A speaker independent word recognition system based on phoneme recognition for a large size (212 words) vocabulary

A SPEAKER INDEPENDENT WORD RECOGNITION SYSTEM BASED ON PHONEME RECOGNITION FOR A LARGE SIZE (212 WORDS) VOCABULARY

Shozo Makino and Ken'iti Kido

Research Center for Applied Information Sciences,
Tohoku University,Sendai 980, Japan.

ABSTRACT

This paper describes the speaker-independent spoken word recognition system for a large size vocabulary. Speech is analyzed by the filter bank, from whose logarithmic spectrum the 11 features are extracted every 10 ms. Using the features the speech is first segmented and the primary phoneme recognition is carried out for every segment using the Bayes decision method. After correcting errors in segmentation and phoneme recognition, the secondary recognition of part of the consonants is carried out and the phonemic sequence is determined。 The word dictionary item having maximum likelihood to the sequence is chosen as the recognition output. The 75.9% score for the phoneme recognition and the 92.4% score for the word recognition are obtained for the training samples in the 212 words uttered by 10 male and 10 female speakers. For the same words uttered by 30 male and 20 female speakers different from the above speakers, the 88.1% word recognition score is obtained.

1. INTRODUCTION

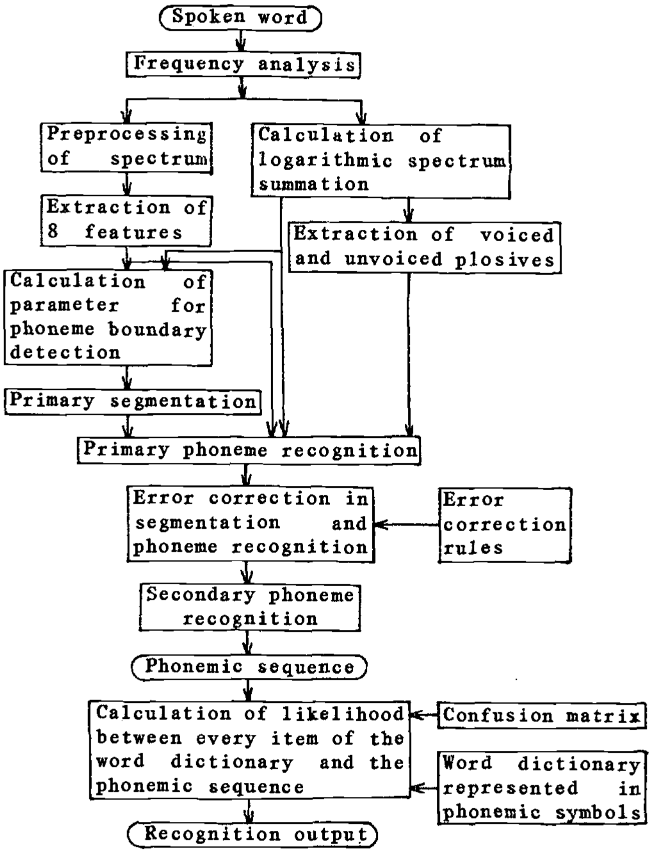
A number of speaker independent systems[1)-7)] has been presented so far。 All systems, except for those developed by Chiba et al。[4)] and Rabiner et al。[6),7)], were based on the phoneme recognition, where recognized speech samples were uttered only by adult male speakers. The systems developed by Chiba et al. and Rabiner et al. require a large amount of data uttered by a number of speakers, in the case where the change in vocabulary is required.

On the contrary, the new spoken word recognition system described in this paper can recognize a large size vocabulary uttered by unspecified adult male and female speakers at a high performance. The system is composed of the phoneme recognition part and the word recognition part, where segmentation and phoneme recognition are carried out. The vocabulary can easily be changed from the key-board since the word dictionary is represented in phonemic symbols.

2. SYSTEM OUTLINE

Figure 1 shows a schematic diagram of the spoken word recognition system based on the phoneme recognition.

