Component reduction technique for covariance matrix of multidimensional Gaussian distribution in Speech Recognition

Eiichi Seyoshi, Kazumasa Yamamoto and Seiichi Nakagawa
Toyohashi University of Technology, Toyohashi 441-8580 Aichi Japan
E-mail: sueyoshi@slp.ics.tut.ac.jp, kyama@slp.ics.tut.ac.jp, nakagawa@slp.ics.tut.ac.jp Tel: 0532-44-6777

***Abstract*—Recently, speech recognition systems are in practical use as an input device of a car navigation system, etc. However, since the computational and storage resources are usually limited for a car navigation system, these systems are required to achieve high recognition performance with the limited resource. In this paper, we propose a method that reduces the components of full covariance matrix considering only the dominant correlation components between static and dynamic feature parameters. The recognition experiment was performed by using the conventional and the proposed method, and both were compared. From the continuous syllable recognition result, we confirmed the effectiveness of proposed component reduction technique considering the correlation between parameters.**

I. Introduction

Hidden Markov model (HMM) has been used as acoustic models of speech recognition systems. An output probability in each state of HMM is represented by a mixture of multidimensional Gaussian distributions with mean vectors and covariance matrices of speech features. Recently, the increase of speech corpus size makes precise training of the covariance matrix be possible. Therefore, instead of diagonal covariance matrices, full covariance matrices are sometimes used[1]. However, the use of all components of the full covariance matrix is difficult due to the computational complexity, thus, a block-type full covariance matrix is usually used which considers only correlation within static and its dynamic (Δ and $\Delta\Delta$) feature parameters, respectively [2][3]. In the field of handwritten character recognition, the computational complexity has been reduced without degrading the recognition performance by relocating only the components to the block with strong correlation of the covariance matrix [4]. In the field of speech recognition, Schuster et al. reported that a full covariance matrix can be approximated by a constrained covariance matrix which has only small number of effective parameters by using MPPCA (Mixture of Probabilistic Principal Component Analysis) [1]. From these reasons, we think that the block type component reduction method which completely disregards the correlation between static and dynamic parameters is not preferable. Moreover, Kusama et al. reported the result of recognition experiments with changing the number of mixtures of the block-type full covariance matrix and diagonal covariance matrix. They concluded that the recognition accuracy became almost the same level when the number of parameters was almost the same.

We evaluated the computational complexity reduction without degrading the recognition performance with reduced components of covariance matrix. We also evaluated the effectiveness of the correlation among static and its dynamic (Δ and $\Delta\Delta$; velocity and acceleration in time sequences) parameters. Then, the recognition performance for the combination of various component was compared. The block-type full matrix that considered only the correlation within static, Δ and $\Delta\Delta$ parameters respectively, and a diagonal covariance matrix were used as traditional techniques. The technique for reduction by using only the components with a strong correlation was used as a proposal method. The above mentioned technique was compared in the point of the recognition rate of large vocabulary continuous speech recognition and continuous syllable recognition.

II. Component reduction technique for covariance matrix

A. Fundamental distribution model

The output probability in each state of HMM is represented by a mixture of multidimensional Gaussian distributions with mean vectors, μ, and covariance matrices of speech features, Σ, as shown in Eq.(1);

$$p(\mathbf{o}) = \sum_{w=1}^{W} \lambda_w \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \times \exp\{-\frac{1}{2}(\mathbf{o}-\mu)^T\Sigma^{-1}(\mathbf{o}-\mu)\} \quad (1)$$

where W is the number of mixtures, λ_w is a mixture weight coefficient of the w-th mixture component, and D shows the number of dimension of the feature vector. Conventionally, training data were few, so that it was difficult to train enough the covariance matrix. Thus the diagonal covariance matrix like Fig.1 has been used. Fig.2 shows the block-type full covariance matrix. However, such a component reduction method disregarding the correlation between parameters has the less possibility that the effective elements are used well for speech recognition.

