A Study of Noise Robustness for Speaker Independent Speech Recognition Method using Phoneme Similarity Vector

Masakatsu HOSHIMI[1)2)]*, Maki YAMADA*[1)]*, Katsuyuki NIYADA*[3)] *and Shozo MAKINO*[2)]

[1)]Matsushita Research Institute Tokyo, Inc.
3-10-1 Higashimita, Tama-ku, Kawasaki, 214-8501, JAPAN
hoshimi@mrit.mei.co.jp

[2)] Tohoku University
04 Aramaki Aza Aoba, Aoba-ku, Sendai, 980-8579, JAPAN

[3)]Central Research Labs., Matsushita Electric Industrial Co., Ltd.
3-4 Hikaridai, Seika, Soraku, Kyoto, 619-0237, JAPAN

ABSTRACT

As an input method for rapidly spreading small portable information devices, development of speaker independent speech recognition technology which can be embedded on a single DSP is now urgently requested. We have reported a speech recognition method using phoneme similarity vector as a feature vector, which is quite robust for reduction of precision of the feature parameter. We've also developed a recognition board with a single DSP, which works 100-word vocabulary using only the internal memory inside the DSP. [1][2]

In this report, we propose a new technique which makes our recognition method more robust, where a newly introduced noise standard template together with traditional phoneme standard templates for calculating phoneme similarity vector realizes precise word-spotting.

When the newly proposed noise robustness method was tested with 100 isolated word vocabulary speech of 50 subjects, recognition accuracy of 94.7% was obtained under various noisy environments.

1. INTRODUCTION

Recently portable information devices such as PDA are rapidly spreading and speech recognition is expected as a useful interfaces for such small and portable devices. To meet the expectation, a reasonable cost and high performance under actual environments are important for a speech reconizer. Regarding a cost problem, speech recognition software for PC can be one solution. But, when it comes to small portable devices which work with a single DSP, hardware resources available for speech recognition process are quite limited and it's difficult to embed traditional ASR algorithm into the small devices. Therefore, we've developed a speech recognition board which works well without extra hardware resources other than a single DSP.

To utilize the board in various applications, the recognizer must keep high performance under various noise environments and word-spotting feature in a babble noise condition. Many solutions have been proposed for noise robustness. Introduction of a distance measure based on posterior-probability enables word spotting, and garbage model is well known to represent unnecessary utterance. HMM decomposition with noise models has been studied for noise robustness as well. We also verified that word spotting is valid for a consumer electronics application in a voice activated VCR. [3]

In this paper, firstly we describe our speech recognition method which employs phoneme similarity as a feature parameter. Secondly we propose a noise robustness method for practical use and report experimental results to verify the validity of the method.

2. Features of Our Speech Recognition Method

2.1 Phoneme Similarity and Distance Measure

We've proved that phoneme similarity vector has smaller individual difference than cepstrum coefficients and also that phoneme similarity vector keeps high recognition accuracy even if it's represented in low precision. [4] Hereafter we describe our speech recognition method using phoneme similarity vector. Fig.1 shows recognition processes in our method. Speech signal

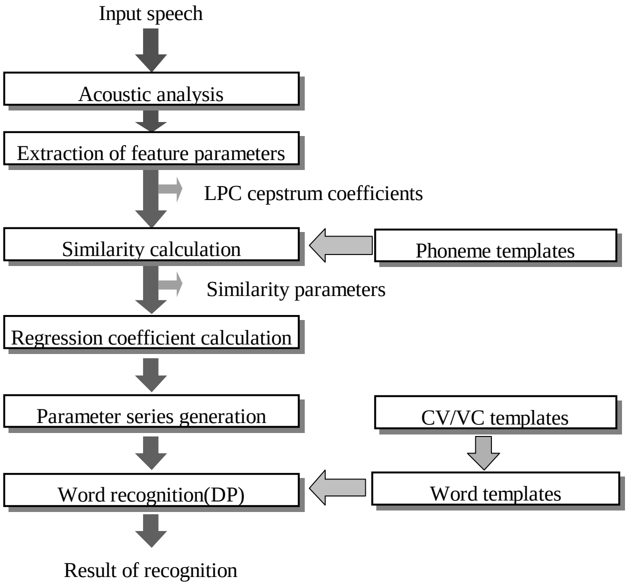

Fig.1 Outline of a speaker independent word recognition

is acquired at 12kHz sampling frequency and analyzed into 15th order of LPC cepstrum coefficients in 10msec frame interval. Then phoneme similarity based on linear discriminant function is obtained by matching the segmented coefficient pattern of test speech and phoneme standard templates, where each phoneme templates consist of cepstrum coefficients over successive 10 frames and totally 150 dimensions. The similarity for a phoneme p, Lp, is obtained as Eq. (1)

