THE IMPORTANCE OF THE FIRST SYLLABLE IN ENGLISH SPOKEN WORD RECOGNITION BY ADULT JAPANESE SPEAKERS

Kazuo Nakayama and Kaoru Tomita-Nakayama

Yamagata University, Yamagata 990-8650, Japan
ek204@kdw.kj.yamagata-u.ac.jp

ABSTRACT

We investigated adult Japanese speakers' deficiencies in English spoken word recognition. We found that the accurate recognition of the first syllable or the initial portion of each word played an important role in recognizing a word correctly. It was implied in the study that their recognition performance would be enhanced by utilizing the speech processing methods, time-scale expansion and/or dynamic range compression. Although approximately 85 percent of English words begin with strong syllables [1], many of them do not carry a sentence stress and they are not pronounced as clearly as isolated words. Moreover, the duration of a word, especially a beginning word is so short that the listener can't recognize it correctly.

Two experiments were administered in the anechoic room. In the first experiment, subjects listened to extracted words and corresponding isolated words of English, which included words without primary stress on the first syllables. We found that they had difficulty in recognizing both isolated words and the extracted words, especially when the word did not begin with a strong syllable, which was sounded somewhat unclear. This is quite frequent in a normal English speech. We confirmed that they had difficulty recognizing the words which began with weak syllables and it is concluded that the first syllable plays an important role in the recognition of words at least for Japanese speakers. In the second experiment, the extracted words and the corresponding time-scale expanded words (henceforth, expanded words) were given. The result indicated that the expanded words were better recognized. It is found that the time-scale modification (henceforth, TSM) of the extracted words didn't lose intelligibility even around the ratio of 2.00, as was clear from the fact that the recognition improved.

1. SPOKEN WORD RECOGNITION

Studies of spoken word recognition can be classified into two kinds. One deals with an isolated word and the other deals with an extracted word.

1.1. Isolated spoken word recognition

There seems to be more acoustic advantage for an isolated spoken word recognition because the duration of an isolated spoken word is much longer than corresponding extracted word.

It was reported on spoken word recognition of English by Japanese university students that if the first syllable is strong, then subjects could identify the word accurately and that as the strong syllable moved rightward, identification rate decreased [2]. The results obtained there implies that the success of identification of the first syllable is crucial for sentence recognition under the supposition that the word can be a beginning word of a sentence.

The previous study which investigated spoken word recognition of English by Japanese university students reported that when the first syllable was amplified, the subjects could identify the word correctly [3]. It employed isolated word stimuli and pointed out that exact perception of auditory information of the first syllable appeared to determine the success of identification. Also, the rate of identification decreased as the strong syllable moved rightward. The results obtained there implies that the success of identification of the first syllable is crucial for sentence recognition.

1.2. Extracted word recognition

An extracted word is the one that is truncated from the sentence. This is an initial attempt to show that extracted beginning word (speech) recognition is difficult for Japanese learners of English. We would also like to suggest that the importance of teaching "actual" pronunciation of a word as well as corresponding isolated word. Actual pronunciation is reflected in eye dialect such as 'bout (about), 'nd (and), prob'ly (probably) and so on.

The quality of spoken signals means that listeners would not be able to perceive speech as successfully as they do if they were engaged in a process of building up the recognition of words solely by attempting to identify their constituent phonemes. Adult native listeners do not perceive speech phoneme by phoneme, or word by word. Instead, they use their knowledge of the phonological regularities of their language, its lexicon and its syntactic and semantic properties, to compensate for the shortcomings of the acoustic signal [4, 5].

Previous studies revealed that when an individual word is extracted from tape recordings of conversations and is played for listeners to identify, only about half of the extracted words can be recognized. If listeners hear them in their original context of utterance, they are readily identified [4, 5], though speech flows are often inseparable and in fact it is often very difficult to truncate a word from the rest of the sentence.

1.3. English spoken word recognition

The familiarity of pronunciation of an isolated word does not seem to contribute to the recognition of a word in a sentence, for example.

2. EXPERIMENT 1

2.1. Purpose

The purposes of experiment 1 are to demonstrate that the subjects have difficulty of recognizing the words which begin with weak syllables.

2.2. Methods

2.2.1 Subjects

Twenty undergraduate students participated in the experiment. They were randomly classified into two groups. The first group listened to the extracted word and the second group listened to the isolated word stimuli.

2.2.2 Stimuli

The stimulus was so made that 1,000 [ms] silent section is inserted before and after the articulation section. Speech materials are read-out fluent speech in General American accent, which are classified into two kinds: First, the sentence stress is on the first syllable of the beginning word in the context, which were from English textbook [Data 1]. Second, sentence stress is on the second syllable and afterwards or nowhere of the beginning word of the sentence. The order of presentation was randomized among subjects.