Speech is passed through the 29channel digital bandpass filters (single tuned circuits of Q=6), whose center frequencies are arranged every 1/6 octave between 250Hz and 6300 Hz. The power of every channel is computed for every frame of 10 ms duration and logarithmically transformed. The 9 features are extracted from the logarithmic spectrum of each 10 ms-frame。 The primary segmentation is carried out using the first order time-derivative of the features. Using the 11 features composed of the above 9 and other two features extracted from the temporal pattern of logarithmic power, the phoneme of the detected segment is recognized by applying the Bayes decision method, on the assumption of multi-dimensional normal distribution. Errors in the primary segmentation and phoneme recognition are corrected by the error correction rules.


Spoken word
Frequency analysis
Preprocessing of spectrum
Calculation of logarithmic spectrum summation
Extraction of 8 features
Extraction of voiced and unvoiced plosives
Calculation of parameter for phoneme boundary detection
Primary segmentation
Primary phoneme recognition
Error correction in segmentation and phoneme recognition
Error correction rules
Secondary phoneme recognition
Phonemic sequence
Calculation of likelihood between every item of the word dictionary and the phonemic sequence
Confusion matrix
Word dictionary represented in phonemic symbols
Recognition output

Fig。1 Schematic diagram of spoken word recognition system

Meanwhile, the secondary phoneme recognition is carried out using the time spectrum pattern method previously proposed[8] for nasals, unvoiced and voiced plosives and unvoiced fricative. The likelihood between the phonemic sequence and each item of the word dictionary is calculated, where the item with maximum likelihood is chosen as the recognition output.

3. SEGMENTATION AND PHONEME RECOGNITION

3.1 Feature extraction using discriminant analysis[9]

The eight features are computed for every frame from the logarithmic spectrum using the discriminant analysis.

For the feature expressed by the weighted sum of p parameters the condition for the optimal discrimination of a specified phoneme group from the other is given by the maximization of Fisher ratio θ:

$$\theta = \mathbf{a}^t\mathbf{B}\mathbf{a}/\mathbf{a}^t\mathbf{W}\mathbf{a} \longrightarrow \max \tag{1}$$

$\mathbf{W}$: Covariance matrix within class
$\mathbf{B}$: Covariance matrix between classes
$\mathbf{W}$ is rewritten using the KL expansion

$$\mathbf{W} = \mathbf{F}\,\Lambda\,\mathbf{F}^t \tag{2}$$

$$\Lambda = \left\{\begin{matrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ & & \ddots & \\ 0 & & & \lambda_m \end{matrix}\right\}$$

$$\mathbf{F} = \{\, f_1, f_2, \text{---}, f_i, \text{---}, f_m \,\}$$

λ_i : Eigen value of $\mathbf{W}$
f_i : Eigen vector of $\mathbf{W}$

where rank of $\mathbf{W}$ is m less than p.
Solution weight vector $\mathbf{a}$ is given by Eq.(3).

$$\mathbf{a} = \mathbf{F}\Lambda^{-1}\mathbf{F}^t\mathbf{d} \tag{3}$$

$\mathbf{d}$: The distance vector between mean vectors.

3.2 Phoneme boundary detection

The following power parameter is used for the detection of phoneme boundary.
(1) Logarithmic spectrum summation LS

$$LS = \sum_{i=1}^{29} S_i \tag{4}$$

S_i : Logarithmic output of the i-th channel

The eight features and the logarithmic spectrum summation LS show the large changes at the phoneme boundaries. For that reason they are passed through the first order time-derivative filters. The solution weight vector for phoneme boundary detection is computed using the discriminant analysis. The weighted summation of the absolute values of the time-derivative filter output is computed frame by frame. The frame with the local maximum of the temporal pattern of the weighted summation is assumed to be the segment boundary. On the other hand, the frame with the local minimum of the temporal pattern is also assumed to be the typical frame of the segment. There are a number of segment insertions which are mostly eliminated using the higher level information.

