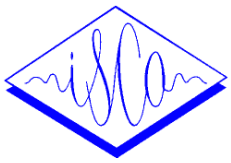

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Speaker Independent Speech Recognition Method Using Phoneme Similarity Vector

Masakatsu Hoshimi, Maki Yamada and Katsuyuki Niyada

Matsushita Research Institute Tokyo Inc.
3-10-1 Higashimita Tama-ku , Kawasaki , 214 JAPAN

ABSTRACT

We are developing a speaker independent speech recognition method using the similarity vectors as feature parameters. In this paper we present the feasibility of the technique in practical usage by downsizing the algorithm and improving the word spotting technique.

When tested with 100 Japanese city names in noisy environments, a word recognition rate 95.9% was obtained with reduced memory size and computation amount. We expect that the method can be implemented in a small hardware.

1.INTRODUCTION

Usually standard patterns for speaker-independent speech recognition are made by statistically processing speech data of speakers. There are several matching methods: for example, a method using the statistical distance measure, and a method applying models such as the neural net and HMM. Especially, numbers of successful HMM are reported using the continuous mixture Gaussian density models. With these methods, spectral parameters are used in speech recognition as a feature parameter and an enormous number of speakers are generally required for training.

If the standard patterns for speaker independent speech recognition can be produced from a small number of speakers, man power and computer power are saved and speech recognition technique can be easy to handle in various applications.

For the purpose mentioned above, we have already studied a speech recognition method using the similarity vectors as feature parameters[1]. In this method, word templates trained with a small number of speakers yield high recognition rates in speaker-independent recognition. We've also reported the high performance by using the word templates made by concatenating sub-word units such as CV and VC, extracted from a small number of speakers[2].

To realize the speech recognition technology in real applications, speech recognizer must be robust to noisy environments and spot intended words from background noise and unintended utterances. Furthermore speech recognizer must retain high quality performance on portable devices. For these reasons, we focused on word spotting technique and downsizing of the algorithm in the recognition method.

2.RECOGNITION METHOD

2.1 Concept

Speech signal is generated according to a shape of a vocal tract and its temporal transition. The shape of the vocal tract, which depends on the shape or size of the vocal organs, inevitably shows individual differences. On the other hand, the patterns of time sequence of the vocal tract, which depend on an uttered word, show a small individual difference. Therefore features of utterance should divided into two factors: the shape of the vocal tract and its temporal pattern. The former shows large difference from speaker to speaker whereas the latter shows small difference. So, if the difference based on the shape of the vocal tract is somehow normalized, the speech of unspecified speakers can be recognized using only the utterances of a small number of speakers.

The difference in the shape of the vocal tracts causes different frequency spectra. One of the methods to normalize the spectral difference among speakers is to classify voice input by matching it with phoneme templates which are made for unspecified speakers. This

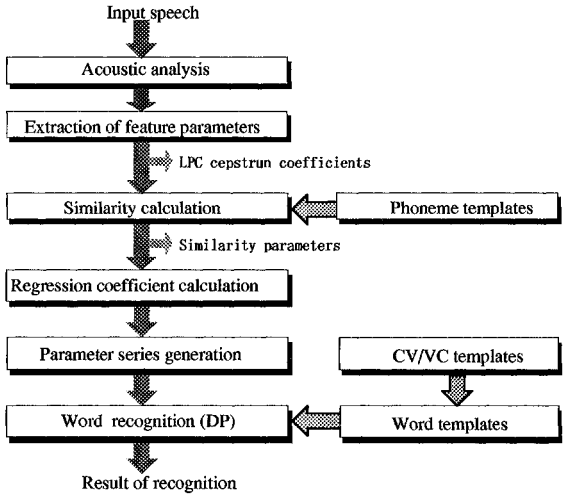

Fig.1 Outline of a speaker independent word recognition

operation provides similarities which does not depend very much on the differences among speakers. On the other hand, the temporal pattern of vocal tract is considered to have small individual differences[3].

