SPEECH RECOGNITION BY USE OF WORD DICTIONARY WRITTEN IN LINGUISTIC UNIT

Ken'iti KIDO, Shozo MAKINO, Michio OKADA,
Satoshi MORIAI and Tetsuo KOSAKA

Research Center for Applied Information Sciences
Tohoku University, Katahira, Sendai 980, JAPAN

INTRODUCTION

We have carried out the researches on speaker independent recognition of words [1] by use of word dictionary which is composed of the sequences of phonemic symbols. The phonemic symbols are derived from linguistic representation of Japanese language. In the system, the spoken word is transformed into the sequence of phonemic symbols and the item of the word dictionary most similar to the input sequence is chosen as the recognition output. That is, the system uses the phoneme as the linguistic unit for the recognition of word.

SPEECH RECOGNITION SCHEME

The unit in speech recognition can be classified into two groups: one is based on articulatory model and the second one is not so. The purely acoustical units and the units which refer to the characteristics of auditory organ belong the second group. And the size of unit is also divided into several groups: least one is the segment of speech which is the minimum unit to express word or speech and the maximum one is the word. Figure 1 shows the hierarchical relation between those units. The thick lines between two boxes in Fig. 1 denotes the relations which are considered to be important but difficult to formulate.

Figure 2 shows the schematic diagram of speaker independent spoken word recognition system we have developed. In the system, the recognition is to find out the item of word dictionary which corresponds to the input speech. And the system is equipped with word dictionary which contains all the words to be recognized.

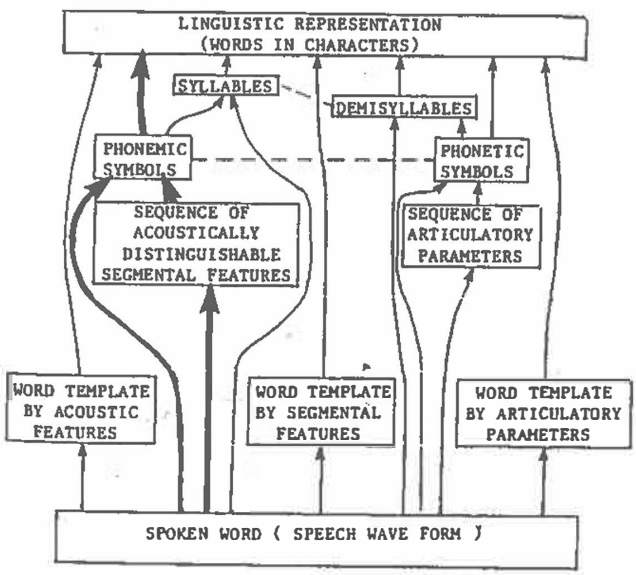

Fig. 1 Hierarchy in the unit of representation of spoken words for speech recognition

In the system, the input speech is transformed into a sequence of phonemic symbols. And the similarity of the content of word dictionary to the input speech is computed for every item. The recognition output is the dictionary item of maximum similarity.

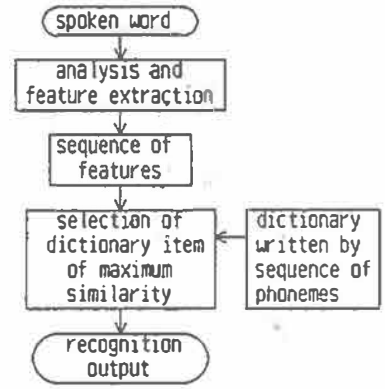

Fig. 2 Schematic diagram of the spoken word recognition system

PHONEME AS LINGUISTIC UNIT

The most important problem in such the system is how to describe the contents of word dictionary. If the contents are described by phonemic symbols, it may be very simple to make the word dictionary especially in Japanese as all the Japanese words are in the form "CVCVCV..." where C denotes the consonant and V the vowel. But the transformation of speech into the sequence of phonemic symbols is not easy because the acoustic characteristics of speech segment does not always correspond to the phonemic symbol which are derived from the linguistic representation.

If the contents are the standard patterns composed of acoustic features directly obtained by analyzing the spoken words, it would be easy to transform the input speech to the patterns for the comparison with the standard patterns. But, a lot of computation is necessary for making the standard patterns common to all the possible speakers especially in the case of large vocabulary.[2]

And there is intermediate system[3] in which the word dictionary is composed of the sequences of acoustic features which are defined by classifying the words uttered by a number of speakers. The classification is based on the differences in acoustic characteristics of speech segments. Such the features may be able to express the acoustic characteristics of words more exactly than the phonemic symbols. The phonetic transcription may be exactly carried out using such the features, and we call the features as the phonetic features in this paper. The transformation of the input speech into the sequence of the phonetic features may be easier than the transformation into the sequence of phonemic symbols. But, a lot of computation and a number of speech samples will be necessary for making the word dictionary composed of such the phonetic features and it may be difficult problem to compose a set of phonetic features which can be used for many vocabulary regardless of speakers.

Therefore, we have used the phonemic symbols for the description of dictionary items and now we are trying to use the acoustic features of segments to derive the sequence of phonemic symbols.

CONVERSION OF SPEECH INTO PHONEMIC SYMBOLS

The input speech is passed through a 29 channel band pass filter bank which is composed of single tuned circuit of Q=6 and the center frequencies are at every 1/6 octave between 250 Hz and 6 300 Hz. The power of every channel is computed for every frame of 10 ms duration and logarithmically transformed.

