A - All prosodic words separated and randomized.

Stimuli B - All phrases separated and randomized.

Stimuli C - All sentences intact.

These stimuli were superposed by a white noise at a signal-to-noise ratio of −5dB, and were presented to two groups of seven subjects each, using the method of constant stimuli. In order to avoid familiarization with the stimuli, each subject listened to the stimuli only once. One group of subjects sat for two sessions: stimuli A for the first session, and stimuli B for the second session. The other group of subjects sat for only one session using stimuli C. The subjects were asked to repeat what they heard, and their utterances were recorded and analyzed.

3.2 Results and Interpretation

Table 1 lists the rate of correct recognition of each unit averaged over all subjects and all sentence types.

Table 1. Averaged rates of correct recognition (in per cent).

Unit	Prosodic word (Stimulus A)	Phrase (Stimulus B)	Sentence (Stimulus C)
Sentense Type 1	12.9	39.7	88.6
Sentense Type 2	3.6	14.3	10.7
Sentense Type 3	15.1	20.0	10.7

Whereas the recognition rate is quite low at the level of prosodic word under stimulus condition A, there is a significant improvement at the level of phrase under stimulus condition B, indicating that a phrase-sized context facilitates word recognition. On the other hand, when going from the level of phrase to the level of sentence under condition C, the recognition rate is drastically increased only for sentence Type 1 (proverbs), but decreases for sentence Type 2 (modified proverbs) and sentence Type 3 (non-proverbs). The high rate of correct recognition for Type 1 sentences indicates that each proverb is probably stored as a unit in the mental lexicon so that recognition dose not require correct recognition of its constituent words. On the other hand, the lower rates of sentence recognition for Type 2 and Type 3 sentences indicate the upper limit of utilization of syntactic and semantic constraints available in ordinary meaningful sentences.

4. FAMILIARITY AND LEXICAL ACCESS

4.1 Objective and Method

While successful lexical access presupposes that the specific item represented by a stimulus should be in the listener's mental lexicon, the ease and accuracy of access will depend on one's familiarity with the specific item [4]. In order to be able to control the degree of familiarity, we used a list of 6000 Japanese family names rank-orderd according to the number of people bearing the family name [5]. For the sake of uniformity of stimuli, only 3-mora names were selected. Seven groups of such names were selected which differ in their rate of occurrence. Each group consisted of ten 3-mora names of approximately equal occurrence rate. These names were pronounced and recorded by a male speaker, and presented to 10 subjects under four noise conditions, i.e., at signal-to-noise ratios of −7.5, −5.0, −2.5, and 0 [dB]. Each subject sat for 4 sessions. The subjects were asked to repeat the names and their utterances were recorded for analysis.

4.2 Results and Interpretation

Figure 3 shows the recognition rate of names as a function of familiarity represented by the average number of people, averaged over all subjects. The four curves correspond to four different S/N ratios. The figure clearly shows that the recognition rate depends heavily on the familiarity, especially under severe noise conditions. Among lexical items of

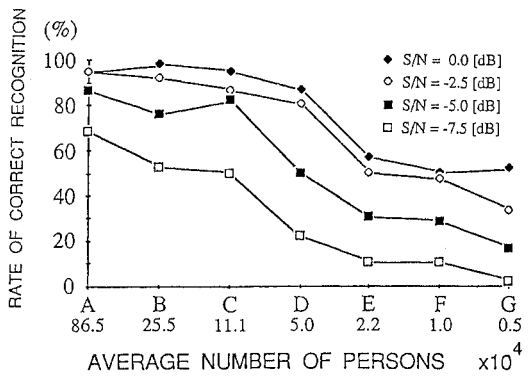

Fig. 3. Familiarity of Japanese family names and their rate of correct recognition in noise.

the same level and same size, the information necessary for correct lexical access is much smaller for familiar items than for unfamiliar items.

5. A MODEL FOR THE HUMAN PROCESSES OF SPOKEN LANGUAGE PERCEPTION

Figure 4 shows a tentative model for the human processes of spoken language perception that is consistent with the findings of the above-mentioned experiments as well as other experiments by the present authors [6-8]. It presupposes the existence of multiple recognition units based on multiple levels of acoustic-phonetic analysis, as well as the dependency of lexical access on familiarity.

6. CONCLUSIONS

Considering the great variability of acoustic-phonetic characteristics of continuous speech and the ability of human listeners to cope with the variability, speech perception in human listeners may not always start with phonetic recognition. A series of experiments has thus been conducted to investigate the size of perceptual unit, the influences of context and knowledge, and the effect of familiarity on lexical access. It was found that human speech perception does not necessarily depend on perception of the smallest units, but often starts with perception of words. It was also found that the perceptual unit can be as large as a sentence if the listener is quite familiar with it as in the case of proverbs. It was further found that the degree of listener's familiarity with a lexical item has a significant influence on the accuracy of lexical access. Finally, a tentative model for the human processes of spoken language perception was presented that is consistent with the findings of these and other experiments conducted by the authors.

