# A study on voice recognition as human interface for Japanese word-processor

SANNOMIY, Machik
Naruto University of Education

KABASAWA, Satoshi
Matsushita Electric Industrial Co., Ltd.

高木, 英行
Matsushita Electric Industrial Co., Ltd.



出版情報：ヒューマンインタフェースシンポジウム論文集. 2, pp.545-552, 1986-10-29. ヒューマンインターフェース学会
バージョン：
権利関係：

1421

2nd Symposium on Human Interface
Oct.29-30, 1986 Tokyo

# A STUDY ON VOICE RECOGNITION AS HUMAN INTERFACE FOR JAPANESE WORD-PROCESSOR

Machiko SANNOMIYA *
Satoshi KABASAWA **
Hideyuki TAKAGI **
Atsushi YOSHIYA *

* Naruto University of Education
(Naruto, Tokushima 772 JAPAN)
** Matsushita Electric Industrial Co., Ltd.
(Moriguchi, Osaka 570 JAPAN)

Abstract: This paper discusses how the recognition error of a voice-activated word-processor(VAWP) psychologically relates to our document-composing behavior. Using a simulated VAWP, we experimented on how we are affected by input methods; monosyllable input and continuous speech input, from the view-point of recognition accuracy and response time. Furthermore we estimated the VAWP of simulated continuous speech input and the word-processor(WP) connected to a monosyllable recognition system we have developed.

Keywords: voice recognition, word-processor, monosyllable input, continuous speech input

## INTRODUCTION

There have been much efforts for voice recognition system because the system would be helpful for us to compose documents, letters, and memoranda. People can compose both any Japanese words and words of foreign origin by 114 monosyllables. Hence a recognition of 114 monosyllables is enough for a voice-activated word-processor (VAWP) to substitute an ordinary word-processor (WP) of manual typing.

However recognition error always affects us because the technology is not yet advanced enough for VAWP to have 100% accuracy. Recent research for monosyllable recognition gives us an accuracy of about 80% (Kabasawa, et al. 1983). Furthermore we must utter syllables separately in the monosyllable mode. Continuous speech does not have such constraint, but the continuous is more difficult to recognize than the monosyllable because of co-articulation, devocalization, and so on.

Therefore the following discussion are important for us to develop VAWP;

(1) Can the monosyllable recognition be used for composition if it has a high recognition accuracy, for instance, more than 95% ?

(2) Can the monosyllable recognition be used for composition if it activates in real time ?

(3) Is the continuous speech indispensable for VAWP even if its recognition accuracy is not so high ?

(4) How much of the accuracy should the continuous speech recognition have, at least ?

J.H.Gould, et al. studied on the human behavior of composing letters when people used a simulated word-recognition system of large vocabulary (Gould, et al. 1983). Their study is based on the word recognition system. Their results are insufficient for us because we are developing a voice recognition system of infinite vocabulary for VAWP.

Using a simulated VAWP, we studied on how we are affected by input methods; monosyllable input and continuous speech input, from the view-point of recognition accuracy and response time.

Considering the items above, the fol-

lowing hypothesises were built up in our experiment;

(Hypothesis-1) The continuous speech input would be preferable to the monosyllable.

(Hypothesis-2) A difference in feeling between the continuous speech input and the monosyllable would depend on recognition accuracy and response time, and the difference would become little when recognition accuracy is high enough and response time is sufficiently short.

Recognition accuracy were 100%, 95%, and 80%, and response time were 0.025 sec and 1 sec in the experiment. A performance on the 100% accuracy gives a basis for our experiment. The accuracy of 95% is the target in our development, and the 80% is the state of current technology, as described above.

In the following section, we describe our experiment, its results and some discussion.

Furthermore we continue estimating the VAWP of simulated continuous speech input and the WP connected to a monosyllable recognition system we have developed (Maehara, et al. 1983). We also describe the further experiment and discussion.

## EXPERIMENT

By a simulated VAWP, we examined how recognition accuracy and response time relate to input modes: monosyllable input and continuous speech input. We introduced the hypothesises described before in the experiment.

Subject: Twenty-four graduate school students of Naruto University of Education served as subjects. They consists of 14 males and 10 females. The age was about 28.5 year old in average. A half of them participated the monosyllable condition, and the another half was for the continuous speech. Both the ratio of males to females and the average age were equal in the two conditions.