$$\mathbf{L}_p = \mathbf{a}_p \cdot \mathbf{c} - b_p \qquad (1)$$

where

$$\mathbf{a}_p = 2\Sigma^{-1} \cdot \mu_p$$

$$b_p = \mu_p \cdot \Sigma^{-1} \cdot \mu_p$$

Here c is presents segmented pattern of test speech, μ_p is a mean vector of the standard pattern of phoneme p, and Σ is a covariance matrix common for all the phoneme categories. These phoneme standard templates are trained from a lot of training subjects' data so that it can be used for speaker independent recognition. 24 phoneme similarities and their time-derivatives, which are obtained at every frame by matching test speech and phoneme standard templates, compose feature parameters for word recognition. Word models are composed by concatenating phoneme similarity pattern of CV/VC sub-word units based on Japanese syllable "KANA" representation.

Total 383 units of sub-word word patterns are trained from phoneme balanced 543 word set uttered by 16 male and 16 female subjects respectively.

The phoneme similarity vector of i-th frame, di, is presented as Eq.(2).

$$\mathbf{d}_i = (\mathbf{L}_1, \mathbf{L}_2, \cdots, \mathbf{L}_{24}) \qquad (2)$$

where Lj is similarity of phoneme j at the i-th frame.

Similarly, time derivatives of phoneme similarity vector are presented as Eq.(3).

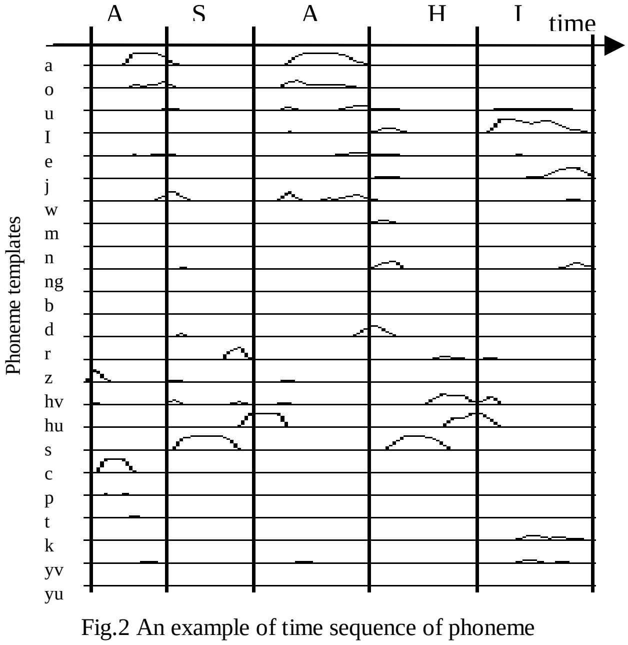

Fig.2 An example of time sequence of phoneme similarities (uttered "ASAHI")

$$\Delta\mathbf{d}_i = (\Delta\mathbf{L}_1, \Delta\mathbf{L}_2, \cdots, \Delta\mathbf{L}_{24}) \qquad (3)$$

DTW is conducted in time-alignment of word matching. Partial score s(i,j) between test speech and word model is calculated by correlation cosine distance as Eq.(4), where. $\mathbf{d}_i$ is the i-th frame similarity vector of test speech, $\Delta\mathbf{d}_i$ it's time derivative, e_j the j-th frame similarity pattern of a word model, and Δe_j it's time derivatives.

$$s(i,j) = w\frac{\mathbf{d}_i \cdot \mathbf{d}_j}{|\mathbf{d}_j|\ |\mathbf{e}_j|} + (1-w)\frac{\Delta\mathbf{d}_i \cdot \Delta\mathbf{e}_j}{|\Delta\mathbf{d}_i|\ |\Delta\mathbf{e}_j|} \qquad (4)$$

W is a weighting ratio and fixed as 0.5 in this study. Correlation cosine is an inner product of normalized similarity vector.