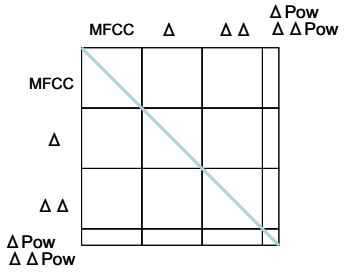

Fig. 1. diagonal covariance matrix

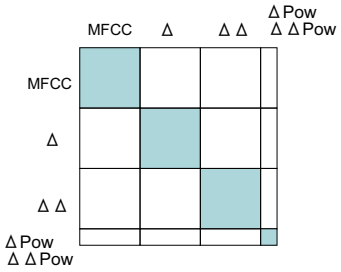

Fig. 2. block full covariance matrix

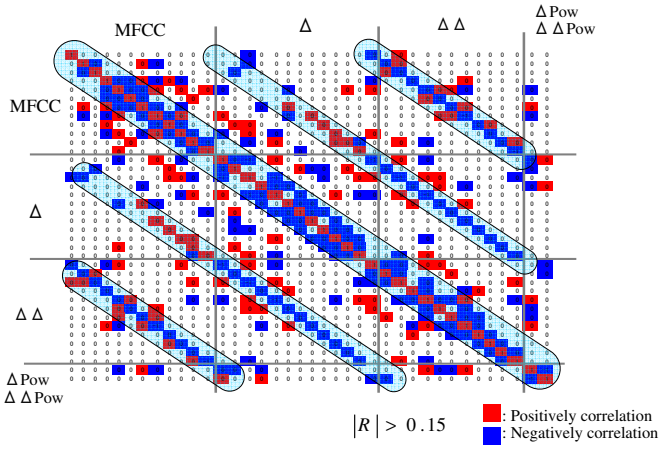

Fig. 3. Actual covariance matrix

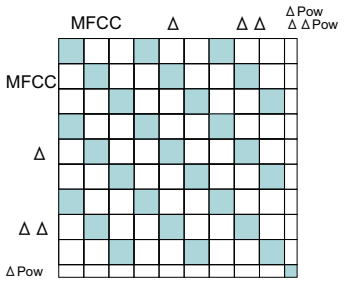

Fig. 4. pattern A

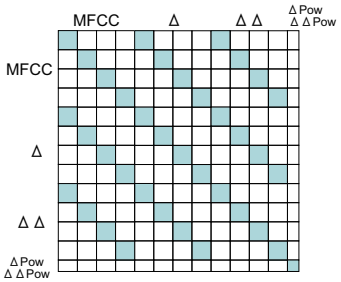

Fig. 5. pattern B

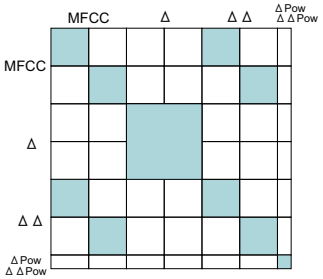

Fig. 6. pattern C

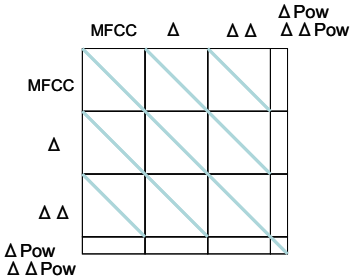

Fig. 7. pattern D

Therefore, we propose the technique for reducing the component for speech recognition to use only the effective components among full covariance matrix components considering the correlation between parameters. And, we try to reduce the computational complexity without degrading the recognition rate.

B. Positive-Definiteness of covariance matrix

The covariance matrix is symmetrical, and it has a positive-definiteness. A positive-definiteness is a property that the determinant is always positive. If a positive definiteness of the covariance matrix is not kept by component reduction, the determinant becomes negative, and the output probability of HMM is not computable. Therefore, It is necessary to note that a positive definiteness of the covariance matrix must be kept by the component reduction.

C. Block type component reduction

Covariance matrix components are reduced to become a block-type matrix, so that the covariance matrix guarantees a positive definiteness. The block full covariance is made by gathering the component with a strong correlation together in a block, and substituting the element with a weak correlation to 0.

The computational complexity of full covariance matrix can be decreased by reducing the component in the block. Moreover, when the training data has the same amount, it can be expected that the parameter's precision of the covariance matrix rises, because the ratio of the number of training data to the number of dimensions will increase, compared with the case to use a full covariance matrix.

