Improved Covariance Modeling for GMM in Speaker Identification

Xi Zhou, Zhi-qiang Yao, Bei-qian Dai

Dept. of Electronic Science and Technology,
University of Science and Technology of China
zhouxi@mail.ustc.edu.cn, zqyao@ustc.edu

Abstract

Gaussian Mixture Model (GMM) with diagonal covariance matrix is commonly used in text-independent speaker identification. However, diagonal covariance matrix implies strong assumption that the feature elements are independent. Even Gaussian mixtures with diagonal covariance can model the correlation to some extent; the model precision is still limited. To alleviate this problem, this paper proposes a framework for sharing linear transformations among the components and introduces a new unsupervised hierarchical clustering algorithm to implement it. In the framework, the full covariance of each component is represented by shared linear transformation and component-specific diagonal covariance. Different linear transformation estimation approaches, i.e., PCA, LDA and MLLT, are proposed and compared. Experiments show that our algorithm using each of the approaches has achieved significant identification error reduction over the best diagonal covariance models.

1. Introduction

Modeling data using Gaussian or Gaussian mixture distributions is very common in pattern recognition, such as text-independent speaker identification. In GMM-based systems there is a basic choice in the form of covariance matrix to be used. It may either be diagonal, block-diagonal, or full. Diagonal covariance matrix implies strong assumption that elements of the feature vector are independent; while full or block-diagonal covariance matrix can explicitly model all or some of the correlations, but it suffers from a dramatic parameters increase, and the increasing parameters are hard to be estimated robustly by limited data.

Though some feature-space based linear transformations schemes, such as Karhunen-Loeve transform [1], Maximum Likelihood Linear Transform (MLLT) [2] have been used for decorrelating the feature, it is hard to find a single transform which is able to decorrelate all elements of the feature vector.

Sharing parameters across components of GMM (or constraining the parameters) to describe the correlation of inter feature-vector element is more flexible to alleviate this problem. The approach can acquire a tradeoff between correlation's modeling and the number of parameters. Moreover, it also decreases storage requirement and computational requirement [2, 4].

Semi-Tied Covariance (STC) [5] which is a successful application of sharing parameters across components of HMM has achieved significant improvement in Large Vocabulary Continuous Speech Recognition (LVCSR). This paper adopts the thought of STC and applies it to GMM in speaker identification. Firstly, all the Gaussian components are classified as multiple components sets. Then each set estimates a linear transformation shared by all components of this set.

For our sharing algorithm to be effective it is desirable to put all the components that have similar correlation of inter feature-vector element into the same set. This paper introduces an unsupervised hierarchical cluster algorithm to classify the components. A distance measure formula based on maximum likelihood (ML) is used for clustering, which primary considers the comparability of covariance of each pair component. Reasonable classification helps those full covariance matrices to describe correlation modeling more precisely. Through our new clustering algorithm, the components with similar covariance matrices are classified to the same set, which is effective to estimate a linear transformation shared by these components.

Since our algorithm provides a good framework for sharing linear transformations among the components, we adopt three different approaches PCA [6], LDA [7] and MLLT [2] respectively to estimate transforms. Experiments show that our algorithm using each of the approaches could achieve above 30% identification error reduction over the best diagonal covariance models on the MSRA mandarin task [10], which strongly indicates the stability and robustness of our algorithm.

The paper is organized as following: Section 2 describes the format of GMM when sharing parameters across components. In section 3, the new unsupervised hierarchical cluster algorithm is shown with the deduction of distance measure. The new technique is evaluated on a text-independent speaker identification task and corresponding experiments results is shown in section 4. Finally, the conclusion and the future work are presented.

2. Sharing Full Covariance Matrices in GMM

As mentioned in the introduction, in GMM based systems, the form of covariance matrix could be diagonal, block diagonal, or full. Compared to the diagonal case, the full covariance matrix case has the advantage that it models inter feature-vector element correlation. However this is at the cost of a greatly increased number of parameters, $n(n+3)/2$ compared to $2n$ per component, including the mean vector and the covariance matrix, where n is the dimensionality. Due to this massive increase in the number of parameters, diagonal covariance matrices are commonly used in text-independent speaker identification.