2.2 Algorithm

This algorithm of speaker independent word recognition method is outlined in Fig.1. First, the acoustic analysis part analyzes speech inputs and extracts LPC cepstrum coefficients and delta power. The extracted parameters are matched with 24 kinds of phoneme templates, and static phoneme similarity and the first order regression coefficients of phoneme similarity are calculated in the similarity calculation part. Then a time sequence of 24-dimensional similarity coefficient vectors and 24-dimensional regression coefficient vectors is obtained.

In the similarity calculation part simplified mahalanobis' distance employed for distance measure, where covariance matrixes for all of the phonemes are assumed to be identical. The input vector $\boldsymbol{c}$ is composed of LPC cepstrum coefficients and delta power in 10 frames. The input vector $\boldsymbol{c}$ is expressed as

$$\boldsymbol{c} = (v^1, c_0^1, c_1^1, \cdots, c_{13}^1, \cdots, v^{10}, \cdots, c_{13}^{10})^t \quad (1)$$

where c_i^k denotes the i-th LPC cepstrum coefficient of the k-th frame and v^k denotes delta power of the k-th frame.

Standard phoneme templates are trained by 212 word set spoken by 20 speakers. They are made from time-spectral patterns around distinctive frames for each phoneme. We define the distinctive frame as epoch frame. For example, the epoch frames of vowels are in the middle of duration and those of unvoiced consonant are at the end of duration. The phoneme similarity between input vector $\boldsymbol{c}$ and phoneme template (phoneme p) is calculated as

$$L_p = \boldsymbol{a}_p \cdot \boldsymbol{c} - b_p \quad (2)$$
$$\boldsymbol{a}_p = 2\sum^{-1} \cdot \mu_p$$
$$b_p = \mu_p \cdot \sum^{-1} \cdot \mu_p$$

where μ_p is a mean vector of phoneme p, and $\sum$ is the covariance matrix.

Fig.2 shows an example of time sequence of phoneme similarities (uttered "ASAHI"). While increase of similarity scores are observed for phonemes /a/ , and /o/ and /u/ which are phonetically close to /a/ in a segment of /a/ , there is almost no change in similarities of the other phonemes in the same segment. In a transient part from /a/ to /s/, similarity of /a/ is gradually decreasing and similarities of phonemes /s/ and /h/ are increasing contrary. After the static phoneme similarities are obtained, regression coefficients of the phoneme similarities are computed using static phoneme similarities over 50 msec.

Next, word templates are produced by concatenating sub-word units such as CV and VC trained from a few speakers' speech. There are 563 kinds of sub-word units according to our definitions. The trajectory of the similarity and regression coefficients are averaged for each sub-word unit and stored in a sub-word dictionary.

When recognizing input speech, the time sequences of the similarity vector and regression coefficients vector for each frame are calculated as feature parameters. These time sequences of the feature parameters of input speech and reference in the dictionary are compared in DP matching and the most similar word is selected as a recognition result. The

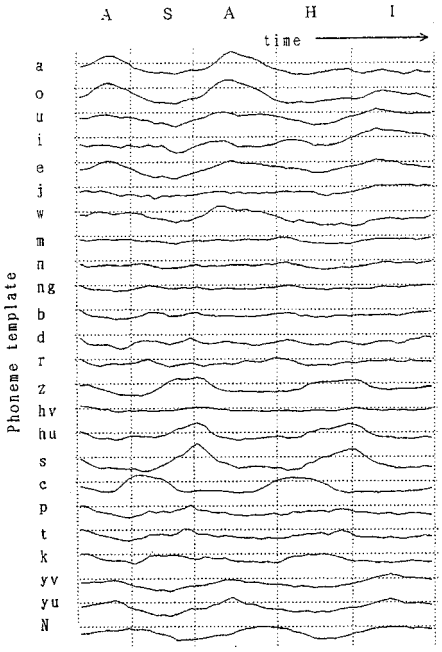

Fig.2 An example of time sequence of phoneme similarities (uttered "ASAHI")