3.3 Recognition of phonemes

The two parameters for unvoiced and voiced plosives are computed using the discriminant filters from the temporal envelope of the logarithmic spectrum summation LS. The output y_j of the discriminant filter for the j-th frame is defined as follows:

$$y_j = \sum_{i=-N}^{-N} c_i v_{j+i} \tag{5}$$
$$= \mathbf{c}\cdot\mathbf{v}_j$$

$$\mathbf{c} = (\, c_{-N}, \cdots, c_0, \cdots, c_N \,)$$

$$\mathbf{v}_j = (\, v_{j-N}, \cdots, v_j, \cdots, v_{j+N} \,),\ N=7$$

$\mathbf{c}$: Solution weight vector
$\mathbf{v}_j$: Sequence of LS for the j-th frame
v_j : Logarithmic spectrum summation LS for the j-th frame

where the solution weight vector $\mathbf{c}$ is computed using the discriminant analysis for optimally discriminating between plosive and the others.

The eight features, the logarithmic spectrum summation LS and the two parameters for plosives are used for phoneme recognition. Figure 2 shows the solution weight vectors for part of the features. The shapes show the good correspondence to the physical images. The feature for voiced plosive is useful for discriminating between consonants and vowels. Figure 3 shows the temporal pattern of the 11 features and the detection parameter of the phoneme boundary for the word /rakuda/ uttered by a adult male, where each of the horizontal lines indicate the threshold for the feature. All features are correctly extracted.

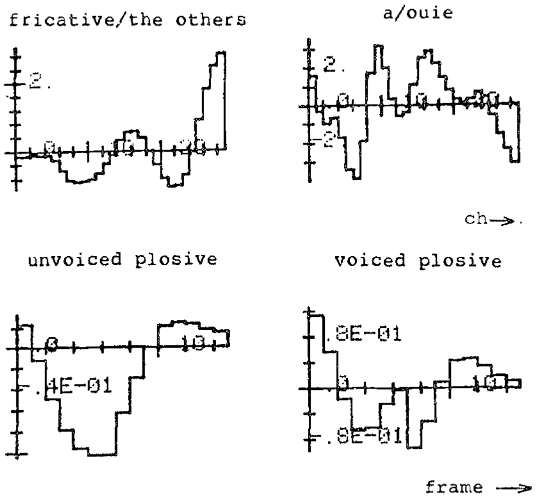


Fig. 2 Solution weight vectors for part of the features

Using aforementioned 11 parameters, the phoneme recognition is carried out for the typical frame by applying the Bayes decision method, assuming multi-dimensional normal distribution. Standard pattern of each phoneme is made using the samples manually extracted by visual inspection.

4. ERROR CORRECTION IN SEGMENTATION AND RECOGNITION

In order to correct the errors in segmentation and phoneme recognition, the following information is used:
(1) phoneme contexts,
(2) the contexts where the omissions occur, and
(3) phoneme duration.

Based on these information, the five correction rules are defined as follows:

(1) When the succeeding two or more segments are recognized as the same phoneme, those segments are merged into one segment unless the total duration exceed the upper limit of the duration.
(2) The segment with duration shorter than the lower limit is absorbed by the surrounding segment, if one of the top-three recognition results for the shorter segment is identical with one of the surrounding segment.
(3) When the vocalic segment appears after the non-sound segment, the plosive segment is inserted between the two segments.
(4) When the succeeding two segments are recognized as unvoiced phoneme and each of these segments satisfies the condition of duration, the devocalized vocalic segment is inserted between the segments.
(5) When the succeeding two segments are recognized as vowels, phoneme recognition is carried out for the frames around the boundary between these segments; a new phoneme is inserted if the result satisfies the certain conditions.

Using the rules above, most of the errors in segmentation and recognition of phonemes are corrected.