2.2 Algorithm improvement for Embedding in Compact Hardware Resources

Our speech recognition method is characterized by unique feature parameter of phoneme similarity vector and correlation cosine as a distance measure. Hereafter, we describe algorithm improvements in reduction of calculation processes and required memory size especially for embedding it in a compact hardware.

2.2.1 Reduction of Parameter Dimension

Phoneme standard templates are trained by segmented speech patterns which are discriminative for each phoneme. Therefore, phoneme similarity produced by matching test speech and these phoneme templates has large value in a phonetically discriminative section in test speech. For example, Fig.2 shows time sequence of normalized phoneme similarity of a sample speech "ASAHI." Similarity of phoneme /a/ has large value in the head and middle of the word utterance. While that of phoneme /o/, which is similar to /a/, also has relatively large value in the same sections, similarities of other phonemes have negligible small values. In a transitional section from phoneme /a/ to /s/, similarity of /a/ gradually decreases, and those of /s/ and similar phonemes such as /c/ and /z/ increase. Observation of phoneme similarity over the other section also shows that, in each frame, few kinds of phonemes have large values in their similarities and most of other phonemes have zero value approximately.

This suggests us that only top N phonemes with large values in each frame should be taken account in when calculating a

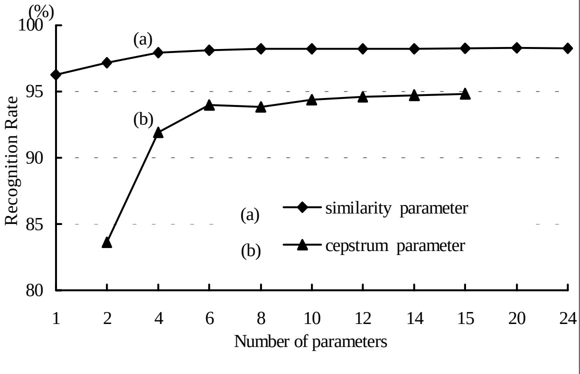

Fig.3 Recognition rate versus number of selected parameters

partial score of each frame. In other words, we can assume that score calculation corresponding to the other phonemes is not necessary.

To verify the validity of the above assumption, we conducted comparison tests where phoneme similarity vector in our method and traditional cepstrum coefficients are tested to compare their robustness in dimension reduction. Fig.3 shows experimental results where only top N feature parameters are taken account for partial score calculation. There horizontal axis means N, the number of selected elements in feature parameters, and vertical one does word accuracy rates. 212-word set uttered by 20 subjects from Tohoku Univ. and Matsushita speech databases were tested. In cepstrum coefficient case, order-base weighting by RPS was employed.

While the normalized phoneme similarity maintains high accuracy even if the number of dimension decrease to six, the cepstrum coefficient loses recognition accuracy rapidly when the dimension comes to six or less. Japanese has 24 phoneme categories and original phoneme similarity vector has 24 dimension before reduction. Therefore, if we select six phonemes for normalized similarity, we can reduce the process amount of inner product and required memory by one fourth (6/24).

2.2.2 Reduction of Bit Precision for Feature Parameter Representation

In word recognition stage, correlation cosine is employed for a distance measure. Feature parameter in the word matching is not absolute similarity value but normalized one. It means that we focus on relative relation of similarities among phonemes rather than similarity values themselves. This assumption originated from the observation of similarity values that only top N phonemes, which may affect partial score calculation, have significant values and difference among them (such as top-1 0.8, top-2 0.4, top-3 0.2), but the others have almost zero values. This suggests us that high precision may not be necessary to represent of the normalized similarity vectors.