Eight features are extracted by using the discriminant filters which are designed by use of speech samples of 212 words uttered by 10 male and 10 female speakers. Figure 3 shows the examples of the solution weight vectors for eight discriminant functions. Another feature is the logarithmic spectrum summation which is the sum of logarithmic power of all the channels.

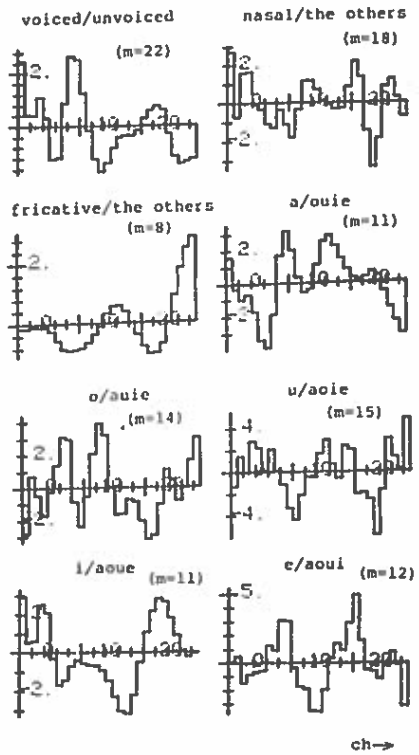

Fig. 3 Examples of solution weight vectors for extracting the eight features

Table 1 Discriminant functions

function X_1/X_2	X_1	X_2
voiced/ unvoiced	/a/,/o/,/u/,/i/, /e/,/j/,/w/,/m/, /n/,/n/,/N/,/b/, /d/,/g/,/r/,/z/	/h/,/s/,/c/ /p/,/t/,/k/
nasal/ the others	/m/,/n/,/n/,/N/	/a/,/o/,/u/,/i/,/e/ /j/,/w/,/b/,/d/,/g/ /r/,/z/,/h/,/s/,/c/ /p/,/t/,/k/
fricative/ the others	/z/,/h/,/s/,/c/	/a/,/o/,/u/,/i/,/e/ /j/,/w/,/m/,/n/,/n/ /N/,/b/,/d/,/g/,/r/ /p/,/t/,/k/
a/ouie	/a/	/o/,/u/,/i/,/e/
o/auie	/o/	/a/,/u/,/i/,/e/
u/aoie	/u/	/a/,/o/,/i/,/e/
i/aoue	/i/	/a/,/o/,/u/,/e/
e/aoui	/e/	/a/,/o/,/u/,/i/

5 vowels, 2 semi vowels, 15 consonants

The functions of the discriminant filters for the eight features are listed in Tab. 1. The phoneme boundaries are assumed to be the frame where the weighted sum of absolute values of the first order time-derivatives of the features takes maximum value exceeding a threshold. The frames of unvoiced and voiced plosives are detected using the discriminant filters. The primary phoneme recognition is carried out for every assumed segment using the outputs of discriminant filters and the standard patterns for phonemes which are made using the 212 spoken words.

After correcting errors by the errorcorrection rules, the secondary phoneme recognition is carried out. Here, the nasals and the voiced and unvoiced plosives are recognized.

WORD RECOGNITION USING LINGUISTIC UNIT

In the word recognition part, a number of subitems are generated referring to the confusion matrices of phoneme recognition for initial-, mid- and final positions of words. The confusion matrices includes the probabilities of insertion, omission and substitution of phoneme. The computation of similarity between the phonemic sequence with top three recognition results and each sub-item follows. The dynamic programming algorithm is used to reduce the time for similarity computation.

The dictionary item having maximum similarity to the input sequence is chosen as the recognition output.

RECOGNITION EXPERIMENTS

Word recognition experiments were carried out using the speech samples used to design the discriminant functions, standard patterns and confusion matrices and the same 212 words uttered by the other 30 males and 20 females. Table 2 shows the summary of the results.

Table 2 Word recognition score

Training set	10 males	93.7%	Average
	10 females	91.3%	92.4%
Test set	30 males	87.0%	Average
	20 females	89.6%	88.1%

CONCLUSION

This paper describes the use of phoneme as the linguistic unit of speech in the spoken word recognition system for a large vocabulary. In the system, the phoneme recognition is first carried out and the word dictionary item with the maximum similarity to the sequence of recognized phonemes is chosen as the recognition output. The score of word recognition is 92.4% in the experiment which is much higher than that of phoneme recognition (75.9%) due to the utilization of word dictionary as the linguistic information source. Studies on phoneme recognition is now continued to improve the word recognition. The vocabulary to be recognized can easily be altered and expanded by changing the dictionary item from key board.

LITERATURES

1) H.Suzuki,S.Itahashi and K.Kido: TheEffectiveness of Utilization of Word Lexicon in Recognition of Japanese Spoken Language, Proc.1967 Conference on Speech Communication and Processing,B12,p.128(1967)
2) S.Chiba: Recognition of Spoken Words,J.Info.Proc. Jpn.19-7(1978)
3)N.Sugamura,S.Furui: Large Size Vocabulary Spoken Word Recognition by Use of Pseudo-phoneme Standard Patterns,J.IECE.Jpn,65-D,8(1982)