SPATIAL CORRELATION TRANSFORMATION BASED ON MINIMUM COVARIANCE

Tengrong Su, Ji Wu, Zuoying Wang

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

ABSTRACT

In speech recognition, acoustic units are highly related. Different from some adaptation methods, such as Reference Speaker Weighting (RSW) and Eigenvoice, the correlation between different acoustic units in the feature space, which is called Spatial Correlation, focuses on the correlation information among different acoustic units of the same speaker. In this paper, a novel scheme using spatial correlation is proposed. In speech recognition system, with the spatial correlation information, the refined acoustic models are trained, and the transformation matrices are determined based on Minimum Covariance criteria. Experiments of this new algorithm show a significant improvement on speaker independent recognition systems.

Index Terms— Speech recognition, spatial correlation, feature transformation, minimum covariance

1. INTRODUCTION

As is well known, acoustic units vary greatly with speakers. But they are far from independent. Conversely, since they are all from the speech organs of human being, and some fixed pronunciation rules should be obeyed when they are produced, there must be some relationship between the acoustic units. What's more, the relationship between different acoustic units of the same speaker might be stable. From the viewpoint of speech recognition, this relationship between acoustic units can be described by the correlation among acoustic model parameters in the feature space, so we call it Spatial Correlation.

Hazen [1] used some relation information of acoustic models in his Ph.D. work. He focused on the relationship between the acoustic models of different speakers, which he called as Speaker Correlation. He proposed an adaptation technique called Reference Speaker Weighting (RSW). In this technique, each speaker is represented by a speaker vector, which is made up of his acoustic model parameters. The basic idea of RSW is that a new speaker vector can be constructed from a weighted combination of a set of individual reference speaker vectors. Eigenvoice [2] improved the idea of RSW. It applies principal component analysis (PCA) to the covariance or correlation matrix calculated between the reference speaker vectors, to find a set of eigenvectors-eigenvoices. Then the new speaker vector is represented by a linear combination of the eigenvoices. With little adaptation data, the Eigenvoice approach significantly outperforms the traditional adaptation algorithms such as maximum a posteriori (MAP) [3] and maximum likelihood linear regression (MLLR) [4].

Yu [5] analyzed the spatial dependence between acoustic units quantitatively in his Ph.D. work. He proposed the "Dependence Coefficient" method to analyze the relationship between parameters of different acoustic models of the same speaker. One of his important conclusions is that when the parameters of two acoustic units are strongly dependent, the linear dependence, which is called Spatial Correlation, is dominant, while the nonlinear dependence can be ignored. Based on his analysis on the spatial dependence, 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 the two methods achieve quite good performance. It is shown that spatial correlation can be very useful in speech recognition.

In previous work, the spatial correlation information is only applied to acoustic model training with limited data and speaker adaptation. It seems to be comparatively difficult to use it in decoding process of speech recognition. In this paper, a novel scheme is elaborately designed to solve this problem. A refined acoustic model using the spatial correlation information and the corresponding training algorithm are proposed in this scheme to achieve much better discrimination, and a feature transformation based on Minimum Covariance criteria is introduced to derive new acoustic feature in decoding process.

This paper is organized as follows. In Section 2, we describe the scheme to get a refined acoustic model. In Section 3, the method to calculate transformation matrix of acoustic feature in decoding process is introduced. In Section 4, we discuss how to apply the algorithm in speech

recognizer implementation. In Section 5, the experiments and results are presented. Finally, we summarize our findings and outline our future work.

2. LINEAR DESCRIPTION OF ACOUSTIC UNITS

In order to use the spatial correlation information in the decoding process of speech recognition, we propose a novel scheme to get a refined acoustic model. To begin, assume that the recognition system has got a set of history frames, $x_1, x_2, \cdots, x_n$, and the current frame y. And assume that all the frames are Gaussian with zero mean. Let the dimension of the frames be defined as D. Then let supervector $x = (x_1^T, x_2^T, \cdots, x_n^T)^T$ represent all the history frames. Thus the dimension of x is nD. Use x and y to construct a new feature vector

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

where W is a $D \times nD$ matrix. Obviously, the new vector z is also a Gaussian vector with zero mean. And the covariance matrix of z is expressed as

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

Define the comparison of two covariance matrices as:

$$R_1 < R_2, \quad if \quad \alpha^T R_1 \alpha < \alpha^T R_2 \alpha, \forall \alpha \neq 0 \tag{3}$$

If the covariance matrix of the new feature z is less than the original feature y in the sense of Equation (3), the new feature and the corresponding acoustic model are supposed to achieve better discriminative performance, as shown in the following sketch:

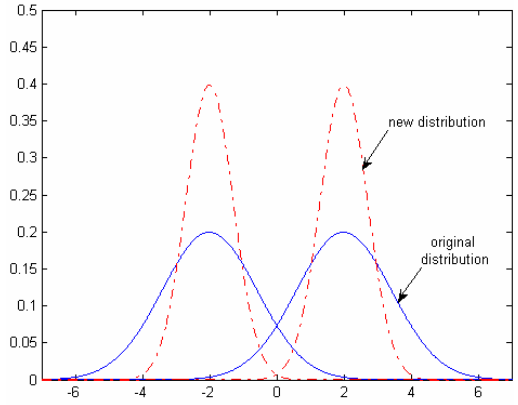

Figure 1 Sketch of performance of less covariance

Choose the transformation matrix W to minimize the covariance matrix R_z. As a result, the distribution of the vector z will be the "narrowest", and consequently the discriminative performance of the new vector z and the new model R_z will be the best.