Design: A two X three X two experimental design with 12 subjects per group was used. The first factor was input mode: monosyllable input and continuous speech input. The second was recognition accuracy: 100%, 95%, and 80%. The last was response time: 0.025 sec and 1 sec. Only the first factor was among subjects because it was a heavy burden for a subject to participate both the monosyllable and the continuous speech.

Material: We prepared six texts. Each text was picked out in a book for child psychology, edited to consist of 50 BUNSETSUs(*), and made up of 246.5 syllables in average. Table 1 shows an example of the texts.

Equipment: We used a personal computer: PC-9801 of NEC as a simulated VAWP, an audio tape deck: TECHNICS M77 of National as a simulated voice input system, and a microphone: SM12A of SURE.

Procedure: Subjects were instructed to read out the texts and to input it to the simulated VAWP. The input method was syllable by syllable, or BUNSETSU by BUNSETSU. In the monosyllable condition, when they uttered a syllable and pressed a key appointed for input, a character came out on a display of the VAWP after an arranged period: 0.025 sec or 1 sec. In the continuous speech input, when they read out the BUNSETSU and pushed the input key, a strings of characters appeared on the display with the fixed time delay. If the transaction was correct, they could go to next input. When the presentation was wrong, they had to repeat the same part and to press a key for revision. An accurate expression was surely given once corrected. For the monosyllable input, a key was assigned to an input of SOKUON, for example, which exists in GAKKOU ( meaning school ) in Japanese. Recognition errors of the 80% and 95% simulations happened randomly according to pseudo random numbers.

(*) In this paper BUNSETSU means a smallest group of words with which we can recognize what it means.

Table 1 An example of texts

子ども時代の／経験が／いつまでも／忘れられないのは／、記憶力が／さかんな／年頃の／
せいでもあるが／、見たり／聞いたりする／ことに／はじめての／経験が／多いので／、
強い／刺激だった／ためではないだろうか／。それから／、どうしても／必要な／ものは／
忘れにくい／。サルを／つかった／実験であるが／、えさを／かくした／場所を／１度／
見せておいて／、数分後に／えさを／さがさせたところ／、空腹な／サルほど／記憶が／
よかったと／報告されている／。ところが／、いくら／必要に／せまられていても／記憶した／
すぐ／あとに／強い／ショックを／あたえると／、記憶は／保たれない／。



Table 2 Experimental Procedure

(1) Guidance from the experimenter

(2) Exercise

(3) Work1 → Work2 → Work3 → Work4 → Work5 → Work6
(each Work followed by Rating)

(4) Introspection

## RESULTS AND DISCUSSION

Figs. 1-9 show the mean rating value of each item in Table 5. Table 6 shows significant main effects and interaction given by analysis of variance.

The input mode was significant in all items except "difficult."

The monosyllable input was a heavier load psychologically than the continuous speech. Moreover it is noticeable that the interactions between the input mode and other factors were also significant. Psychological effect of the input mode varied with the recognition accuracy and the response time.

Figs. 1-9 show little difference between the loads of the monosyllable input and the continuous speech in the condition of 100% and 0.025 sec. The tendency in 95% and 0.025 sec is the same as the above. These two conditions exhibited no statistically significant difference. When the recognition accuracy was 80%, the load of the monosyllable input became large regardless of the response time. On the other hand, the continuous speech was not under the influence of the recognition accuracy and the response time.

Moreover there were no significant interaction between the recognition accuracy and the response time. This was an unexpected outcome. This means that each of those factors affects us independently, and that they do not compensate each other. In other words, inaccuracy would not be allowable even if the response time is quite short, and delay would not be excusable even when the accuracy is high enough.

The above discussion is summarized as follows;

(1) Usually the monosyllable input is a heavier load to us than the continuous speech. This verifies the hypothesis-1.

(2) However, there is little difference between the loads of the monosyllable and the continuous speech if the recognition accuracy is high enough and the response time is quite short. This verifies the hypothesis-2. The difference becomes little enough when the accuracy is 95%. The accuracy is not necessarily 100% for disregard of the difference. People

can probably tolerate some recognition errors.

(3) A slow response was a load to people in some degree even if the recognition is complete.
A worse recognition than 80% would not be allowable in spite of a real time response.