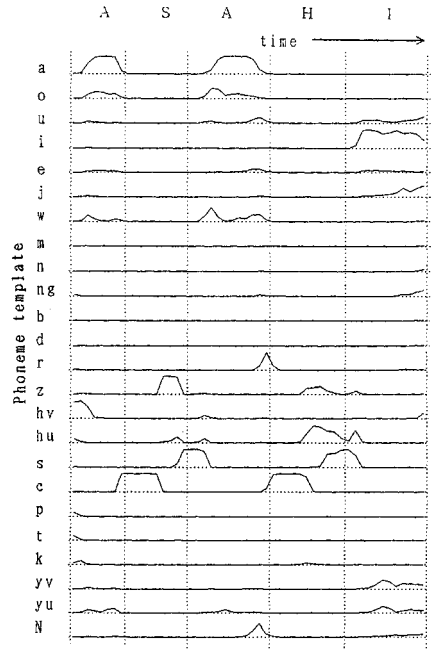

Fig.3 An example of time sequence of normalized phoneme similarities (uttered "ASAHI")

correlation cosine as shown in eq.(3) is used to calculate the partial similarity score function of $s(i,j)$.

$$s(i,j) = w\frac{\boldsymbol{d}^i \cdot \boldsymbol{e}^j}{|\boldsymbol{d}^i| \cdot |\boldsymbol{e}^j|} + (1-w)\frac{\Delta\boldsymbol{d}^i \cdot \Delta\boldsymbol{e}^j}{|\Delta\boldsymbol{d}^i| \cdot |\Delta\boldsymbol{e}^j|} \quad (3)$$

where $\boldsymbol{d}^i$ denotes a similarity vector in the i-th frame of input, $\boldsymbol{e}^j$ denotes an similarity vector in the j-th frame of reference, and $\Delta\boldsymbol{d}^i$ and $\Delta\boldsymbol{e}^j$ are the respective regression coefficient vectors, and "w" is the mixing ratio between scores from the similarity vector and its regression coefficient vector. There are wide individual difference in the absolute values of similarity. However, since there is less individual difference in the correlation of similarity values, the correlation cosine is useful to eliminate individual differences.

3. ALGORITHM DOWNSIZING

3.1 Concept

This recognition method is characterized by using phoneme similarity as feature parameter and correlation cosine as distance measure. This section presents methods of efficient computation based on these characteristics of the recognition method.

Phoneme similarity which is produced by matching between a particular phoneme standard pattern and input speech at each frame has a large value at an epoch frame and a trivial value at other frames. Therefore, it is reasonable to consider that, among many components of similarity vector, few components with large values has dominant effect on recognition performance and that omission of the other components with trivial values doesn't cause bad effect.

Furthermore, since phoneme similarity vector is normalized by its norm and has only relative value, it's quite natural to assume that it doesn't need to be expressed with high precision.

Fig.3 shows time sequences of normalized similarities of a speech example "A SA HI." Near the epoch frame of a certain phoneme, similarity of the corresponding phoneme is particularly large and those of similar phonemes are a little large. On the other hand similarities of other phonemes remain quite low. These observations suggest that, to compute partial scores, a certain number of components of a similarity vector which has N largest similarities in a word template are sufficient and other components can be ignored.