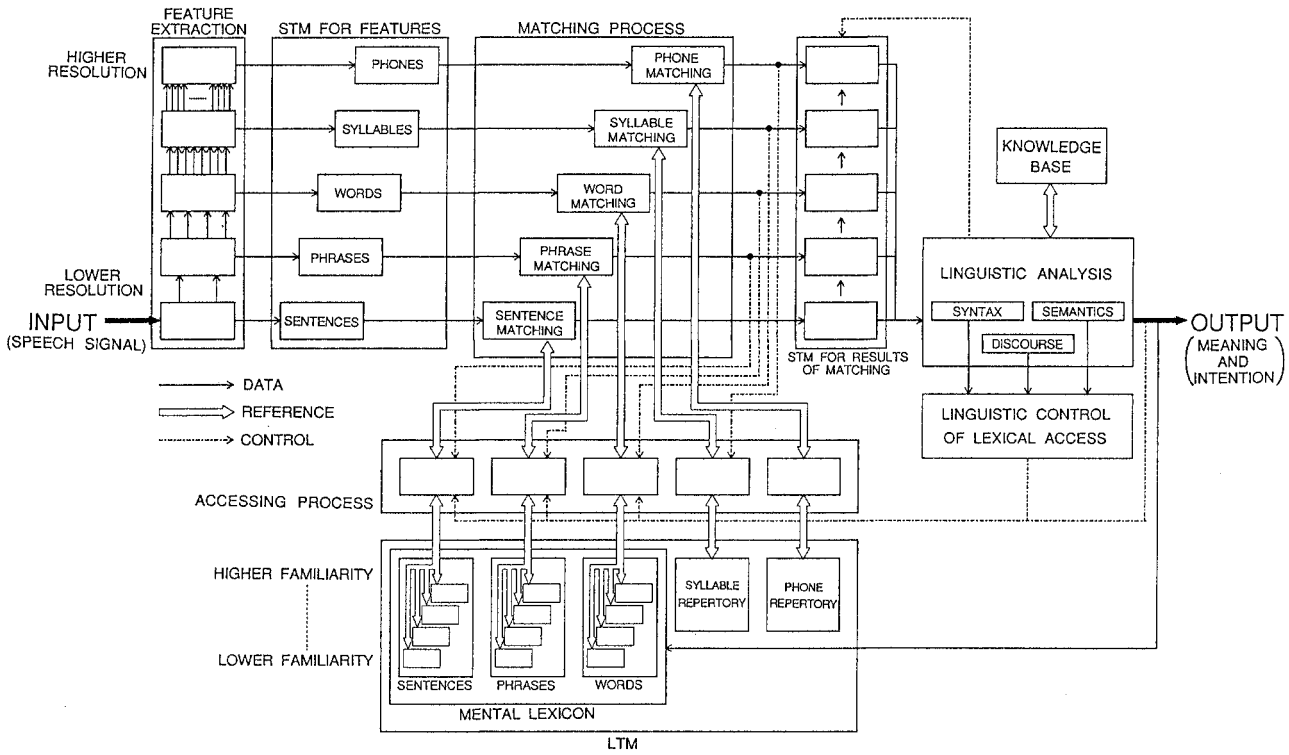

Fig. 4. A tentative model for the human processes of spoken language perception.

REFERENCES

[1] H. Fujisaki, 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 Committe on Speech Research, Acoust. Soc. Jpn.*, **S84-69**, pp.541–548, 1984.

[2] H. Fujisaki, K. Hirose and H. Udagawa, "A study on units of processing in the perception of continuous speech," *Tech. Rep. Inst. Electron. Commun. Eng.*, **SP86-53**, pp.15–22, 1986.

[3] H. Fujisaki, K. Hirose, H. Udagawa, N. Kanedera and K. Hirata, "A study on role of context in human processes of speech perception," *Reports of Spring Meet. Acoust. Soc. Jpn.*, 3-5-7, pp.103–104, 1987.

[4] H. Fujisaki, K. Hirose, S. Ohno and N. Minematsu, "A study on the organization and the mode of access of the mental lexicon," *Reports of Spring Meet. Acoust. Soc. Jpn.*, 3-3-17, pp.103–104, 1990.

[5] H. Fujisaki, K. Hirose, S. Ohno and N. Minematsu, "Experimental study on the structure of and access to the mental lexicon," *Reports of Autumn Meet. Acoust. Soc. Jpn.*, 3-8-2, pp.97–98, 1990.

[6] H. Fujisaki, K. Hirose, H. Udagawa and N. Kanedera, "A new approach to continuous speech recognition based on considerations on human processes of speech perception," *Proc. IEEE ICASSP 86*, 37.2.1, pp.1959–1962, 1986.

[7] H. Udagawa and H. Fujisaki, "An experimental study on lexical matching in human processes of speech perception," *Reports of Autumn Meet. Acoust. Soc. Jpn.*, 3-5-20, pp.151–152, 1987.

[8] H. Udagawa, S. Ohno and H. Fujisaki, "An automatic speech recognition system based on human processes of speech perception," *Tech. Rep. Inst. Electron. Inf. Commun. Eng. Jpn.*, **SP87-90**, pp.25–32, 1987.