Raw data of the first kind are the followings.
Try to make persuasive arguments.
What would you think?
Hold a debate on the subject.
My heart is weak
And another thing, Carlos.
Why would we want to do that?
When are you doing the deed?
Choose one position in the dialogue and prepare an argument in favor of that position from the point of view of society.
Take pornography, for example.
One to one, yes.
There you go, Harman.
Let me tell you something.
Now picture yourself at the same rally.
Rank your arguments from the most important to least
Think about the following situations.
Of course not.
Find an apartment with some other student.
Some people think wearing secondhand clothes is a disgrace.
Study the meaning of the expressions in boldface as they are used in the text.
Believe me.

Raw data of the second kind are the followings.
If I had known I was going to get a lecture about drinking, I wouldn't have invited you two for a drink
But I was wondering if there was any possibility I could live with an American family in New York.
But how can I find a good host family?
Imagine that a younger friend asks you for advice about going to study English abroad. Imagine that you are asked to advise high school students about drinking.
Suppose he didn't have very much money but was quite interested in typical Japanese lifestyle. Suppose your classmate invite you to have a drink, but you have recently quit drinking.
Assuming that we will have to make some adjustments to get used to living together, we might as well do that now so that we will live harmoniously as soon as we're married.
Perhaps you guys can advise me about housing.
Repeat them after the tape and memorize them.
Rearrange the words in the parentheses so that each dialog will be meaningful according to the context.

The duration of each beginning word is shown in table 1.

Word	Extracted	Isolated
Take	109	342
Of	110	558
Let	129	614
When	135	444
Find	141	735
Some	149	508
What	173	503
Hold	173	608
Choose	181	641
Why	203	471
Rank	208	728
And	223	476
Try	228	640
Now	228	677
Think	238	471
There	245	598
Study	246	431
One	262	711
My	266	601
Discuss	367	894
Believe	398	729
Consider	473	865
(Average Duration)	222.0	602.0

Table 1: Duration of stressed extracted/isolated word [ms]

Word	Extracted	Isolated
But	55	512
But	89	512
If	114	471
Suppose	290	916
Repeat	303	757
Suppose	312	916
Perhaps	355	660
Assuming	445	752
Imagine	467	742
Imagine	481	742
Rearrange	559	1,137
(Average Duration)	313.6	737.9

Table 2: Duration of word including unstressed extracted /isolated word [ms]

The duration of the extracted words ranges from 55 to 559 [ms]. It should be noted that each duration comes from small samples. We need to examine various data, especially extracted words.

All the extracted words listed above in Table 1 and 2 were separable from the following contexts, they were truncated from the context by Kay's Multi-Speech Model 3700. Judgment of stressed/unstressed was determined by the present authors on the basis of various kinds of visual data provided by the Multi-Speech.

2.2.3 Procedure

The experiment was administered in the anechoic room of the faculty of medicine, the University of Tokyo. Subject was asked to report the meaning of the word s/he heard. When this response was not exact, the subject was then asked to write down the spelling to make sure. The idea of requiring the subject to answer the meaning is to judge if s/he has an association between sound and meaning. That is, to know if s/he can access the exact word from the speech sound. In addition, spelling and/or repetition are examined to make sure of his or her understanding the word if the answer is not reliable enough.

Subjects was seated in front of a computer display and wore a headphone (STAX Lambda Nova Basic), through which the stimulus was presented binaurally. Stimulus was presented twice. In the warm-up session, subjects adjusted the volume of the speech sound to their Most Comfortable Level.

2.3. Results

Some of the most conspicuous results are shown in table 3.

Word	Extracted	Isolated
What	2	9
Hold	3	8
When	3	8
Choose	4	7
Take	2	8
Let	2	8
Of	0	10

Table 3: Correct answers of experiment 1

2.4. Discussion and Conclusion

Although we don't have space to show all the data, the isolated words were statistically more identified than the extracted ones (t(19)=2.07, p<0.05).

It was observed that if a subject was uncertain about the answer, s/he tended to identify the speech sound with the familiar word which sounded most approximate to the subject. The results seems to imply that the exact recognition of the initial portion leads to the exact identification at least for the Japanese speakers.

Subjects' poor recognition of extracted words seems to be reflected in poor performance in listening comprehension. If the subjects' performance on spoken word recognition can be generalized, especially when extracted from a fluent speech and if the word is the first word of a sentence, then their poor performance on listening comprehension is at least in part rendered to the inability to the correct recognition of a beginning word of a sentence. Consequently, Japanese learners appear to be unfamiliar with "actual" pronunciation of a word.

It should be noted that we didn't take the following context effect into consideration here, though the promotable role of subsequent context (the rest of the sentence, here) in the auditory recognition of the beginning word was reported [6].

Not only the duration but articulation (clarity) is another important factor to affect identification of a word. Auditory speech recognition can be improved through training. Materials for speech audiometry would also be helpful. More importantly, aid in training speech recognition. If s/he will be skillful of it, we consider it would affect so-called listening comprehension considerably. In other way, the recognition test as like the experiment could be used in part as listening comprehension.