Define

$$J(W, \alpha) = \alpha^T R_z \alpha = E[(\alpha^T (y - Wx))^2] \tag{4}$$

Then the determination of the transformation matrix W is an optimum problem as following:

$$\min_W J(W, \alpha), \forall \alpha \neq 0 \tag{5}$$

Let $\partial J / \partial W = 0$, we get

$$E[\alpha \alpha^T (y - Wx) x^T] = 0 \tag{6}$$

Since α can be any nonzero vectors, the equation can be equivalently written as:

$$E[(y - Wx)x^T] = E[yx^T] - WE[xx^T] = 0 \tag{7}$$

Finally the optimum transformation matrix W is expressed as:

$$W = E[yx^T]E[xx^T]^{-1} = R_{yx} R_x^{-1} \tag{8}$$

Thus, the new vector and its covariance are expressed as:

$$z = y - R_{yx} R_x^{-1} x \tag{9}$$

$$R_z = R_y - R_{yx} R_x^{-1} R_{xy} \tag{10}$$

where R_x is the autocorrelation matrix of x, while R_{yx} is the correlation matrix between y and x, and R_{xy} is the transpose of R_{yx}. It can be proved that R_z is less than R_y in the sense of Equation (3). Further more, it can also be proved that the more history frames are used, the less R_z will be.

3. CALCULATION OF TRANSFORMATION MATRIX

To calculate the transformation matrix W, we need to calculate the autocorrelation matrix R_x and the correlation matrix R_{yx}. The two matrices can be estimated from a set of supervectors, which are constructed from the acoustic model parameters according to the state labels. Suppose we have already trained a set of speaker-dependant (SD) acoustic models. For each speaker, a supervector corresponding to the history data x is constructed from his SD acoustic model parameters. The supervector $U^{(p)}$ for speaker p is defined as:

$$U^{(p)} = \begin{pmatrix} c_{s_1}^{(p)} \\ \vdots \\ c_{s_n}^{(p)} \end{pmatrix} \tag{11}$$

where $c_{s_i}^{(p)} = \mu_{s_i}^{(p)} - \mu_{s_i}, i = 1, \cdots, n$, while s_i denoting the state of the history data x_i, and $\mu_{s_i}^{(p)}$, μ_{s_i} denoting the mean vectors of state s_i of speaker p and the speaker-independent (SI) model separately.

Let the number of speakers be defined as P. Then the autocorrelation matrix R_x can be expressed as:

$$R_x = \frac{1}{P}\sum_{p=1}^{P} U^{(p)}U^{(p)T} = \frac{1}{P}UU^T \tag{12}$$

where $U = [U^{(1)}, U^{(2)}, \cdots, U^{(P)}]$.

Define a parameter matrix U_{s_i} for state s_i, which is given as:

$$U_{s_i} = [c_{s_i}^{(1)}, c_{s_i}^{(2)}, \cdots, c_{s_i}^{(P)}] \tag{13}$$

Then the super matrix U can be rewritten as:

$$U = \begin{pmatrix} U_{s_1} \\ \vdots \\ U_{s_n} \end{pmatrix} \tag{14}$$

Thus the correlation matrix R_{yx} can be expressed as:

$$R_{yx} = \frac{1}{P}U_{s_y}U^T \tag{15}$$

Consequently the transformation matrix W is expressed as:

$$\begin{aligned} W &= \frac{1}{P}U_{s_y}U^T \cdot (\frac{1}{P}UU^T)^{-1} \\ &= U_{s_y}U^T(UU^T)^{-1} \end{aligned} \tag{16}$$

Since the number of speakers is much less than the number of history frames, the autocorrelation matrix R_x calculated by Equation (12) is always rank-deficient, that is, non-invertible. But we can use the Moore-Penrose inverse of this matrix to substitute its inverse matrix.

Assume we have a full column rank matrix $F_{m\times n}$, and a full row rank matrix $G_{r\times m}$, then

$$(FG)^{-} = G^H(F^HFGG^H)^{-1}F^H \tag{17}$$

is a Moore-Penrose inverse of the matrix FG [8].

