This paper describes the development of a portable speech interpreter by which the user can look for words and relating sentences easily using his/her voice. The prototype uses a RISC microprocessor for the embedded use, resulting in a compact body of 15cm(W) and 6cm(H). The voice user interface is helpful, especially in small devices, but it has a serious problem related to the recognition errors. The total user interface must be designed carefully to minimize the inconvenience caused by the recognition errors. We have developed an interface in which the confinement of the dictionary realizes the quick and accurate recognition, and the syllable correction procedure provides an easy way to correct recognition errors. These features make it possible to input any words, which are found in a standard Japanese dictionary. Experimental evaluations show that the N-best recognition rate for words and syllables are acceptable.

1. INTRODUCTION

The progress of the speech recognition technology has created a new big market of portable electronic tools with the voice input function. Some of them are already realized and appeared in the market, but the size of the recognition dictionary is still small in those products. However, it is not convenient for the user to be required the knowledge about the dictionary. If the user does not have knowledge about the dictionary, it could not be judged whether the input word is misrecognized or is not included in the dictionary. This causes serious difficulties in the user interface. The user may repeat the input of an out-of-vocabulary word, believing that the machine is repeating the recognition errors. One of the solutions for this problem is to prepare a dictionary of enough size. Normally, a standard Japanese dictionary contains 30,000 to 100,000 words, so we decided to set the dictionary size to 30,000. If the voice input for the dictionary of this size is realized, one can input any words which appear in his/her mind. At least, the probability that the word is not included in the recognition dictionary is almost same as the probability that the word is not included in a (small) paper dictionary, which is supposed to be acceptable for the user.

Among many applications which are suitable for the above-mentioned specification, we chose a portable interpreter which has the voice-activated interface for the retrieval of the word and relating sentences. We begin this work with Japanese-to-English one-way interpretation, but the extension to other languages is also in our scope.

An interpreter requires a large and general dictionary for input, and has to be small enough for the travel use. The voice input is very useful because buttons and keyboards are inconvenient in such a small device. However, it is technically difficult to recognize the input word quickly and correctly from the 30,000 word dictionary. Therefore, the user interface should be helpful for the reduction of the computational burden, and robust for the recognition errors. In this work, we propose an interface, in which the confinement of the dictionary realizes the quick response of the recognition, and the syllable correction procedure makes it easier to correct the recognition errors. Using combination of these two schemes, the input of any word is guaranteed by the fewest utterances and the fewest key strokes.

The rest of the paper is organized as follows. Section 2 describes the overview of the portable interpreter. Section 3 describes the detail of the proposed user interface. The specification and the estimated performance of the voice recognition are given in Section 4. Finally, conclusions and discussions are given in the last section.
2. PORTABLE INTERPRETER

We have made a prototype of a portable interpreter. The size of the prototype is 15cm (width), 6cm (height), and 3cm (thickness). Figure 1 shows the system architecture of the prototype of the portable interpreter. The main CPU is SH-3 (Super H® Risc Engine series) with the clock speed 60MHz. Program and data are stored in the 16Mbyte flash memory, which are loaded in the 8Mbyte SDRAM for execution. Voice input and output are processed through the AD/DA converter. All the character-based information is displayed on the LCD (30 characters / 4 lines), whereas the detailed monitoring through the RS232C interface is possible.

3. SOPHISTICATED ERROR CORRECTION

Figure 2 shows the action flow of the user interface of the proposed system. First, the user speaks a word to the microphone. The input is recognized and the N-best result is shown on the LCD. At this time, the target dictionary for the recognition is confined within the 3,000 important words, to save the calculation time and to keep the recognition rate high. If the correct word is placed on the higher rank of the N-best result, it is easily selected by the button. Each entry of the N-best list includes the translation of the word, so the process can be terminated if the user is not interested in related sentences. If the correct word can not be found in the list, the user is recommended to proceed to the syllable correction procedure, although it is possible to scroll to the lower rank of the N-best list. At this point, all the feature vectors extracted from the input voice are stored for the purpose of the re-recognition in the later step.

In the syllable correction procedure, one character of the recognized word is selected by the cursor. It means that the left side of the selected character is regarded to be correct. For example, if the user says "YO-YA-KU" (means "reservation" in Japanese) and the recognized word is "YO-U-YA-KU" (means "abstract"), then the user should move the cursor to the second character "U", that means the first character "YO" is correct. After that, the user speaks a syllable "YA" which corresponds to the correct character. Since Japanese "kana" characters basically correspond one-to-one to syllables, this scheme works well for the purpose of character correction. To modify the system to deal with English and other western languages, syllable recognition has to be replaced by the alphabet recognition.