Sharing parameters across components clearly tends to acquire a tradeoff between correlation's modeling and the number of parameters, which therefore improves covariance modeling for GMM. By increasing few parameters upon diagonal covariance modeling, we expect that the effect of modeling inter feature-vector element correlation can approach to that of full covariance situation. Moreover, sharing parameters also decreases storage requirement and computational requirement.

This paper adopts the thought of a covariance matrices sharing scheme, Semi-tied covariance matrices (STC) [5] to share the parameters. STC is a simple extension to the standard diagonal, block-diagonal, or full covariance matrices used with HMM's. Instead of having a distinct covariance matrix for every component, the components classified into a set can share a linear transformation. Then, for a specific component, the covariance matrix consists of two elements, a component specific diagonal covariance element, $\Sigma_{diag}^{(m)}$, and a *semi-tied* class-dependent, nondiagonal matrix, $F^{(\gamma_m)}$ (referred to as the semi-tied transform). The form of the covariance matrix is then

$$\breve{\Sigma}^{(m)} = F^{(\gamma_m)}\Sigma_{diag}^{(m)}F^{(\gamma_m)T} \quad (1)$$

$F^{(\gamma_m)}$ describes the inter feature-vector element correlation which may be tied over components of a set ; while $\Sigma_{diag}^{(m)}$ describes the scale of every dimension for a specific component.

To estimate $F^{(\gamma_m)}$, we need firstly to classify components into several sets, which is an important issue to be considered. From Equ (1)， we can see that if two components have the same full covariance matrix, then we can easily acquire the semi-tied transform $F^{(\gamma_m)}$ shared by the two components and it can get the same effect of modeling inter feature-vector element correlation as respectively using two full covariance matrices. The classification algorithm will be described in the next section.

3. Components Classification

We do the classification as follows. The first stage is to compute some distance-like measure between each pair of components. Since the function of the semi-tied transform $F^{(\gamma_m)}$ is to describe the inter feature-vector element correlation of components in the same set, the distance measure should considers the comparability of covariance of each pair components. Then we create a hierarchical clustering tree by starting with each component in its own cluster and recursively merging clusters according to some distance-related criterion. At last, this paper uses the level cutting to select the best partition of the candidate clusters.

3.1 Distance Measure

Because we only care about the distance of covariance matrices, mean of all components are firstly normalized to zero. Then the distance between two components M_1 and M_2 can be derived from a likelihood ratio of the likelihood H_0 that all features belonging to either of the two components are described with a single Gaussian distribution and the likelihood H_1 that the features of per component are described separately.

Let x_1 denote the features belonging to component M_1 and $L(x_1:\omega_1)$ be the likelihood of the x_1 vectors, where ω_1 denotes maximum likelihood estimates for the single gaussian based on samples in the x_1. $L(x_0:\omega_0)$ and $L(x_2:\omega_2)$ are similarly defined, where x_0 denotes the features belonging to the combined collection of x_1 and x_2 . Then the likelihood $\Pr(H_0)=L(x_0:\omega_0)$ and the likelihood $\Pr(H_1)=L(x_1:\omega_1)L(x_2:\omega_2)$. Let λ_L denote the likelihood ratio, thus

$$\lambda_L = \frac{\Pr(H_0)}{\Pr(H_1)} = \frac{L(x_0:\omega_0)}{L(x_1:\omega_1)L(x_2:\omega_2)} \quad (2)$$

The distance measure between two components used in the clustering is

$$d_L(\omega_1,\omega_2) = -\log(\lambda_L)\,. \quad (3)$$

According to the formula in [2], the expression of $L(x_1:\omega_1)$ is given by

$$L(x_1:\omega_1) = \prod_{i=1}^{N_1}\frac{1}{(2\pi)^{d/2}\left|\Sigma_1\right|^{1/2}}\exp\left\{-\frac{1}{2}o_i{}'\Sigma_1^{-1}o_i\right\} = g(N_1,d)*\left|\Sigma_1\right|^{-\frac{N_1}{2}} \quad (4)$$