Table1 shows results of recognition experiments where the bit precision in phoneme similarity representation is gradually decreased. Bit precision 32 means a floating-point representation and other smaller numbers do fixed-point. It shows that reduction of bit precision does not cause drastic affection on recognition accuracy. If similarities of top six phonemes are presented in four bits precision as feature parameter, memory amount required to store sub-word units (CV/VC) can be reduced by 1/32 comparing floating presentation of all the 24 phonemes' similarities. (6/24x4/32)

Table 1 Recognition rate versus precision of phoneme similarity vectors
(Top 6 phonemes are selected.)

Bit Precision	*32*	*8*	*6*	*4*
Recognition Rate(%)	97.5	97.5	97.5	97.3

Total memory amount for our recognizer is estimated as follows. In each frame of a sub-word unit, eight bits are necessary to store similarity of one phoneme. Four bits out of the eight bits are consumed to store information about which phoneme is selected and the other four bits to represent the similarity value. Considering that similarities of top six phonemes and six time-derivatives of similarities per a frame are stored as feature parameter and that sub-word units have 1923 frames in total, reduced memory size for sub-word template storage is only 23 k bytes. Considering that 370 k bytes are necessary without reduction, the memory reduction is quite efficient. In addition to 7 k bytes for phoneme standard templates, memory requirement for our recognizer come to be 30 k byte in total.

Together with the reasonable memory requirement, reduction in computation process enables us to embed the recognition algorithm in a low-cost fixed-point DSP for vocabulary size of 100 words.

3. Improvement for Noise Robustness

3.1 Word Spotting Method

In practical usage of a speech recognizer, high performance under various noise conditions is strongly desired. Furthermore unknown endpoint condition is essential in a practical noisy environment. Therefore, we developed a unique word-spotting method to improve noise robustness in our method.

In noise section proceeding/following speech, phoneme similarities before normalization should have small values and it can be an important cue to discriminates speech/non-speech section. However, in our recognition method, similarity vector is normalized when used as feature parameter. It means that normalized similarities in noise sections have some amplitude similar to that in a speech section. Due to the amplified phoneme similarity, miss-spotting between word templates and noise section decreases accuracy of word-spotting functionality.

To solve such a problem in word spotting, we propose to add a noise standard template to 24 phoneme standard templates and introduce similarity toward the noise templates as one of elements in phoneme similarity vector. Then Eq.(2) and (3) are updated to (5) and (6) after introduction of the noise similarity Ln.

$$\mathbf{d}_{\mathrm{i}} = (\mathbf{L}_1, \mathbf{L}_2, \cdots, \mathbf{L}_{24}, \mathbf{L}_{\mathrm{n}}) \qquad (5)$$

$$\Delta\mathbf{d}_{\mathrm{i}} = (\Delta\mathbf{L}_1, \Delta\mathbf{L}_2, \cdots, \Delta\mathbf{L}_{24}, \Delta\mathbf{L}_{\mathrm{n}}) \qquad (6)$$

Dominant similarity toward the noise standard template decreases matching scores between word templates and a noise section in a test sample, and then reduces recognition errors which were caused by low accuracy of word-spotting functionality without the noise standard template.

The noise standard template should be trained in a specific noise sound data if the noise environment can be exactly specified. However, if it can't be specified, a mixture density trained by plural sorts of typical noise sounds should be a substitute of the specific noise When a mixture density is employed for the noise similarity, the maximum similarity over the all noise categories represents the noise similarity Ln as Eq.(7) shows.

$$\mathbf{L}_{\mathrm{n}} = \max_{\mathrm{i}} \left\{ \mathbf{L}_{\mathrm{n_i}} \right\} \qquad (7)$$

3.2 Experimental Results

Noisy test speech data was simulated by merging noise sound to clean speech data. The clean speech data (100 Japanese city name) was sampled in a sound proof room, and three kinds of noise, exhibition show site, car, office environment, were added.

Firstly basic validity of introduction of the noise standard template was tested. With a test speech data corrupted by exhibition show site noise, we tested recognition accuracy by following three conditions.

1)Endpoint Fixed (No Noise Standard Template)
2)Endpoint Free (No Noise Standard Template)
3)Endpoint Free (With Noise Standard Template trained by exhibition show site noise)

Table2 shows the test result. Recognition rate in endpoint free condition without a noise standard template decreased 2.2 % from endpoint fixed condition. On the other hand, decrease with a noise standard template was only 1.6 %. That means that introduction of a noise standard reduced error rate by 27% and validity of proposed method for noise robustness was proved.