## FURTHER EXPERIMENT

The experiment described above contrasted the continuous speech input with the monosyllable input. It is mainly to get a target for our development of VAWP. Furthermore, to compare the continuous speech and the monosyllable with manual typing, we continue evaluation of the VAWP of a simulated continuous speech input and the WP connected to a monosyllable recognition system we have developed.

In this new experiment we compare the monosyllable input and the continuous speech with the input through a keyboard. The recognition accuracy of the monosyllable depends on participants because the monosyllable recognition system reacts to their utterance. The average accuracy of the system is about 70% in the present system. We fix the accuracy of the continuous speech about 60% in the simulated continuous speech recognition system.

Table 3 Guidance for a subject in monosyllable condition

(1) This system transforms your utterance through the microphone into letters. However, KANA(*) - KANJI(**) translation can not be performed; KANA only, and it is without commas and periods.

(2) Press "Input-Key" after uttering each syllable. The corresponding letter will appear on the CRT display.

(3) When you fail to utter the syllable correctly, utter correctly again and press "Input-Key."

(4) Always examine the correctness of the letter on the display, because the system sometimes misrecognizes your utterance.

(5) If recognition error occurs, utter the syllable once again and press "Correction-Key." The correction is effective only for each syllable just after uttering it.

If you missed the recognition error or noticed it, afterwards, leave it.

(6) You can input small "tsu" ( SOKUON ) by pressing "Input-Key" after "tsu"-Key. That is, you have to press two keys.

(7) You do not necessarily have to finish the task quickly. Be careful not to utter a wrong syllable and not to press a wrong key.

Table 4 Guidance for a subject in BUNSETSU condition

(1) This system transforms your utterance through the microphone into letters. However, KANA(*) - KANJI(**) translation can not be performed; KANA only, and it is without commas and periods.

(2) Press "Input-Key" after uttering each BUNSETSU. The corresponding letters will appear on the CRT display.

(3) When you fail to utter the BUNSETSU correctly, utter correctly again and press "Input-Key."

(4) Always examine the correctness of the letters on the display, because the system sometimes misrecognizes your utterance.

(5) If recognition error occurs, utter the BUNSETSU once again and press "Correction-Key." The correction is effective only for each BUNSETSU just after uttering it. You have to utter the whole BUNSETSU even if only one syllable was wrong.

If you missed the recognition error or noticed it afterwards, leave it.

(6) You do not necessarily have to finish the task quickly. Be careful not to utter a wrong BUNSETSU and not to press a wrong key.

(*) KANA means Japanese characters.
(**) KANJI means Chinese characters.



At present, we get the following views;

(1) Most participants prefer to the typing in comparison of the monosyllable to the manual typing.

(2) The evaluation of the continuous speech vs. the typing is better than that of the monosyllable vs. the typing in spite of the inferiority of the accuracy of the continuous speech.

## CONCLUSION

Voice recognition is helpful to us when we compose documents, but recognition error always affects us because the technology is not yet advanced enough for a voice-activated word-processor to have 100% accuracy. It is important to know how the error relates to our composing behavior.

Using a simulated voice-activated word-processor, we experimented on how we are affected by input methods; monosyllable input and continuous speech input, from the view-point of recognition accuracy and response time. Recognition accuracy were 100%, 95%, and 80%, and response time were 0.025 sec and 1 sec.

We got the following experimental results;

(1) When recognition accuracy is 95% and response time is 0.025 sec, the monosyllable input would be acceptable.

(2) When recognition accuracy is 80%, the monosyllable input would not be acceptable regardless of response time.

(3) The continuous speech input would be acceptable even when recognition accuracy is 80% and response time is 1 sec.

(4) Recognition accuracy and response time are psychologically independent factors; we could not say that some error would be acceptable if response time is short enough.

Furthermore we continue estimating the voice-activated word-processor of simulated continuous speech input and the word-processor connected to a monosyllable recognition system we have developed. At present it can be said that the continuous speech input is preferable to the monosyllable input even if the recognition accuracy of the continuous type is worse by about 10% than the monosyllable type. This further experiment is still continued, and we will present the results on the next opportunity.

