Speech Recognition Mechanism of Second Language Learners in Short-Term Memory

Kazuko TOJO
Department of English Literature, Kyushu Women's University
Yahatanishi-ku, Kitakyushu, Fukuoka 807 JAPAN

In instantaneous information processing as required in human dialogue, speech recognition mechanism functions within the realm of one's short-term memory. When a second language learner listens to the speech in the target language, a great portion of the difficulty is posed by the limitation of short-term memory. This is due to the facts that speech sounds are not retrievable and that the processing should be completed within a given time period. The purpose of this study is to clarify the speech recognition mechanism of second language learners in short-term memory. Two experiments were conducted to see how Japanese learners of English in various stages of listening proficiency recognize spoken English sentences and improve their comprehension. The first experiment was concerned with natural speech utterances and the second with acoustically manipulated utterances.

Experiment 1

Seven plus or minus two is a widely known number in the treatment of human perception within short-term memory (Miller, 1956). This being applied to the listening process of an English sentences by second language learners, a word is taken as the smallest information unit for the present study. A finite human channel capacity is then imposed by short-term memory which allows a second language learner to process up to seven-plus-or-minus-two words in a sentence. After the number of the words reaches the upper limit, it is hypothesized that more words are received by expanding the capacity within each unit. For example, the sentence "Why don't you meet me at 10 o'clock in front of the library" contains 13 words which presumably goes beyond the theorized number of words in a sentence. However, if the words are recognized in chunks as "Why don't you meet me/ at 10 o'clock/ in front of the library," the number of units is reduced to 3 and hence the burden of handling them within short- term memory is greatly minimized. The first experiment was conducted to find out the channel capacity of the second language learners at various proficiency levels and other possible restrictions posed by short-term memory. It was also aimed to gain the underlying factors of their chunk-building.

In order to embody the latent and receptive process of speech recognition, an elicited repetition task was given to the subjects. The subjects were instructed to listen to an English sentence and immediately repeat it. The task was regarded as most relevant, among other listening comprehension tests, to instantaneous sentence processing. The channel capacity in the experiment was observed in the overlap between input (a taped sentence) and output (the subject's reproduction). The subjects chosen for the study were 80 Japanese college students learning English, which were divided into three groups according to their proficiency levels. Thirty English sentences were sampled from natural conversational American English. The number of the words in a sentence varied form 3 to 17 words. Vocabulary, grammar and the topic of the sentences were carefully controlled so that they would not cause any difficulties to the subjects. The test sentences were tape-recorded by a native speaker (an American female). The test was conducted in a language laboratory and the reproduced sentences

by the subjects were tape-recorded. The recorded reproductions were scored by the tester according to the number of words correctly reproduced. The acceptability criterion was set by the minimum recognizability necessary to retrieve the original words, as pronunciation accuracy being of secondary importance. Reproduction errors were also recorded.

The following results were obtained. (1) The channel capacity was approximately assumed to be four words for the lower level group, six words for intermediate level group and ten words for the upper level group. (See Figure 1.) The upper level group's reproductions were characterized with a longer reproduction span, which was considered to be realized by chunking the words.

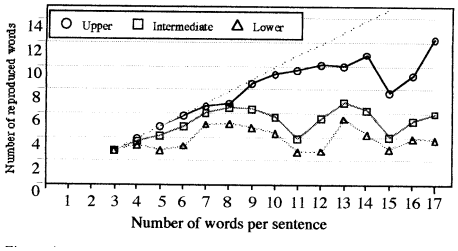

Figure 1: Channel Capacity of Word Reproduction at Three Proficiency Levels

(2) Consistent features were found in the subjects' responses to the given sentences. The subjects reproduced the head segment in a sentence most successfully, followed by the tail segment. The middle segment of the sentence was the first to be dropped. Their characteristic responses to the segments in a sentence were confirmed with the sentence variations where an adverbial phrase was moved from the middle or the tail segment to the head respectively. The variation of "I went to the zoo with my young nephew the day before yesterday," for example, was "The day before yesterday, I went to the zoo with my young nephew." The reproduction of the word 'nephew' was greatly improved from 15.3% (the average of total subjects) to 40.0% in the latter sentence where the word was placed at the sentence end. Comparison of the scores with the variations proved that the subjects' reproductions were greatly affected by the location of the words in a sentence.

