IMPROVEMENTS ON MINIMUM COVARIANCE BASED SPATIAL CORRELATION TRANSFORMATION

Tengrong Su, Ji Wu, Zuoying Wang

Department of Electronic Engineering
Tsinghua University, Haidian District,
Beijing, 100084, P.R.China
str03@mails.tsinghua.edu.cn

Jie Hao

Toshiba (China) CO. LTD.
Tower W2, Oriental Plaza, Dong Cheng District,
Beiijing, 100738, P.R. China
haojie@rdc.toshiba.com.cn

ABSTRACT

In order to take advantage of the correlation information among different acoustic units in speech recognition, a novel approach named Minimum Covariance based Spatial Correlation Transformation was proposed in [8], which achieves satisfactory performance. However, there are two issues of this approach which can still be improved, 1) the estimation of the transformation matrix; 2) the construction of the history data. In this paper, a new algorithm of estimating the transformation matrix and a new strategy of constructing history supervector are proposed. Experimental results show that the improved approach achieves better performance than the original one.

Index Terms—Speech recognition, spatial correlation, feature transformation, history data

1. INTRODUCTION

The Hidden Markov Model (HMM) has been successfully applied in the area of speech recognition. However, one of its key assumptions named "frame-independence" ignores the correlation existing in real speech [1]. Since both the vocal organ of human being and the pronunciation rules of languages are almost fixed, for a specific speaker, strong correlation exists among different acoustic units such as phones and, what's more, the correlation might be stable. The correlation among acoustic units can be described by the correlation among acoustic model parameters in the feature space, so we call it Spatial Correlation.

In literature, the correlation among different models has been used in some model adaptation approaches. In Maximum Likelihood Linear Regression (MLLR) [2], different Gaussian components are tied with each others by a regression tree, to share the same transformation matrix. The MLLR approach yields good performance on significant amounts of adaptation data.

Reference Speaker Weighting (RSW) [3] focuses on the correlation among different speakers. In this approach, each speaker is represented by a supervector, which is constructed from the speaker-dependent (SD) model parameters for the speaker. And a new speaker is considered to be a weighted combination of a set of training speakers (reference speakers). Eigenvoice [4] improved the idea of RSW. It applies principal component analysis (PCA) to either the covariance or the correlation matrix calculated from the reference speakers, to find a set of eigenvectors (eigenvoices). Then the new speaker is represented by a linear combination of the eigenvoices. When there is a small amount of adaptation data, the Eigenvoice approach significantly outperforms the MLLR approach.

Based on quantitative analysis on the spatial correlation among different acoustic units [5], Yu proposed a training algorithm named "Spatial Constrained Training (SCT)" [6], which applies a set of Spatial Constraints to the traditional K-Mean Segmental algorithm, and a new adaptation algorithm named "Spatial Correlated Maximum a Posteriori Adaptation (SC-MAP)" [7], which applies Spatial Correlation Assumption to the traditional Maximum a Posteriori criteria. Both approaches achieve a good performance.

All the previous approaches focus on the acoustic model training or the model adaptation. Our approach, named Minimum Covariance based Spatial Correlation Transformation (MC-SCT) [8], instead applies the spatial correlation information in the decoding process. Based on minimum covariance criteria, a transformation matrix is determined to find new acoustic features and the corresponding models which can achieve better discriminative performance. Though the original algorithm of this approach achieves competitive performance, two issues of the approach can still be improved, 1) the estimation of the transformation matrix; 2) the construction of the history superverctor. In this paper, a new algorithm for estimating the transformation matrix is proposed, in which the spatial correlation information among history data is utilized in estimating the covariance matrices of the new features. Furthermore, a new strategy for constructing the

history supervector is applied to the approach, to reduce the influence of incorrect state labels.

This paper is organized as follows. In section 2, we review the basic idea and the original algorithm of MC-SCT. In section 3, the improvements on the approach are introduced. In section 4, we discuss the combination of the adaptation approaches and MC-SCT. In section 5, the experiment results are presented. Finally, we summarize this paper and outline our future work.