Table 5 Rating Items

Rate the work just done by 5-point scales on the following items.
Choose one alternative and encircle it.

| | | | | | |
|---|---|---|---|---|---|
| | very much | | | | 5 |
| | rather | | | | 4 |
| | a little | | | | 3 |
| | barely | | | | 2 |
| | never | | | | 1 |
| unfamiliar | 5 | 4 | 3 | 2 | 1 |
| slow | 5 | 4 | 3 | 2 | 1 |
| difficult | 5 | 4 | 3 | 2 | 1 |
| irritating | 5 | 4 | 3 | 2 | 1 |
| boring | 5 | 4 | 3 | 2 | 1 |
| inconvenient | 5 | 4 | 3 | 2 | 1 |
| troublesome | 5 | 4 | 3 | 2 | 1 |
| unencouraging | 5 | 4 | 3 | 2 | 1 |
| tiring | 5 | 4 | 3 | 2 | 1 |

Acknowledgement

The authors would like to thank our colleagues for their helpful discussion. They also thank to directors in Matsushita Electric Industrial Co., Ltd. for their encouragement.

References

[1] Kabasawa, S., et al, (1983), "A study in the error factors in Japanese mono-syllable recognition," Autumn Meeting, Acoust. Soc. Jpn., 3-1-2, pp.183-4, October, ( in Japanese ).

[2] Maehara, F., et al, (1983) "Development of a small size Japanese monosyllable recognition system," Trans. of the Committee on Speech, The Acoustical Society of Japan, S83-40, October, ( in Japanese ).

[3] Gould, J. D., et al, (1983) "Composing letters with a simulated listening typewriter," Comm. of the ACM, Vol.26, No.4, pp.295-308, April.

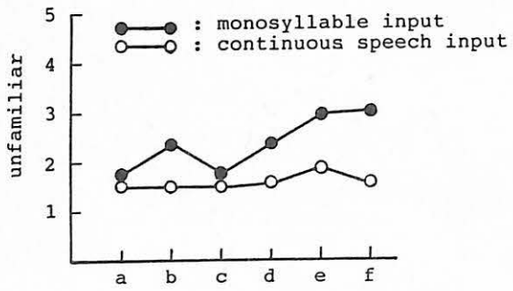


Fig.1 Mean rating value of "unfamiliar"
( see Table 7 )

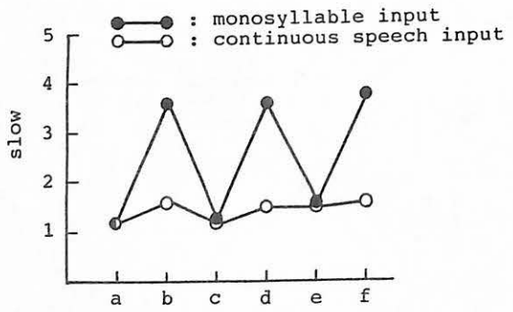


Fig.2 Mean rating value of "slow"
( see Table 7 )

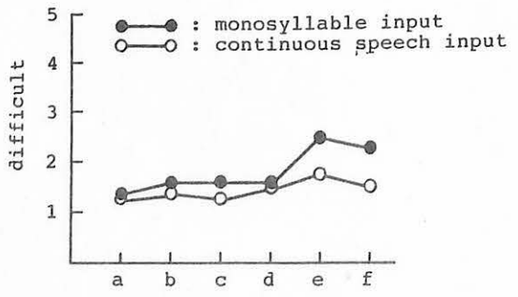


Fig.3 Mean rating value of "difficult"
( see Table 7 )

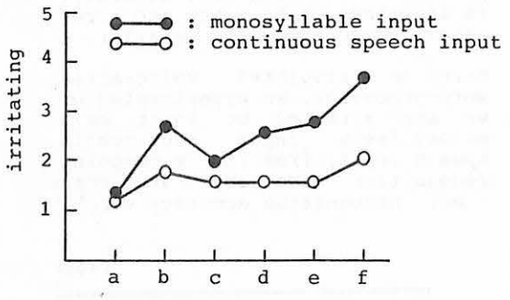


Fig.4 Mean rating value of "irritating"
( see Table 7 )