Since the super matrix U always has full column rank, Equation (16) can be rewritten as:

$$W = U_{s_y}U^T \cdot (UU^T)^{-} = U_{s_y}(U^TU)^{-1}U^T \tag{18}$$

Finally, we can get the expressions of new feature and corresponding autocorrelation matrix as following:

$$z = y - U_{s_y}(\sum_{i=1}^{n} U_{s_i}^T U_{s_i})^{-1}(\sum_{i=1}^{n} U_{s_i}^T x_i) \tag{19}$$

$$R_z = R_y - \frac{1}{P}U_{s_y}U_{s_y}^T \tag{20}$$

In order to obtain the new feature in the time-synchronous decoding process, we define:

$$A_n = \sum_{i=1}^{n} U_{S_i}^T U_{S_i} = A_{n-1} + U_{s_n}U_{s_n}^T \tag{21}$$

$$b_n = \sum_{i=1}^{n} U_{S_i}^T x_i = b_{n-1} + U_{s_n}^T x_n \tag{22}$$

$$\beta_n = A_n^{-1}b_n \tag{23}$$

Equation (19) can then be rewritten as:

$$z = y - U_{s_y}\beta_n \tag{24}$$

According to Woodbury Formula [9], we get

$$\begin{aligned} A_n^{-1} &= (A_{n-1} + U_{s_n}U_{s_n}^T)^{-1} \\ &= A_{n-1}^{-1} - A_{n-1}^{-1}U_{s_n}^T(I + U_{s_n}A_{n-1}^{-1}U_{s_n}^T)^{-1}U_{s_n}A_{n-1}^{-1} \end{aligned} \tag{25}$$

Thus Equation (23) can be rewritten as:

$$\beta_n = \beta_{n-1} + A_{n-1}^{-1}U_{s_n}^T(I + U_{s_n}A_{n-1}^{-1}U_{s_n}^T)^{-1}(x_n - U_{s_n}\beta_{n-1}) \tag{26}$$

Eventually, the values of the new feature and its corresponding model parameters can be calculated instantaneously. And the computation cost is always the same even if the number of the history frames increases. It provides a good way of utilizing the spatial correlation information in decoding process of speech recognition. The computation cost mainly depends on the computation of the inverse of the $D\times D$ matrix $(I + U_{s_n}A_{n-1}^{-1}U_{s_n}^T)$.

4. APPLICATION IN RECOGNITION

As Equation (18) shows, the transformation matrix is determined by the state labels of the current frame and the previous frames in the history. In a speech recognition system, however, the state labels are unknown. We can take the recognition result as the state labels of the previous frames, but how can we decide the state label of the current frame? It's the problem we should solve before the algorithm can be applied.

Fortunately, in the frame-synchronous decoding process, we should search all the states and calculate the output probability of the vector. Thus, when the state being searched is s, we apply the transformation as following:

$$z_s = y - W_s x \tag{27}$$

where $W_s = R_{sx}R_x^{-1}$. Then the new features are applied in the Viterbi search progress, using the output probability $p(z_s \mid \theta_s)$, where θ_s denotes the acoustic model of the new feature z_s.

5. EXPERIMENT RESULTS

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 1st and 2nd derivatives and 1st and 2nd derivative of the frame energy.

To evaluate the performance of spatial correlation transformation, experiments were carried out on a Chinese LVCSR task, and the speech database is provided by National 863 High Technology Project. The training data comprises of 650 sentences each from 76 female speakers, the same amount of data from another 7 female speakers are used as the testing data.

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.

Eigenvoice (EV) has good performance in speaker adaptation, especially when the adaptation data is very limited.

In order to evaluate our new algorithm (MC-SCT) for utilizing spatial correlation information, we compared its performance with EV. These experiments were carried on enrolled and batch mode, although MC-SCT can be implemented easily on instantaneous and on-line mode. For each test speaker, an increasing number of sentences were used as history data, with the recognition result of 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 and EV

nSent	MC-SCT	EV
0	30.26%	30.26%
10	25.77%	25.74%
20	25.65%	25.61%
40	25.44%	25.47%
60	25.27%	25.34%
80	25.16%	25.33%

As shown in Table 1, MC-SCT obtains obvious SER decline with few history sentences. And the SER keeps on descending as the history sentence number increases. It shows that MC-SCT and EV have similar performance when the adaptation data is very limited. Further more, the asymptotic property of MC-SCT is better than EV. When the sentence number is 10, the relative decline of SER is 14.8% for MC-SCT and 14.9% for EV. When the sentence number is 80, it is 17.6% for MC-SCT and 17.1% for EV.

6. CONCLUSION

Spatial correlation is very important in describing the relationship among the acoustic units. In this paper, a scheme based on minimum covariance to utilize the spatial correlation information efficiently is proposed. Instead of adapting the acoustic models to fit for the speaker and environment as the speaker adaptation methods do, this approach is applied to find refined acoustic features and corresponding models which can achieve better discriminative performance. Experimental results show that there is observable SER decline over the SI recognition system in unsupervised mode. And compared with the Eigenvoice approach, it obtains better asymptotic property. Further more, it is easily to be applied in instantaneous and on-line mode with low computation cost. With the recognition process going on, and longer history and more data gained, lower SER can be obtained.

It is just the beginning for us to utilize spatial correlation information in decoding process of speech recognition. There is still much further work to do in the future. From the result we already have, we believe that this approach has big potential.

7. REFERENCES

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

[2] 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.

[3] J. Gauvain and C. Lee, "Maximum a Posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. Speech Audio Proc.*, vol. 2, pp. 291–298, April 1994.

[4] 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.

[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] Zhang Xianda, *Matrix analysis and applications*, Tsinghua University Press, Beijing, P.R.China, 2004.

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