2. BASIC IDEA AND ORIGINAL ALGORITHM OF MC-SCT

Let's assume that the recognition system has got a set of observed frame vectors with state labels, $x_1, x_2, \cdots, x_n$, called history data, and the current frame vector y. After mean normalization, we can assume that all the frames are Gaussian random vectors with zero means. And we assume that they have a joint Gaussian distribution. Let supervector $x = (x_1^T, x_2^T, \cdots, x_n^T)^T$ represent all the history data. And use x and y to construct a new feature vector

$$z = y - Wx \tag{1}$$

where W is the transformation matrix. Obviously, the new vector z is also a Gaussian vector with zero mean. And the covariance matrix of z is expressed as:

$$R_z = E(zz^T) = E[(y - Wx)(y - Wx)^T] \tag{2}$$

According to the minimum covariance criteria, the transformation matrix W is optimized to minimize the covariance of vector z, in order that the new feature will have better discriminative performance than the original feature. The optimum transformation matrix can be expressed as:

$$W = E[yx^T]E[xx^T]^{-1} = R_{yx}R_x^{-1} \tag{3}$$

So the corresponding vector and its covariance can be expressed as:

$$z = y - R_{yx}R_x^{-1}x \tag{4}$$

$$R_z = R_y - R_{yx}R_x^{-1}R_{xy} \tag{5}$$

If we take frame vector x_i as a sample of its corresponding state's observation distribution, we can use a set of SD models trained previously to estimate the correlation matrices R_x and R_{yx}. For each speaker, a supervector is constructed from his SD acoustic model parameters according to the state label sequence of x. The supervector $U^{(p)}$ for speaker p is defined as:

$$U^{(p)} = \begin{pmatrix} c_{s_1}^{(p)} \\ \vdots \\ c_{s_n}^{(p)} \end{pmatrix} \tag{6}$$

where $c_{s_i}^{(p)} = \mu_{s_i}^{(p)} - \mu_{s_i}$, $i = 1, \cdots, n$, with s_i denoting the state of frame vector x_i, and $\mu_{s_i}^{(p)}$, μ_{s_i} denoting the mean vectors of state s_i of speaker p and the SI model separately. Let the number of speaker be defined as P. And define a parameter matrix U_{s_i} for state s_i, which is given as:

$$U_{s_i} = [c_{s_i}^{(1)}, c_{s_i}^{(2)}, \cdots, c_{s_i}^{(P)}] \tag{7}$$

Then the autocorrelation matrix R_x and the correlation matrix R_{yx} can be expressed as:

$$R_x = \frac{1}{P}\sum_{p=1}^{P} U^{(p)}U^{(p)T} = \frac{1}{P}UU^T \tag{8}$$

$$R_{yx} = \frac{1}{P}U_{s_y}U^T \tag{9}$$

where s_y denotes the state of y, and

$$U = [U^{(1)}, U^{(2)}, \cdots, U^{(P)}] = \begin{pmatrix} U_{s_1} \\ \vdots \\ U_{s_n} \end{pmatrix} \tag{10}$$

The final expression of the new vector and its covariance can be given as follows:

$$z = y - U_{s_y}(\sum_{i=1}^{n} U_{s_i}^T U_{s_i})^{-1}(\sum_{i=1}^{n} U_{s_i}^T x_i) \tag{11}$$

$$R_z = R_y - \frac{1}{P}U_{s_y}U_{s_y}^T \tag{12}$$

The detail of the derivation can be found in [8].

3. IMPROVEMENTS ON MC-SCT

3.1. A new algorithm for estimating the transformation matrix

Since the number of speakers is much less than the dimension of the supervector x, the autocorrelation matrix estimated in Equation (8) is always rank-deficient, that is, non-invertible. In [8] the Moore-Penrose inverse of this matrix is adopted to substitute its inverse matrix, which causes the result that the covariance of the vector z is not related to the history data, as shown in Equation (12). In other words, the spatial correlation information among the history data is not used in estimating the new covariance. To solve the problem, we propose a new algorithm to estimate the transformation matrix.

According to Equation (8) and (10), the autocorrelation matrix can be reformulated as:

$$R_x = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1n} \\ R_{21} & R_{22} & \cdots & R_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ R_{n1} & R_{n2} & \cdots & R_{nn} \end{bmatrix} \quad (13)$$

where

$$R_{ij} = \frac{1}{P} U_{s_i} U_{s_j}^T, 1 \le i, j \le n \quad (14)$$

R_{ij} represents the correlation between frame x_i and frame x_j. Obviously, the autocorrelation of frame x_i is represented by R_{ii}, the covariance matrix of the SD model mean vectors of state s_i.