The result of the syllable recognition is also shown as the N-best list. By choosing correct syllable, the left side of the word is fixed. In the above example, the first two characters "YO-YA" are fixed. It means that the target dictionary is confined within the words beginning with "YO-YA." At this time, confinement within the important 3,000 words is unnecessary, because the number of words is small enough if a part of the word is fixed. Even if only the first one character is fixed, the size of the confined dictionary is less than 2,000. After that, re-recognition is carried out using the confined recognition dictionary and the feature vectors.
which are extracted from the input voice and stored in the memory. The result is shown as the N-best list. The user selects the correct word, or goes to the syllable correction procedure if necessary. When the input word is determined, a list of the related sentences is shown. The user can scroll and select a sentence, or input an additional keyword using voice. In the latter case, candidates of the additional keywords are limited and displayed on the LCD, to guarantee the efficient coverage and classification of sentences. In addition, a limited vocabulary of the additional keywords brings high recognition rate, so the syllable correction procedure is not necessary at this step. After the sentence is selected, the user can see the translation and hear the utterance of a native speaker.

4. PERFORMANCE OF VOICE RECOGNITION

Needless to say, the usability of the proposed system is highly dependent on the voice recognition accuracy. In this section, we report the specification and the estimated performance of the voice recognition.

Voice recognition program is based on the software which is developed for general use of Super H microprocessors [1-2], but we have made some modifications to adjust it to the specific application proposed in this work. The word recognition is based on the 2-state 2-mixture Hidden Markov Models (HMMs) for phonetic speech units. There are 464 units to express Japanese pronunciation, and a word consist of 16 units in average. The syllable recognition is based on the 7-state 6-mixture syllable HMMs. There are 68 syllables in Japanese. Both HMMs are semi-continuous, and they share a codebook which has 128 codewords. Input voice is sampled by 12kHz/16bit and processed by the 20ms-long frames (10ms period) to extract feature vectors. A feature vector consists of 14 LPC cepstrums and their time differentials, and the time differential of the log power.

Figure 3 shows the results of recognition experiments. The recognition rate of the word is shown on the left. This experiment corresponds to the first step of the user interface, where the dictionary includes only 3147 important words. The recognition rate for the 1st rank is not high (71.8%) compared with similar experiments [1-2], because the contents of the dictionary is rather general, and there are many pairs of similar pronunciation. Sometimes a word consists of one syllable (There are 17 single syllable words in the dictionary), which is very difficult to recognize. In spite of these problems, the recognition rate within 4th rank is 92.7%, which seems to be acceptable. In this case, the 4th rank corresponds to the display size, which means the user can find the correct answer without scrolling to the next page.

The recognition rate of the syllable is shown on the right side of Figure 3. As mentioned above, the single syllable recognition is more difficult than the word recognition. Therefore, the 1st rank recognition rate is much lower than the word recognition (59.3%), but it rises quickly when the lower ranks are allowed. The
recognition rate is 90.8% within 4th rank, and 97.0% within 8th rank. In addition, the character to be corrected is apparently wrong, so it can be eliminated beforehand from the syllable list. It would make a little improvement of the recognition rate. If we use the knowledge about the connection of characters in words, it would also raise the recognition rate. However, it must be implemented carefully, because the uttered syllable could not recognized if the user is trying to input a forbidden syllable, due to the lack of knowledge about the dictionary.

5. CONCLUSION

We have developed a prototype of portable speech interpreter. The word dictionary of the interpreter contains 30,000 words, but the confinement of the contents realizes the quick response and high recognition rate. Even if the recognition result is wrong, syllable correction procedure using Japanese syllable recognition guarantees the input of 30,000 general words. Sample sentences are bound to the keywords, so the user can easily find a sentence from the input word. Seeing the translation and hearing the utterance is straightforward. Even though the system is working well with the current simple recognition scheme, our future improvement with speaker and environment adaptation, vocabulary learning, and precise HMMs from larger database will create even more convenient user interfaces.

ACKNOWLEDGMENTS

The authors would like to thank Yoshiaki Kitadume, Makoto Tanaka and Hideki Uchidate of Hitachi ULSI Systems Co, Ltd., and Taisuke Kashima, Yukinobu Maruyama and Hitoshi Yamadera of Design Center, Hitachi, Ltd. for designing and developing the hardware of the proposed portable interpreter.

REFERENCES