3. EXPERIMENT 2

3.1. Purpose

Experiment 2 is focused on the effect of time-scale expansion developed by Misaki, M and his co-workers [7].

3.2. Methods

3.2.1 Subjects

Twenty undergraduate students participated in the experiment 2. They were randomly grouped into two groups, each group consisted of ten subjects. First group listened to the extracted words stimuli and second group listened to the corresponding expanded word stimuli.

3.2.2 Stimuli

The same stimuli as experiment 1 were employed.

3.2.3 Procedure

Experiment 2 was carried out with the same procedures as the experiment 1, except that the subject could choose the TSM ratio in the warm-up session, too.

3.3. Results

Some of the most conspicuous results among Japanese speakers are shown in table 4.

Word	Extracted	Expanded
What	2	4
Hold	3	5
When	3	5
Choose	4	7
Take	2	3
Let	2	2
Of	0	1

Table 4: Correct answers of experiment 2

3.4. Discussion and conclusion

The results obtained in this experiment show that overall performance on the expanded stimuli was statistically better than the extracted ones (t(19)=1.75, $p<0.05$).

It might be concluded that the first syllable plays an important role in recognition of word at least for Japanese speakers. If it can be extended into processing sentences, speech techniques mentioned above would be considered to enhance speech processing, too [8].

The present findings have important implications for understanding the course of foreign language acquisition. It is also confirmed in this experiment dealing with foreign language speech sound that the subject tended to access his or her familiar lexical item if s/he was not much certain what exactly s/he heard. As far as experiment 2 was concerned, time-scale expansion (henceforth, TSE) was effective. We judged that the expansion ratio of 1.50 was the most appropriate rate for the intelligibility. It tends to decrease if unstatic section like plosive consonant is expanded so much (more than 1.50) when the subject is young. On the contrary, when the elderly person is the subject, it does not decrease.

Computer-based testing and learning system of spoken word recognition is now underway, in which response time can be obtained. Dynamic range compression is another candidate for enhancing speech. If we have difficulty identifying the beginning word of a sentence, then we can amplify it or beginning 500[ms] for example when the target word is separable from the preceding utterance. Thus we could enhance not only identification of the word but also processing the sentence. The target beginning portion of a sentence is amplified in reverse proportion to incoming speech levels. Speech signals more than or equal to 2 [kHz] is a good candidate for amplification at present. Without a steep fade-in/face-out in amplitude, the transients sound smooth [9]. Together with TSE, lexical identification would be more accelerated, which would decrease the processing load at the beginning of the sentence. It would then enhance processing of the sentence. Optimization of parameters such as the scope of amplification, the frequencies amplification is applied to and the rate of time-scale modification should further be studied to reinforce the demonstration.

Acknowledgements

We would like to acknowledge Professor Shigeru Kiritani of the University of Tokyo, Japan for his invariable comments as well as giving us permission to use the facilities of the university. Thanks also go to Prof. Donna Erickson of Ohio State University who was kind enough to work as an informant. This research is partially supported by the Grant-in–Aid for exploratory Research (09878038) by Mombusho, the Japan Ministry of Education, Science, Sports and Culture.

1. REFERENCES

1. Cutler, A. and D.M. Curter 1987 “The predominance of strong initial syllables in the English vocabulary,” *Computer Speech and Language*, 2, 133-142.
2. Nakayama, K. 1994 “The Ability of Adults to Acquire a Foreign Language Phonology and Remedies for Foreign Language Listening Deficiencies,” *Applicability of Time-Scale Modification of Speech System to Teaching English*, 111-189.
3. Nakayama, K. and M. Misaki 1996 “Recognition of continuous speech enhanced with time-scale modification and/or amplification of the first syllables by postpubescent learners of EFL,” *ASJ Proc.* September 1996. 435-436.
4. Pickett, J.M. and I. Pollack 1963 “Intelligibility of Excerpts from Fluent Speech: Effects of Rate of Utterance and Duration of Excerpts,” *Language and Speech,* 6, 3, 165-171.
5. Bard, E.G. and A. H. Anderson 1983 “The unintelligibility of parent speech,” *Journal of Child Language,* 10, 1-8.
6. Kawashima, T. and M. Kashino 1997 “The Promotable Role of Subsequent Context in the Perception of the Beginning Word in a Sentence,” A Study Report for Acoustics. H-97-44.
7. Suzuki, R. and M. Misaki 1992 “Time-scale modification of speech signals using cross-correlation functions,” *IEEE Transactions on Consumer Electronics,* 38, 3, August, 357-363.
8. Nakayama, K. et al. 1998 “Enhancing speech perception of Japanese learners of English utilizing time-scale modification of speech and related techniques,” *Speech Technology in Language Learning*. KTH, 123-126.
9. Nakamura, A. et al. 1994 “Real Time Voice Speed Converting System with Small Impairments,” *The Journal of the Acoustical Soc. of Japan.* 50, 7, 509-520.

Data

1. Naff, C. and T. Matsui 1995 *Let’s Enjoy Talking in English*. Kinseido.