To ensure the autocorrelation matrix R_x being full-rank, we substitute R_{ii} with the covariance matrix $\overline{R}_{s_i}$ of state s_i in the SI model, which enhances the pivot elements of R_x. Then Equation (13) can be rewritten as:

$$R_x = \begin{bmatrix} \overline{R}_{s_1} & R_{12} & \cdots & R_{1n} \\ R_{21} & \overline{R}_{s_2} & \cdots & R_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ R_{n1} & R_{n2} & \cdots & \overline{R}_{s_n} \end{bmatrix} = R_I + \frac{1}{P} U U^T \quad (15)$$

where

$$R_I = diag(\overline{R}_{s_1} - R_{11}, \overline{R}_{s_2} - R_{22}, \cdots, \overline{R}_{s_n} - R_{nn}) \quad (16)$$

According to Woodbury Formula [9], we get

$$R_x^{-1} = R_I^{-1} - R_I^{-1} \cdot \frac{1}{P} U (I + \frac{1}{P} U^T R_I^{-1} U)^{-1} U^T R_I^{-1} \quad (17)$$

Using Equation (17), Equation (4) and (5) can finally be rewritten as:

$$z = y - \frac{1}{P} U_{s_y} A_n^{-1} b_n \quad (18)$$

$$R_z = R_y - \frac{1}{P} U_{s_y} (I - A_n^{-1}) U_{s_y}^T \quad (19)$$

where

$$\begin{aligned} A_n &= I + \frac{1}{P} U^T R_I^{-1} U \\ &= I + \sum_{i=1}^{n} \frac{1}{P} U_{s_i}^T (\overline{R}_{s_i} - R_{ii})^{-1} U_{s_i} \end{aligned} \quad (20)$$

$$b_n = U^T R_I^{-1} x = \sum_{i=1}^{n} U_{s_i}^T (\overline{R}_{s_i} - R_{ii})^{-1} x_i \quad (21)$$

Both of them can be accumulated iteratively.

To reduce the dimension of the matrix in the inverse calculation, according to Woodbury Formula, we get:

$$\begin{aligned} A_n^{-1} = A_{n-1}^{-1} - \frac{1}{P} A_{n-1}^{-1} U_{s_n}^T \\ \cdot [(\overline{R}_{s_n} - R_{nn}) + \frac{1}{P} U_{s_n} A_{n-1}^{-1} U_{s_n}^T]^{-1} U_{s_n} A_{n-1}^{-1} \end{aligned} \quad (22)$$

Now the covariance of the new vector is related to the history data, as shown in Equation (19). Then the spatial correlation information among the history data can be used in estimating the new covariance.

3.2. A new strategy for constructing the history super-vector

In the original MC-SCT, the history supervector x is constructed by concatenating all the history frame vectors, and the supervectors used to estimate its autocorrelation matrix R_x are constructed from the SD model parameters according to the state sequence of the history data. It means that the correlation between two frames is represented by the correlation between their corresponding states' parameters, as shown in Equation (14). When MC-SCT is applied in the unsupervised mode, since the state labels may be not as precise as we expect, the incorrect state labels may influence the transformation matrix in an incorrect direction.

To tackle the above problem, a new strategy for organizing history data is considered here. The history supervector is not constructed by concatenating all the history frames, but the sample mean vectors in the history data for the states appearing in the state sequence. Then the new history supervector can be expressed as:

$$\overline{x} = (\overline{x}_1^T, \overline{x}_2^T, \cdots, \overline{x}_M^T)^T \quad (23)$$

where $\overline{x}_s$ denotes the sample mean for state s, while M denotes the total number of states appearing in the state sequence. The influence of incorrect state labels is reduced by the sample mean vectors here.

Then the previous algorithms of estimating the transformation matrix can be applied to the new history supervector $\overline{x}$. In the batch mode, the iterations in Equation (20) and (21) should be carried out according to the states appearing in $\overline{x}$, and the total iteration number is M. In the on-line mode, A_n should only be accumulated whenever a new state appears, but b_n should be updated whenever a new frame x_n appears for the sample mean $\overline{x}_{s_n}$ should be updated.

4. COMBINATION OF MC-SCT AND ADAPTATION

The model adaptation approaches utilize the correlation information among different models to adapt the model

parameters to fit the speaker and the environment, while the MC-SCT approach utilizes the spatial correlation information among different acoustic units to find new acoustic features which can achieve better discriminative performance. Therefore it is desirable to combine the two approaches by applying MC-SCT after the model adaptation approaches.

5. EXPERIMENT RESULTS

In order to evaluate the performance of MC-SCT, experiments were carried out on a Chinese LVCSR task. The speech database was provided by National 863 High Technology Project. The training data was collected from 76 female speakers each with 650 sentences, and the testing data from another 7 female speakers, each with the same amount of sentences.