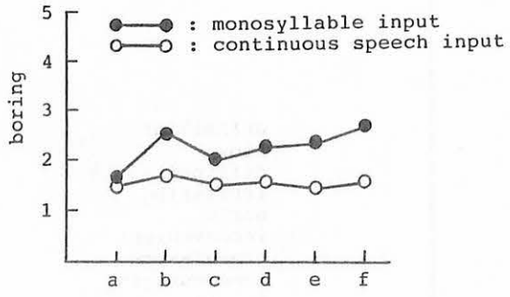


Fig.5 Mean rating value of "boring"
( see Table 7 )



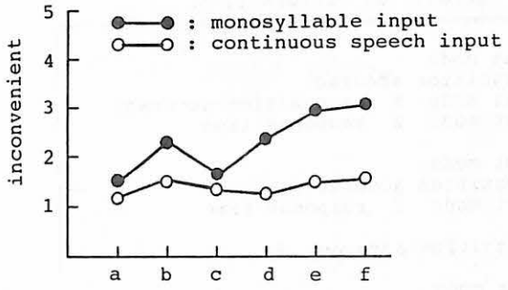


Fig.6 Mean rating value of "inconvenient"
( see Table 7 )

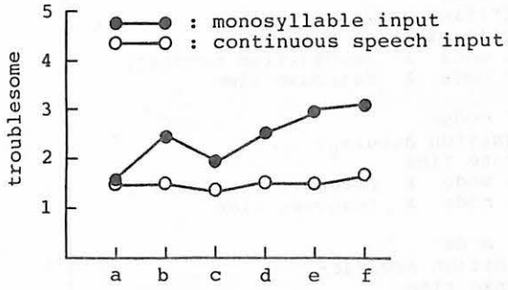


Fig.7 Mean rating value of "troublesome"
( see Table 7 )

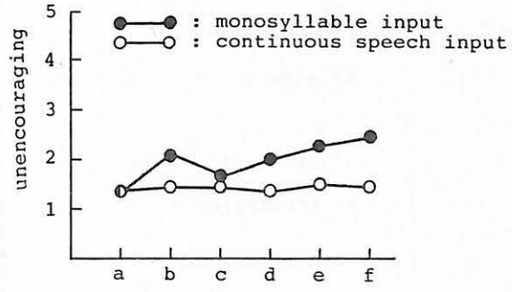


Fig.8 Mean rating value of "unencouraging"
( see Table 7 )

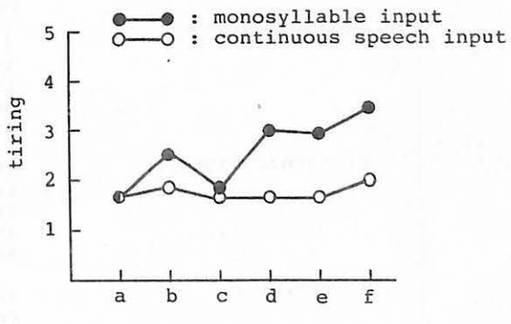


Fig.9 Mean rating value of "tiring"
( see Table 7 )

Table 7 Description of conditions in Figs.1-9

| conditions | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| recognition accuracy(%) | 100 | 100 | 95 | 95 | 80 | 80 |
| response time(sec) | 0.025 | 1 | 0.025 | 1 | 0.025 | 1 |

Table 6 Significant main effects and interactions

| | |
|---|---|
| 1) unfamiliar : | input mode<br>recognition accuracy<br>input mode X recognition accuracy<br>input mode X response time |
| 2) slow : | input mode<br>recognition accuracy<br>input mode X response time |
| 3) difficult : | recognition accuracy |
| 4) irritating : | input mode<br>recognition accuracy<br>response time<br>input mode X recognition accuracy<br>input mode X response time |
| 5) boring : | input mode<br>input mode X recognition accuracy<br>input mode X response time |
| 6) inconvenient : | input mode<br>recognition accuracy<br>response time<br>input mode X recognition accuracy<br>input mode X response time |
| 7) troublesome : | input mode<br>recognition accuracy<br>response time<br>input mode X recognition accuracy<br>input mode X response time |
| 8) unencouraging : | input mode<br>recognition accuracy<br>response time<br>input mode X recognition accuracy<br>input mode X response time |
| 9) tiring : | input mode<br>recognition accuracy<br>response time<br>input mode X recognition accuracy<br>input mode X response time |