D. Component reduction pattern

To reduce the component of a full covariance matrix effectively, the correlation between components of an actual covariance matrix was investigated. The correlation of the component is calculated for the distributions of covariance matrices of all over HMMs. and Fig.3 shows the average correlation calculated for all syllable models. If the absolute value of the correlation is higher than 0.15, it is highlighted. We can consider that covariance matrix tends to retain the high correlation in its diagonal component between static and dynamic parameters such as, MFCC and Δ Δ MFCC. Then, the highlighted boxes are around those components in Fig.3. Thus, we propose the component reduction by four patterns such as Fig.4, Fig.5, Fig.6, and Fig.7 based on the actual covariance matrix.

The number of components of a block-type full covariance matrix is MFCC(12×12) + Δ MFCC(12×12) + Δ Δ MFCC(12×12) + Δ log pow and Δ Δ log pow(2×2) per mixture. It has 436 components in total. However, the actual number of components is 238 because the covariance matrix is symmetrical. Comparing with a block full covariance matrix, pattern A shown in Fig.4 consists of a small block of 4×4, and the number of parameters per distribution (mean + covariance matrix) becomes 276. Pattern B chooses the component of the small block of 3×3, which is similar to the pattern A, and is comparable with the number of components and the recognition performance of pattern A. As we can see in Fig.3 again in detail, it has been found that there is no element with the high correlation between Δ and MFCC or Δ and ΔΔ.

TABLE I
NUMBER OF ELEMENTS OF MEAN AND COVARIANCE MATRICES PER DISTRIBUTION OF EACH ELEMENT RESTRICTION PATTERN

Mix.	diag.	block	pattern			
			A	B	C	D
1	76	276	276	222	276	112
2	152	552	552	444	552	224
4	304	1104	1104	888	1104	448
8	608	2208	2208	1776	2208	896
16	1216	-	-	-	-	1792
32	2432	-	-	-	-	-

Based on this fact, pattern C reduces the component considering only the correlation of MFCC and ΔΔ. The number of mixtures of each pattern and the number of parameters per distribution are shown in Table I. The number of mixtures used in this paper were 1,2,4,8,16,32 mixtures for diagonal pattern, 1,2,4,8 mixtures for patterns A and B in consideration of the number of components and the amount of the training data. The recognition experiment evaluates the relation between the number of components and the recognition accuracy by the difference of the component reduction patterns. Furthermore, the improvement of the recognition accuracy by considering the correlation between different parameters is evaluated.

III. RECOGNITION EXPERIMENT

A. Experimental setup

The data used in this experiment was Japanese Newspaper Article Sentences (JNAS) corpus.

The speech analysis conditions were sampling frequency of 16kHz, frame length of 25ms, frame shift of 10ms and filterbank channels of 24. The dimension of feature vector was 38 consisting of 12 MFCCs, their Δ and ΔΔ, Δpower and ΔΔpower.

The training data were 12703 sentences in total that 125 male speakers had uttered respectively 100 sentences. The test data were 100 sentences selected from the JNAS corpus, which were uttered by 10 male speakers (these speakers are independent of the training data). The acoustic model was 4 state left-to-right, context independent 116 syllable HMMs and left-context dependent 928 syllable HMMs.

B. Recognition experiment using context independent HMM

The result of the continuous syllable recognition experiment using HTK that uses context independent 116 syllable HMMs is shown in Table II, Fig.8, and Fig.9. The measure for evaluation is correct rate and accuracy of syllables. The accuracy of syllables was 62.1% with 4 mixtures having block-type full covariance and 62.8% with 16 mixtures having diagonal covariance (conventional method), respectively. In contrast, The recognition accuracy of the proposed method were 65.8%, 64.3%, 65.0% with 4 mixtures having the covariance of patterns A, B, and C, respectively, and 63.5% with 8 mixtures for pattern D. The number of components per mixture of patterns A and C was equal to the block type, and the number of components of patterns B and D is less than the block