In our recognition system, there are 1254 Chinese syllables; each syllable is made up of one initial and one final. There are 100 initials and 164 finals in total. As one initial is divided into two states and one final into four, each syllable is modeled as a six-state HMM. Thus, totally, we have 856 states, each being modeled as a single Gaussian with full covariance. The acoustic feature vector consists of 45 features formed by 14 Mel-frequency cepstrum coefficients with their 1^{st} and 2^{nd} derivatives and the frame energy with its 1^{st} and 2^{nd} derivatives.

In the experiments, we focus on the acoustic part. The speech utterances are recognized to be free syllable strings without any grammar constraints, and the result is organized into syllable-lattices. No language model is used, and the Syllable Error Rate (SER) results are reported for performance evaluation.

For convenience, we use SCT1 here to denote the original MC-SCT, and SCT2 to denote the improved approach proposed in this paper. To evaluate the two schemes of MC-SCT, we compared their performance with MLLR (LR) and Eigenvoice (EV), and the combinations of SCT2 and LR/EV. Experiments were carried out in unsupervised, enrolled and batch mode. For each test speaker, an increasing number of sentences were used as history data, with the recognition result of the SI model as the state labels, while all the sentences were used as test data. The average result is shown in Table 1.

Table 1 Comparison of SER (%) for MC-SCT, MLLR and EV

nSent	LR	EV	SCT1	SCT2	LR+ SCT2	EV+ SCT2
0	28.87	28.87	28.87	28.87	28.87	28.87
1	29.11	27.03	29.50	27.69	27.95	28.25
5	28.41	26.10	26.13	26.19	27.36	26.07
10	28.82	25.74	25.77	25.69	27.96	25.50
50	26.85	25.36	25.28	25.04	25.76	24.80
100	27.56	25.29	25.20	24.73	26.69	24.62
200	27.02	25.20	25.09	24.54	25.99	24.45

As shown in Table 1, the new scheme does improve the performance of MC-SCT. Compared with the adaptation approaches, MC-SCT is very competitive. It nearly always outperforms MLLR in the unsupervised mode, and shows more and more advantage over EV when the sentence number is larger than 10. On the other hand, when MC-SCT is applied based on MLLR or EV, it always improves the performance of the baseline. But the combinations' performance is not always better than MC-SCT itself.

6. CONCLUSION

Spatial correlation information is useful knowledge source to improve the performance of the speech recognition system. Minimum Covariance based Spatial Correlation Transformation (MC-SCT) has shown its effectiveness in utilizing spatial correlation information in decoding process. This paper proposes the improvements on two issues of MC-SCT, 1) the estimation of the transformation matrix; 2) the construction of the history superverctor. Experiment results show that the improvements obtain more advantage over adaptation approaches, and that MC-SCT can be combined with adaptation approaches in the same recognition system.

The current MC-SCT approach is sentence-based. Further study is necessary to improve it to be a frame-based approach, which will improve the Viterbi decoding process by using all the frames having appeared.

7. REFERENCES

[1] Steve Young, "Statistical modelling in continuous speech recognition," *Proc. International Conference on Uncertainty in Artificial Intelligence*, Seattle, 2001.

[2] C.J.Leggetter and P.C.Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.

[3] T. Hazen, "The use of speaker correlation information for automatic speech recognition," Ph.D. diss., Mass. Inst. Technol., Cambridge, Jan. 1998.

[4] R. Kuhn, J.C. Junqua, P. Nguyen, et al, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans on Speech and Audio Processing*, vol. 8, no. 6, pp. 695 -707, Nov. 2000.

[5] Yu Peng, "Studies on spatial dependence information in speech recognition," Ph.D. diss., EE dept., Tsinghua University, Apr. 2002.

[6] Yu Peng, Wang Zuoying, "Using spatial correlation information in speech recognition," in *Eurospeech* 2001, Scandinavia, vol. 3, pp. 1629-1632.

[7] Yu Peng, Wang Zuoying, "Spatial correlated maximum a posteriori adaptation algorithm," *Chinese Journal of Electronics*, vol. 11, no. 3, pp. 336-340, Jul. 2002.

[8] Tengrong Su, Ji Wu, Zuoying Wang, "Spatial correlation transformation based on minimum covariance," in *ICASSP* 2008, Las Vegas, pp. 4697-4700.

[9] M.A. Woodbury, "Inverting modified matrices," *Memorandum Report* 42, Statistical Research Group, NJ, Princeton, 1950.