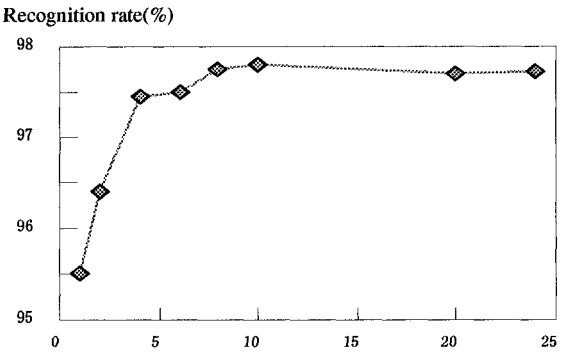

Fig. 4 Recognition rate versus number of selected phonemes

3.2 Experimental Result

Based on the above prediction, recognition experiments were carried out with clean test data(100 Japanese city name). Fig.4 shows recognition rate versus N, the number of components used computation of partial score, where phoneme similarities of N largest values and delta phoneme similarities of N largest absolute values are selected from 24 phonemes respectively. In Fig.4 horizontal axis is for the number of phoneme and vertical axis is for word recognition rate. Fig.4 suggests us that recognition performance remains sufficient when 4 or more phonemes are used. Computation amount with selected 6 phonemes is one-fourth of that with 24 phonemes.

Furthermore, Table.1 shows recognition rates versus precisions of phoneme similarity vector, where selected 6 components of phoneme similarity vector is expressed in the listed precisions. In Table.1 32 bits means floating, and 4, 6 and 8 mean bit numbers in integer expression. It shows that the reduction of precision doesn't decrease recognition rate. Total memory size for a sub-word (CV.VC) dictionary can be reduced to one thirty second of the original one when it consists of selected 6 phonemes out of 24 phonemes with 4 bit integer precision instead of 32 bit floating.

With the reduction of computation amount and memory size, a high quality recognizer can be realized with just one DSP.

Table 1 Recognition rate versus precisions of phoneme similarity vector

Precision of feature parameters	32bit	8bit	6bit	4bit
Recognition rate(%)	97.5	97.5	97.5	97.3

4. WORD SPOTTING

4.1 Word Spotting Algorithm

In this section, word spotting technique in this method is described. This method uses correlation cosine as distance measure and phoneme similarity vector is normalized by it's size at each frame. This normalization operation allocates similarity scores to certain phonemes even in non-speech segment. In other word, the normalization in this method derives non-speech information from sample data including no-speech segment. The lack of non-speech information makes word spotting difficult and reduces recognition performance of this method in practical use.

To deal with such a problem, posteriori probability is introduced to partial score calculated by correlation cosine. To make partial score $s(i,j)$ have posteriori probability from, we need to have probability $P(k/s)$ that a lattice point $k(i,j)$ is on optimal path with partial score s. $P(k/s)$ is provided by Bayes' Rule as

$$P(k|s)=\frac{P(s|k)}{P(s)}P(k) \quad (4)$$

where $P(s/k)$ is the probability that the partial score s is observed on the optimal path, $P(s)$ is a probability that the partial score s is observed all over the lattice points, and $P(k)$ is a priori probability and can be considered as a constant. Then cumulative score g by taking logarithm of eq.(4) and dropping constants can be written as eq.(5)

$$g=\sum \log\frac{P(s(i,j)|k)}{P(s(i,j))} \quad (5)$$

where $s(i,j)$ is partial score along the optimal path.

Thus, g is calculated for all the words in vocabulary and a word with the best score is taken as a recognized word. The score of $P(s/k)$ and $P(s)$ expressed in eq.(4) are statistically obtained by observing speech data. Fig.5 shows $P(s/k)$ and $P(s)$ which are derived from DP matching between input data and correct word template. Curve (a) in Fig.5 is a histogram of partial score on the optimal path and can be considered as probability density function of $P(s/k)$, and (b) is a histogram of partial scores on all lattice points and can be taken as probability density function of $P(s)$. In Fig.5 horizontal axis is for partial scores and vertical one is for frequency distribution. Integration of the histogram is normalized so that it can be considered as probability density. Fig.6 shows the relation between partial score and posterior probability. (a) is the posterior probability