5. SECONDARY PHONEME RECOGNITION USING TIME SPECTRUM PATTERN METHOD[8)]

Secondary phoneme recognition is carried out using the time spectrum pattern(TSP) method for discriminating among the phonemes with the same manner of articulation. The TSP method assumes the time spectrum pattern(29channels by 5frames) of a phoneme as the multi-dimensional vector derived from the normal distribution. Recognition of consonants using the TSP method is carried out for the frames around the epoch point, where the epoch point for nasals is the final stationary frame from the nasal consonant to the following vowel and that for plosives is the burst frame. The epoch point is detected by the change in the logarithmic power of the specified frequency range. The standard patterns are made using samples around the epoch point manually extracted by visual inspection. The dimension of the time spectrum pattern is reduced from 145 to 24 using the Karhunen-Loeve expansion. Secondary recognition is carried out using the TSP method if the result of recognition in the previous stage belongs to one of the four phoneme groups: /m,n,ŋ/, /p,t,k/, /b,d,g/ and /s,c/.

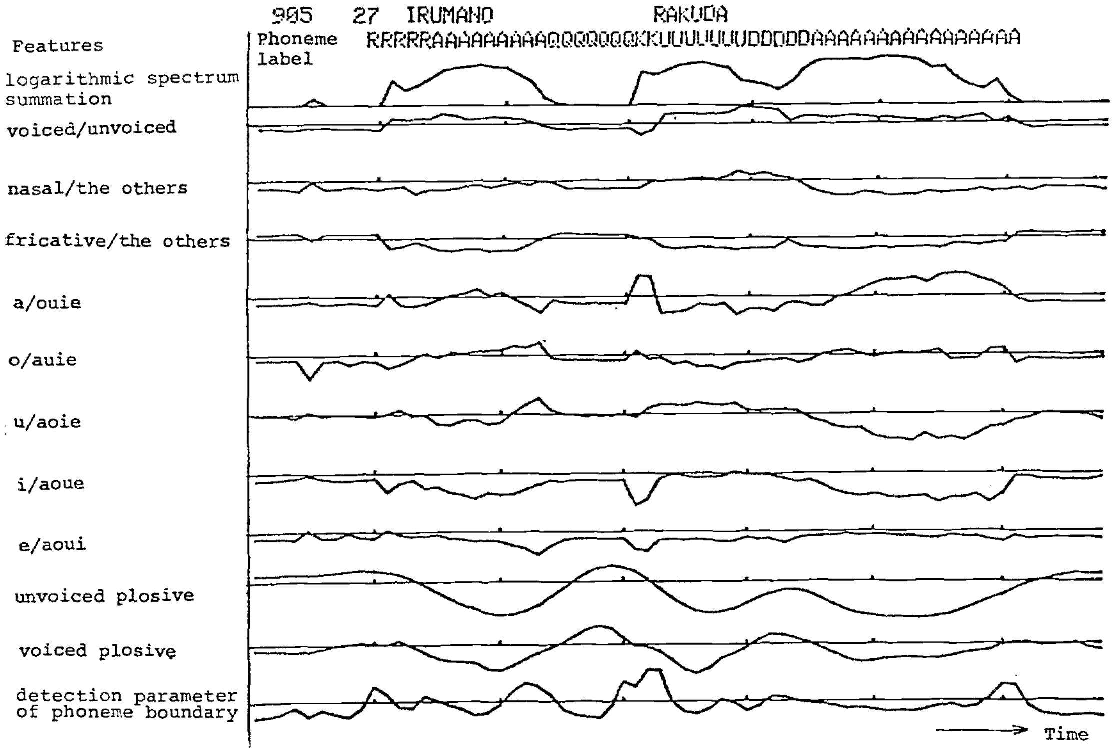


Fig. 3 The temporal pattern of the 11 features and the detection parameter of the phoneme boundary for the word /rakuda/

Thus determined is the phonemic sequence with the top-three results sent to the word recognition stage.