Also $L(x_0 : \omega_0)$ and $L(x_2 : \omega_2)$ can be expressed as

$$L(x_0 : \omega_0) = g(N, d) * \left|\Sigma_0\right|^{-\frac{N}{2}}$$

$$L(x_2 : \omega_2) = g(N_2, d) * \left|\Sigma_2\right|^{-\frac{N_2}{2}}$$

Where $g(N, d) = (2\pi e)^{-\frac{Nd}{2}}$, $N = N_1 + N_2$
Therefore we can get that

$$\lambda_L = \frac{\Pr(H_0)}{\Pr(H_1)} = \frac{\left|\Sigma_0\right|^{-\frac{N}{2}}}{\left|\Sigma_1\right|^{-\frac{N_1}{2}} * \left|\Sigma_2\right|^{-\frac{N_2}{2}}} \quad (5)$$

λ_L is only dominated by the three full covariance matrices. From Equ (3) and (5), we can find that $d_L(\omega_1, \omega_2)$ is close to zero when the covariance matrices Σ_1 and Σ_2 are similar; while if Σ_1 and Σ_2 are very different, then $d_L(\omega_1, \omega_2)$ will be large and therefore the two components have low probability to be classified into a set.

As the inter feature-vector element correlation is more important than the scale of every dimension for our classification, we may adopt correlation matrix R to take place of covariance matrix Σ for distance measure.

3.2 Hierarchical clustering

Hierarchical clustering is a well-known method for generating candidate clusters from a distance matrix which gives the distance between pairs of single components [8]. Agglomerative clustering algorithms first set all components start as singleton clusters and then iteratively merge the closest pair of clusters according to the distance matrix until only a single cluster containing all the components exists.

Some method is requested to construct a distance between clusters as only the distance between pairs of single components can be directly obtained from the distance matrix. Typical methods include [9]:

Minimum pair Pick the smallest distance between a component in one cluster and one in the other cluster.
Maximum pair Pick the largest distance from a between-clusters pair of components.
Average pair This is the mean of all the component distances.

All the three methods can be used here, and very little is known about what qualities make a cluster distance good for clustering. However, the likelihood based distance measure d_L used in this paper relies on estimates of full covariance matrices, which is based on feature-vectors contained in this cluster. These estimates are more reliable when computed using more vectors. So we re-estimate full covariance matrices Σ_0, Σ_1 and Σ_2 for every new merged cluster, then calculate distance using Equ (5).

A tree produced by hierarchical clustering contains exactly $2N-1$ candidate clusters (this includes the N singletons clusters). This paper uses the simplest method, called level cutting, to select the best partition of the candidate clusters. It is equivalent to slicing the tree horizontally at each merge level.

4. Experiment

4.1 Database

All speaker identification experiments were performed on the MSRA's mandarin speech corpora [10], which include 100 male speakers and about 200 sentences per speaker. The duration of each sentence varies from 3 seconds to 9 seconds. All waveforms were recorded at a sampling rate of 16,000 Hertz and 16 bits per sample using close talking microphones connected to Creative Lab Soundblaster cards in quiet office environments.

In all experiments, we choose 20 sentences per speaker to train GMM and other non-overlap 40 sentences, total 4000 for test. The total length of 20 sentences per speaker for training is about 2.5 minutes. The speech was parameterized into 16 MFCC's, C_1 to C_{16}, along with the first differentials of these parameters (computed in 5 consecutive frames). This yielded a 32-dimensional feature vector.

The baseline system used for identification task was standard diagonal covariance GMM system [11]. We investigate the performance of GMM system with variance number of Gaussian components. At last, the baseline system chose 32 Gaussian components to model the distribution of a speaker. The speaker identification error rate (IER) of the baseline system is 4.6%.

4.2 Sharing Covariance Model

Since our algorithm provides a good framework for sharing semi-tied transforms among the components, many linear transformation schemes can be used to estimate the shared transforms. In our experiments, we adopt three different approaches PCA [6], LDA [7] and MLLT [2] respectively to estimate shared semi-tied transforms. The result is shown in table 1. We can see that all the approaches can achieve better performances than the baseline model set with diagonal covariance matrices. There is no significant performance difference among the three approaches. When we classify components into 800 sets, using each of the three approaches can achieve more than 40% identification error reduction compare with that of baseline.