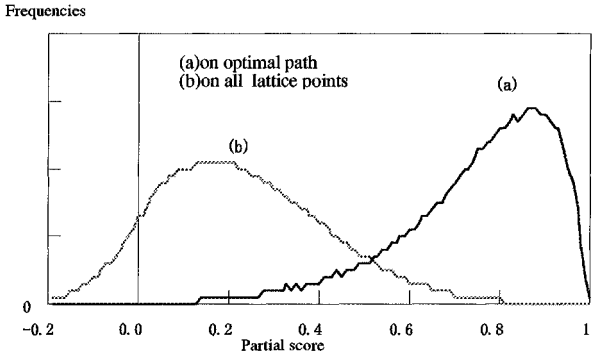

Fig. 5 Distribution of partial scores

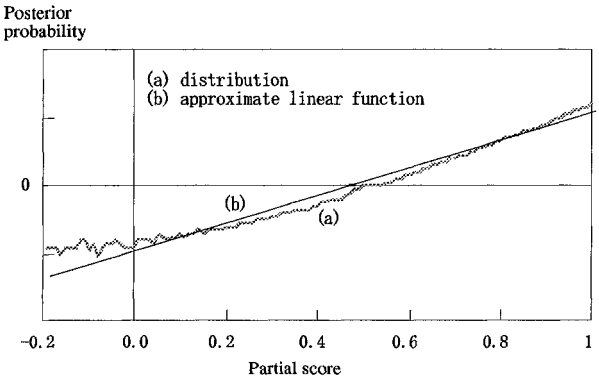

Fig.6 Relation between partial score and posterior probabiliry

of distribution derived from Fig.5. Fig.6 shows that (a) can be approximated by a linear function and eq.(5) is expressed as

$$g=\sum \beta\{s(i,j)-\alpha\} \quad (6)$$

where α and β are constants. Clearly the constant β has no significance to compare word scores. This equation suggests us that posterior probability can be obtained simply by subtracting a constant from partial scores.

4.2 Experimental Result

Recognition experiments were carried out with 100 words data (100 Japanese city names) of 25 male and 25 female speakers. To simulate recognition performance under noisy condition, noise data sampled at exhibition hall was added to speech data to be 20 dB S/N ratio. Six phoneme similarities were selected at each frame in 4 bit precision.

This experimental result is shown in Table 2. Introducing posterior probability, recognition rate under the endpoint free condition was improved from 91.7 % 95.9 %. This result verifies our method of the simple word spotting technique.

Table 2 Relation between endpoint condition and recognition rate

Endpoint condition	*Recognition rate (%)*
Fixed	97.5
Free(baseline)	91.7
Improved free	95.9

5. CONCLUSION

Word spotting technique and reduction of computation in this method were described, which are essential to use it in practical conditions. The points described in this study are:

(1)Feature parameters can be represented approximately with a few selected components of similarity vector, using partial score in DP calculated only with the few selected components yields high performance.

(2)Since correlations between phoneme similarities are used as partial score, 4 bit integer precision is sufficient to present each factor of feature parameters.

(3)Simple posteriori probabilities form were introduced, where a certain value is subtracted from correlation cosine similarity.

From the above viewpoints we can conclude that a high quality recognizer can be realized in a small hardware.

REFERENCES

[1] Hoshimi M, M Miyata, S Hiraoka and K Niyada,"Speaker-Independent Speech Recognition Method Using Training Speech from a Small Number of Speakers", ICASSP-92,ppI-469

[2] Miyata M, M Hoshimi, S Hiraoka and K Niyada,"Speaker Independent Speech Recognition Using Sub-Word Units of Model Speech Uttered by a Small Number of Speakers",IEICE Technical Report SP91-83,(Dec. 1991)

[3] Niyada K, M Hoshimi and M Miyata,"A Speaker Independent Spoken Word Recognition Method Using Phoneme Similarity Vectors which are Robust to Individual Differences",IEICE Trans. Vol. J77-A No.2 pp.135-142(Feb. 1994)

[4] Ukita T,T Nitta and S Watanabe,"Speaker Independent Connected Word Recognition Based on Unit-Word Matching",IEICE Technical Report SP83-19 (June 1983)