(3) The reproduction errors appeared characteristically with the more successful subjects. Some errors were observed as the result of semantic reprocessing (example: 'you should phone call ' for 'you should give us a call') and others as syntactic reprocessing (example: 'If the weather will improve' for 'If the weather improves.') A successful speech recognition was implemented when the processing take place at multiple levels of semantic and syntactic areas simultaneously, rather than at a phonetic level alone.

Experiment 2

In considering some of acoustic characteristics to mark off the clusters of words into an information unit, a pause is one of the most eminent features. Manipulation of pauses is a promising means to help learners to comprehend a sentence more effectively by legitimately chunking words. The second experiment was conducted to see how various types of pauses can increase the listening efficiency. The examination of pauses in natural English speeches reveals the variables of pauses in terms of the location of occurrence, acoustic nature and the duration. In natural speeches, pauses usually occur either at syntactic/semantic unit end or

before the so-called content word (the former was labeled as syntactic/semantic pause and the latter as hesitation pause for this study.) Pauses are also classified in terms of voiced or unvoiced. A voiced pause is usually filled with er or uh sound. As can be seen in the example, as the utterance gets more spontaneous, more voiced hesitation pauses come to appear [example: . . .while Japan has set some /er/ economic problems recently, / it will remain the economic leader in the Asian Pacific /er/ area for / many many years if, / if not indefinitely. (an excerpt from an interview)] As for the duration of pauses, a series of studies by Kohno (1995) suggested that 330 millisecond is a dividing point in considering the quality of auditory processing. He proposed that a pause beyond 330ms in its duration hinders the holistic listening processing as an analytical nature becomes ruling.

The above factors put into consideration, various types of sentences were prepared by artificially inserting pauses in a sentence. Besides the original sentence without pause (a), the sentence with unvoiced syntactic/semantic pauses (b), the sentence with an unvoiced hesitation pause (c) and the sentence with a voiced hesitation pause (c) were prepared. Based upon the recorded sentences with the above mentioned markers, the length of pauses was precisely measured using Sound Scope (GWI-SOS-LC) at 250ms. Also, for type (b), two more variations (0ms pause and 700ms pause) were created. As the result, 6 variations for one test sentence were used for the experiment [example: "I went to the zoo with my young nephew the day before yesterday." (a-original), "I went to the zoo / with my young nephew / the day before yesterday." (b-0) (b-250) (b-700), "I went to the zoo with / my young nephew the day before yesterday." (c-250) and "I went to the zoo with /uh/ my young nephew the day before yesterday (d-uh)."]

Nine original test sentences (7 words to 14 words) were recorded by a native speaker (an American female) for the above mentioned 6 variations. The subjects for the experiment were 150 Japanese college students studying English, which were divided into two groups according to the proficiency level. Each group was further divided into 5 more groups, as each group taking the task with two types of test sentences of the original and one of the other variations. The elicited repetition task was employed with the same procedure and evaluation process as in Experiment 1. The average reproduction rates (%) for each type were compared.

The results showed that the upper proficiency group showed the improvements from the average of a-original (75.9%) to in order of b-250 (91.6%), c-250 (87.7%), b-0 (87.3%), d-uh (84.9%) and b-700 (84.8%) , and that the lower proficiency group showed the improvements from the average of a-original (35.7%) to in order of b-250 (61.5%), b-0 (52.8%), b-700 (50.3%), d-uh (47.6%) and c-250 (42.3%). It is interpreted as follows: (1)The upper group responded flexibly with sentences with a hesitation pause, while the lower level group depended heavily on syntactic/semantic pauses. (2) For both the upper and the lower group, a 250 ms pause showed the most effective results. For the upper group, a 700ms pause showed the least effectiveness in listening, which proved the theory proposed by Kohno that a pause beyond 330ms transfers the listening target into a realm where an analytical rather than a holistic processing becomes dominant.

Discussion

The findings regarding the speech recognition mechanism of second language learners are meaningful for a successful man-machine interaction, in both an indirect and a direct manners. The acquisition process of second language learners in each developmental phase is expected to reveal some important aspects in terms of machine increasing its accuracy

in human speech recognition. In an effort to realize an effective man-machine interaction, the limitations a person is posed by short-term memory and the framework within which human is operating in speech recognition are a dimension that should be taken into account in a man-machine dialogue modeling.

Keywords: speech recognition, short-term memory, information unit, language acquisition