We can also see that the performances are improved consistently with the increase of component sets number

from 100 to 800, which strongly indicates the stability and robustness of our algorithm.

Table 1. Identification result by sharing parameters (The baseline identification error rate 4.6%)

estimating approach	transform number			
	100	300	500	800
PCA	3.575%	3.05%	2.8%	2.675%
LDA	3.575%	3.15%	2.7%	2.675%
MLLT	3.275%	2.875%	2.675%	2.650%

4.3 Compared with diagonal model with more components

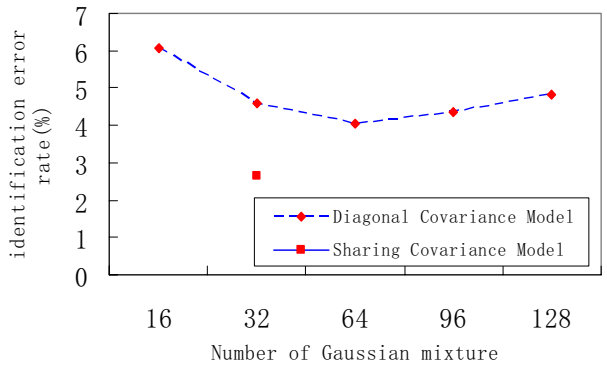

Figure 2. Performance comparison of diagonal models with increasing number of mixtures

In Figure 2, we first show the identification performance (in IER) of several diagonal model sets as a function of their component numbers in each speaker's GMM. When the number of components reaches 64 (the error rate is 4.025%), the identification error rate reduction compared with the baseline is only 12.5%. This shows that it cannot significantly improve the performance over our baseline model by simply increasing the number of diagonal Gaussian mixtures. We also plot the sharing parameters' performance when using MLLT to classify components into 800 sets in the figure as an isolated square point. From the results, it is clear that our algorithm yields much better performance than the diagonal model set with even larger number of mixtures.

5. Conclusions and future work

This paper proposes a framework for sharing linear transformations among the components and introduces a new unsupervised hierarchical clustering algorithm to implement it. Different linear transformation estimation approaches, i.e., PCA, LDA and MLLT, are proposed and compared. We evaluate our algorithm on the MSRA mandarin task. By using each of the three approaches, a significant error reduction over the standard model with diagonal covariance matrices has been observed, which strongly indicates the stability and robustness of our algorithm. In the future, we will the test the effective of the algorithm on larger database.

6 Acknowledgements

The work of this paper was sponsored by the National Science Foundation of China (60272039).

7 References

[1] K. Fukunaga, "Introduction to Statistical Pattern Recognition". *New York: Academic*, 1972.

[2] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. ICASSP*, vol. II, 1998, pp. II-661–II-664.

[3] N. Kumar, "Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition," *Ph.D. dissertation*, Johns Hopkins Univ., Baltimore, MD, 1997.

[4] N. K. Goel, R. Gopinath, "Multiple linear transforms," in *Proceedings of ICASSP*, 2001.

[5] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 272–281, 1999.

[6] M.J.Hunt and C.Lefebre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1989, pp. 262-265

[7] R.Haeb-Umbach, "Linear discriminant analysis for large vocabulary speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, San Francisco, 1992, pp. 13-16

[8] D A Reynolds, E Singer, B A Carlson, G C O'Leary, J J McLaughlin & M A Zissman. "Blind Clustering of Speech Utterances Based on Speaker and Language Characteristics," *Proc. ICSLP'98* Vol. 7 pp. 3193-3196

[9] A Solomonoff, A Mielke, M Schmidt & H Gish. "Clustering Speakers by their Voices," *Proc. ICASSP'98* Vol. 2 pp. 757-760

[10] Eric Chang, Yu Shi, Jianlai Zhou, and Chao Huang. "Speech Lab in a Box: A Mandarin Speech Toolbox to Jumpstart Speech Related Research," *Eurospeech* 2001, Aalborg, Denmark, 2001.

[11] D. A. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture", *J. Acoust. Soc. Amer., Vol. 84, 1995, p91-108.*