6. WORD RECOGNITION BASED ON THE PHONEMIC SEQUENCE

In the word recognition part, a number of sub-items are generated referring to the confusion matrices of phoneme recognition for initial-, mid- and final positions of words. This is followed by the computation of similarity between the phonemic sequence with the top-three recognition results and each sub-item of of the word dictionary item. Probabilities of insertions, omissions and substitutions of phonemes included in confusion matrices are taken into account for computation. The rank of phoneme recognition is also taken into account. The confusion matrices are made according to the results of experiments. The computation time for similarity including generation of sub-items is considerably reduced by the utilization of dynamic programming. The dictionary item having the maximum similarity to the sequence is chosen as the recognition output.

7. RECOGNITION EXPERIMENTS

Phoneme recognition experiments are carried out for 212 words uttered by 10 male and 10 female speakers. The solution weight vectors for the features and the standard patterns for the phonemes are made using the same sample word group. By using the TSP method, the scores for the unvoiced and voiced plosives are increased by 18% and 20%, respectively. The scores of phoneme omission and insertion are 6.4% and 14.7%, respectively, whereas the phoneme recognition score is 75.9%.

Table 1 shows the word recognition scores。 The word recognition score of 92.4% is obtained for the samples from which the confusion matrices for the word recognition are made. For the same words uttered by 30 male and 20 female speakers different from the above-mentioned speakers, the word recognition score of 88.1% is obtained. The scores for female speakers are nearly the same as those for male speakers.

8. CONCLUSION

This paper describes a spoken word recognition system for a large vocabulary recently developed. The score of phoneme and word recognition are 75.9% and 92.4%, respectively. The system recognizes the spoken words based on the phoneme recognition; word recognition is carried out using the likelihood between the phonemic sequence transformed from the input speech and the word dictionary item written in phonemic symbols. The vocabulary to be recognized can easily be altered by changing the dictionary item from the key-board. The large vocabulary size is also one of the features of the system.

Table 1 Word recognition scores

Training set	10 males	93.7%	Average 92.4%	4240 samples
	10 females	91.3%		
Test set	30 males	87.0%	Average 88.1%	10600 samples
	20 females	89.6%		

REFERENCES

(1) Itahashi S., S. Makino and K. Kido, "Discrete-Word Recognition Utilizing a Word Dictionary and Phonological Rules", IEEE Trans., AU-21, 3, pp239-248(June 1973)

(2) Sakai T. and S. Nakagawa, "A Classification Method of Spoken Words in Continuous Speech for Many Speakers" (in Japanese), Information Processings Soc。 Japan, 17, 7, pp650-657(July 1976)

(3) Kido K., T。 Matsuoka, J. Miwa and S。 Makino," Spoken Word Recognition System for Unlimited Adult Male Speakers", Trans. IECE Japan, E61, 8, pp593-598(Aug. 1978)

(4) Chiba S., M。 Watari and T. Watanabe, " A Spoken Word Recognition System for Unlimited Speakers "(in Japanese), presented at the meeting of IECE Japan, 219(Aug. 1977)

(5) Miwa J., Y. Niitsu, S. Makino and K. Kido," Spoken Word Recognition Systems using Gross Features of Speech Spectrum and these Dynamic Properties"(in Japanese), J. Acoust. Soc. Japan, 34, 3, pp186-193(Mar. 1978)

(6) Rabiner L. R.: " On Creating Reference Templates for Speaker Independent Recognition of Isolated Words", IEEE Trans., ASSP-26, 1, pp34-42(Feb. 1978)

(7) Rabiner L. R. and J. G. Wilpon, " Speaker-Independent Isolated Word Recognition for a Moderate Size(54 Word) vocabulary" , IEEE Trans., ASSP-27, 6, pp583-587(Dec. 1979)

(8) Ide K., Makino S. and Kido K., " Recognition of unvoiced plosives using Time Spectrum Pattern " (in Japanese), J。 Acoust. Soc. Japan, 39, 5, pp321-329(Aug. 1983)

(9) Kawabata T。, Makino S. and Kido K。, "Feature Extraction of Phoneme Based on the Discriminant Analysis" (in Japanese), Trans. of the IECE of Japan, J65-A,12, pp1278-1285(Dec. 1982)