Table 2 Relation between endpoint condition and recognition rate

Endpoint Condition	*Noise standard template*	*Recognition rate*
Fixed	NA	97.5 (%)
Free	NA	95.3
Free	Exhibition show site Noise	95.9

Secondary proposed method was evaluated in various noise conditions. Three kinds of noise, exhibition show site, car, office environment, were used for testing and training a noise standard template. Four noise standard templates were trained. Three of them were trained by a specific noise sound, and the other one was trained by all the three noise to be a mixture density form.

Training conditions of noise standard templates are shown as follows.

a) No Noise Template (Baseline)
b) Exhibition Show Site Noise
c) Car Noise
d) Office Environment Noise
e) Mixture Density with above three noises

Fig.4 shows evaluation results. Horizontal axis shows noise category of test speech data, and vertical one does recognition rate. Each line consists of recognition results from the same noise standard template.

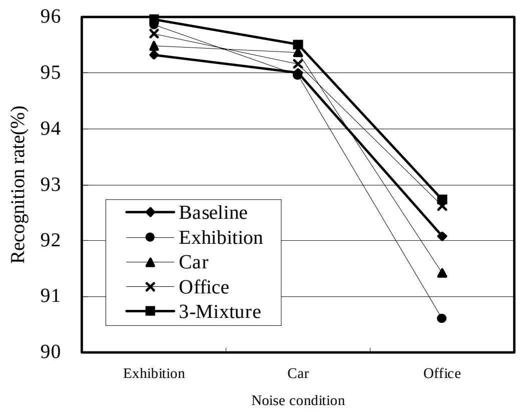

Fig.4 Relation between noise condition and recognition rate

When a noise template was trained in a specific noise sound, improvement was obtained when noise conditions are identical in training and testing (b, c, d). However, when training and testing conditions are different, accuracy is decreased in some cases.

To achieve robustness in unspecific noise conditions, the mixture density noise template was also tested, where the template was trained by three kinds of noise, exhibition show site, car, and office environment. As the case (e) in Fig.4 shows, the mixture noise template keeps high performance over any kinds of testing conditions.

According to these experimental results, we can conclude that a mixture noise standard template trained in typical plural noise conditions is valid to obtain a high performance in various noise conditions.

Furthermore, increase of computation process for the proposed method is quite small since we only need to acquire noise similarity by matching test speech with a noise standard template. In comparison with total amount of process and memory for the recognizer, increase of process and memory extra for the proposed method is quite trivial. Consequently the recognition algorithm with the proposed noise robustness method is practical in the sense that it can be implemented in a conventional single DSP.

4. Conclusion

This paper described outline of our method which works on a single DSP for 100 word recognition. Then improvement method for noise robustness was proposed. With evaluation experiments, we got following conclusions:

1)In addition to conventional 24 phoneme standard templates, introduction of a noise standard template was proved to achieve high accuracy under endpoint free conditions.

2)Mixture density of a noise template trained in typical noise conditions is valid to achieve robustness in various test conditions.

3)Since the proposed method requires little increase in computation process and memory, it's possible to embed the improved recognition method in a conventional single DSP for 100-word vocabulary size.

In a future work, we evaluate the method under farther various noise conditions and improve the algorithm for better performance.

5. REFERENCES

[1]Hoshimi M, M Yamada and K Niyada,"A Practical Speech Recognition Method for Unspecified Speakers on a Single DSP Chip" IEICE Trans. Vol.J79-D-II No.12 pp.2096-2103(Dec. 1996)

[2]Hoshimi M, M Yamada and K Niyada,"Speaker Independent Speech Recognition Method Using Phoneme Similarity Vector",ICSLP94,S31-21

[3]Hiraoka S, K Niyada T Kimura,"A Small Vocaburary Speech Recognizer for Unspecified Speaker Using Word-Spotting Technique"IEICE Technical Report SP88-18(June 1988)

[4]Niyada K, M Hoshimi and M Yamada," 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)