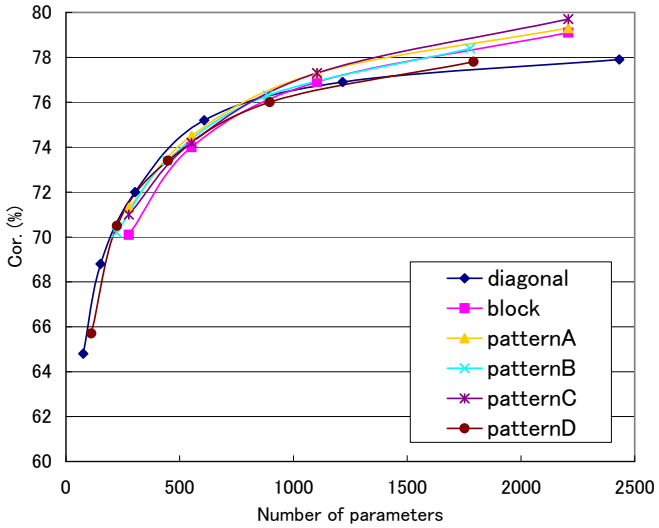

Fig. 8. Result of continuous syllable recognition using context independent HMM -Cor.(%)-

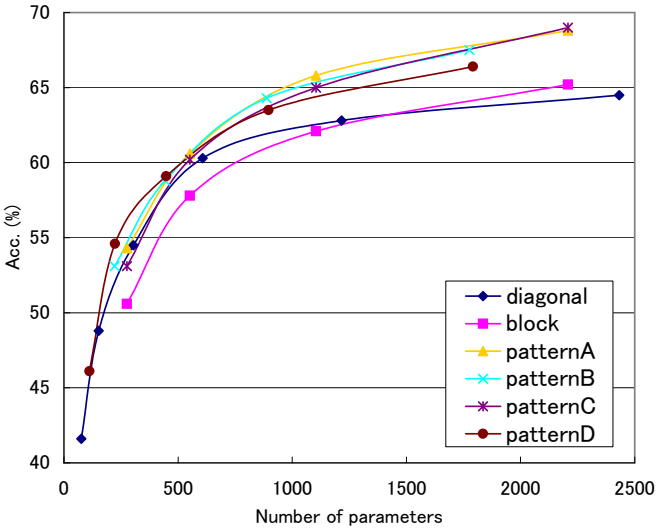

Fig. 9. Result of continuous syllable recognition using context independent HMM -Acc.(%)-

type. Therefore, we considered that these component reduction patterns could use the dominant correlation components for speech recognition better than the conventional methods. As a result, the possibility of improving acoustic model's recognition performance by using the correlation between static and dynamic feature parameters was confirmed. It is expected that to use the correlation between static and dynamic feature parameters can improves LVCSR pefermance.

C. Recognition experiment using context dependent HMM

The component reduction technique is effective for acoustic model's performance improvement from the result of preceding section. For this result, the component reduction technique was evaluated for context dependent HMMs. The number of mixtures of HMM was set to 16 mixtures for diagonal, and 4 mixtures for block type and patterns A and B, so that the number of component of acoustic models become almost the

TABLE II
RESULT OF CONTINUOUS SYLLABLE RECOGNITION USING CONTEXT INDEPENDENT HMM [%]

Mix.	diagonal		block		pattern A		pattern B		pattern C		pattern D	
	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.
1	64.8	41.6	70.1	50.6	71.3	54.3	70.2	53.1	71.0	53.1	65.7	46.1
2	68.8	48.8	74.0	57.8	74.5	60.6	73.4	58.8	74.2	60.2	70.5	54.6
4	72.0	54.5	76.9	62.1	77.3	65.8	76.3	64.3	77.3	65.0	73.4	59.1
8	75.2	60.3	79.1	65.2	79.3	68.8	78.4	67.5	79.7	69.0	76.0	63.5
16	76.9	62.8	-	-	-	-	-	-	-	-	77.8	66.4
32	77.9	64.5	-	-	-	-	-	-	-	-	-	-

TABLE III
RESULT OF CONTINUOUS SYLLABLE RECOGNITION USING CONTEXT DEPENDENT HMM [%]

	diagonal	block	pattern			PCA
			A	B	C	
Mix.	16	4	4	4	4	4
cor.	83.8	83.3	84.7	83.7	85.1	83.5
acc.	70.3	69.9	73.8	73.0	74.4	71.4

TABLE IV
RESULT OF CONTINUOUS WORD RECOGNITION USING CONTEXT DEPENDENT HMM [%](1PASS TRIGRAM)

	diagonal	block	pattern			PCA
			A	B	C	
Mix.	16	4	4	4	4	4
cor.	93.7	94.3	94.5	93.6	94.6	93.0
acc.	92.3	92.2	93.1	91.8	92.9	91.4

same for component reduction patterns.

In addition, we compared the proposed technique with Principal Component Analysis (PCA) that was a statistical technique to reduce the number of components. We compressed MFCC from 36 dimensions (static, Δ, $\Delta\Delta$) into 20 dimensions by PCA (Cumulative Proportion: 0.983), and Δpow and $\Delta\Delta$pow were appended. The number of parameters per Gaussian distribution is 235 and the number of elements is almost equal to the block type.

A trigram model was used as a language model trained with the Mainichi Newspaper articles, and SPOJUS was used as a recognition decoder developed in our laboratory[6].

Continuous syllable recognition and LVCSR (20,000 words) experiments by SPOJUS were performed by using the context dependent HMMs. The recognition result is shown in Tables III and IV. The measure for evaluation is correct rate and accuracy of syllables.

For the continuous syllable recognition, results of the accuracy of syllables were 70.3% with diagonal, 69.9% with block type, 73.8%, 73.0% and 74.4% with patterns A, B and C, respectively.

The recognition performance using patterns A and B considering the correlation between parameters was improved from the conventional method. This result has similar tendency to the case with context independent HMMs in the previous section. Especially, the accuracy of syllable was improved 4.5% in block type and pattern C, to which the number of components is equal. Therefore, the proposed method was effective for speech recognition. In the result of the LVCSR, there was small improvement of recognition performance on the respective reduction patterns.

IV. CONCLUSIONS

To reduce the computational complexity without degrading the recognition rate or to improve the recognition rate, we proposed the technique for reduction the components by using only the effective component of a full covariance matrix considering the correlation between parameters. The recognition experiment was performed by using the conventional and the proposed methods, and both were compared.

From the continuous syllable recognition result with context independent HMMs and context dependent HMMs, the recognition performance of proposed method with patterns A and B was improved in comparison with the block-type full covariance matrix and the diagonal covariance matrix. Therefore, we confirmed the effectiveness of proposed component reduction technique considering the correlation between parameters. In the LVCSR result, there was small improvement of the consideration of the correlation between parameters.

In this study, we proposed the component reduction method by four patterns based on the average of correlation between components of all syllables. As future work, we will try the further evaluation of the component reduction method by another reduction pattern, for instance, that use only the components with high correlation at every syllable.

REFERENCES

[1] M. Schuster, T. Hori and A. Nakamura, “Experiments with Probabilistic Principal Component Analysis in LVCSR”, *Proc. Interspeech’05 (Eurospeech)*, p.1685-1688, 2005.

[2] S. Nakagawa , Y. Hirata , Y. Hashimoto, “Japanese phoneme recognition using continuous parameter hidden Markov models”, *Jour. Acoustical Society of Japan*, Vol.46, no.6, pp.486-496, 1990.6 (in Japanese).

[3] K. Hanai, K. Yamamoto, N. Minematsu and S. Nakagawa, “Continuous speech recognition using segmental unit input HMMs with a mixture of probability density functions and context dependency”, *Proc. ICSLP’98* ,pp.2935-2938, 1998.

[4] F. Sun, S. Omachi, N. Kato and H. Aso, “Fast and Precise Discriminant Function Considering Correlations of Elements of Feature Vectors and Its Application to Character Recognition” ,*Systems and Computers in Japan*, vol.30, no.14, pp.33-42, December 1999.

[5] T. Kusama, Y. Okuyama, M. Katoh, T. Kosaka, M. Kohda, “Improvement of Unsupervised Adaptation in Lecture Speech Recognition” ,*IEICE technical report. Speech* , 107, pp.73-78, 2007 (in Japanese).

[6] J. Zhang, L. Wang, S. Nakagawa, “LVCSR based on context dependent syllable acoustic models”, *Asian Workshop on Speech Science and Technology*, SP2007-200, pp.81-86, 2008.3.