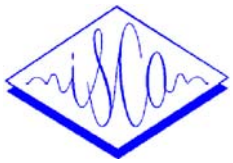
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Speaker Identification Using Probabilistic PCA Model Selection

Jen-Tzung Chien and Chuan-Wei Ting

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, ROC
jtchien@mail.ncku.edu.tw & motor@chien.csie.ncku.edu.tw

10.21437/Interspeech.2004-515

Abstract

Gaussian mixture model (GMM) techniques are popular for speaker identification. Theoretically, each Gaussian function should have a full covariance matrix. However, the diagonal covariance matrix is usually used because the inverse of diagonal covariance matrix can be easily calculated via expectation maximization (EM) algorithm. This paper proposes a new probabilistic principal component analysis (PPCA) model for speaker identification. The full covariance of speaker's data is considered. This model is originated from factor analysis theory. The probability distributions using PPCA are well defined. In particular, GMM and PPCA are found to be equivalent when using diagonal covariance matrix. In this study, we derive a novel PPCA model selection and establish models for different speakers. Applying PPCA model selection, we can dynamically determine the numbers of speech features and mixture components. Experiments show that PPCA achieves desirable speaker recognition performance with proper model regularization.

1. Introduction

Speaker recognition plays an important role for biometric and security applications. Desirable performance can be obtained with moderate computation and storage requirements. In the literature, there are several techniques proposed for speaker recognition such as GMM [9], hidden Markov model (HMM) [12], vector quantization [6] and polynomial classifier [2] based approaches, etc. In speaker identification, when a test speaker is predefined in speaker pool, it is referred as the closed-set identification corresponding to a multiple-choice classification problem. On the contrary, the open-set identification is the case that test speaker does not appear in the pool. In closed-set identification, we can determine the best match speaker according to the input voice. However, in open-set problem, we need to check whether the confidence is high enough to ensure that the best match speaker is the actual speaker. Speaker verification is a kind of simplified open-set problem that only concerns a yes-no hypothesis test problem. On the other hand, speaker identification can be classified into text-dependent and text-independent approaches depend on the contexts of the input voice. In text-dependent speaker identification task, the dynamic time warping or HMMs are used to obtain the phonetic transcriptions of input speech. We then verify whether the recognized transcriptions correspond to predefined words or sentence. For text-independent speaker identification, the spoken words are unknown beforehand. In this study, we adopt the mixture model to construct a closed-set text-independent speaker identification system.

In a speaker identification system, the selection of appropriate features and the determination of model size are also an important topic to construct compact speaker models [5]. The objective of feature selection is to find the salient features from original feature vectors so as to reduce feature dimension. Principal component analysis (PCA) is popular to extract the most expressive features. In this paper, we propose the new speaker identification framework where speaker characteristics are modeled using PPCA [14]. Full covariance matrix is considered in PPCA model rather than assuming diagonal matrix in conventional GMM [10]. We also establish *mixture of PPCA* and relate this model to GMM model. Both models are equivalent when their covariance matrices are diagonal. Basically, mixture of PPCA model constructs the mixture component using full covariance matrix. The original features are transformed into eigenspace to obtain common factors. In this work, we also propose a PPCA model selection method to determine the optimal model structure with appropriate mixture components and feature dimensions.

2. Speaker Identification Framework

2.1. Gaussian Mixture Model (GMM)

GMM is popular to represent speaker characteristics in speaker identification system. Specifically, the distribution of feature vectors extracted from a speaker is modeled using a mixture of Gaussian densities. For a d-dimensional feature vector $\mathbf{y}$, the mixture of Gaussian densities for speaker s is defined as

$$p_{\mathrm{GMM}}(\mathbf{y}|\lambda_s) = \sum_{j=1}^{M} \omega_{s,j} \cdot (2\pi)^{-\frac{d}{2}} |\Sigma_{s,j}|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(\mathbf{y}-\mu_{s,j})^T \Sigma_{s,j}^{-1}(\mathbf{y}-\mu_{s,j})\right\} \quad (1)$$

where speaker model parameter set λ_s consists of mixture weights $\{\omega_{s,j}\}$ with constraint $\sum_{j=1}^{M}\omega_{s,j}=1$, mean vectors $\{\mu_{s,j}\}$ and covariance matrices $\{\Sigma_{s,j}\}$.

Speaker identification using Bayesian decision rule is performed through finding the most likely speaker from his/her input utterance $\mathbf{Y}=\{\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_N\}$, i.e.

$$\hat{s} = \arg\max_s p(\lambda_s|\mathbf{Y}) = \arg\max_s p(\mathbf{Y}|\lambda_s)p(\lambda_s). \quad (2)$$

2.2. Probabilistic Principal Component Analysis (PPCA)

This paper presents a new PPCA model for speaker identification. As we know, principal component analysis (PCA) is popular for dimension reduction in multivariate linear data analysis. This approach has been successfully applied to data compression, image analysis, pattern recognition, regression and time series prediction. Although PCA achieved good performance in many applications, its major disadvantage is the absence of an associated probability density. In the literature, there was an approach connecting the

relation between PCA and GMM. This approach was to use PCA procedure as the post-processor of GMM. Two steps were performed. The first step was to apply vector quantization to split observation data into several disjoint partitions corresponding to different mixture components. Euclidean distance was calculated to measure data closeness. Generally, the GMM was generated in this step. Second, the original features were transformed into PCA eigenspace spanned by the eigenvectors of the covariance matrix of whole observed data or data for each mixture [7]. After these two steps, we could obtain the Gaussian mixture model with transformed principal components. However, this procedure led to neglect some changes after transformation. In this study, we present a new framework merging GMM and PCA via constructing the proba1bility model for PCA. Our procedure to construct probability model is motivated from the *factor analysis* (FA) theory [14]. Using factor analysis, the observation vector **y** is represented by

$$\mathbf{y} = W\mathbf{x} + \mu + \varepsilon, \tag{3}$$

where **x** denotes the q-dimensional latent variable ($q < d$) or called common factors defined to be independent and Gaussian distributed with unit variance $\mathbf{x} \sim N(0, \mathbf{I})$. Common factors are extracted from training data. The noise variable ε is also Gaussian with diagonal covariance matrix Ψ, $\varepsilon \sim N(0, \Psi)$. The $d \times q$ transformation matrix W contains the factor loadings.

Generally, the column space of W is not equivalent to the principal subspace of observation data. In case of isotropic noise $\varepsilon \sim N(0, \sigma^2 \mathbf{I})$, the columns of the maximum-likelihood estimator W_{ML} can span the principal subspace of the data. In this case, the conditional density of **y** given latent variable **x** has the form of [14]

$$p(\mathbf{y} \mid \mathbf{x}) = (2\pi)^{-\frac{d}{2}} (\sigma^2)^{-\frac{d}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \left\| \mathbf{y} - W\mathbf{x} - \mu \right\|^2 \right\}. \tag{4}$$

Because the latent variable has the density

$$p(\mathbf{x}) = (2\pi)^{-\frac{q}{2}} \exp\left\{ -\frac{1}{2} \mathbf{x}^T \mathbf{x} \right\}, \tag{5}$$

the marginal distribution $p(\mathbf{y})$ is obtained by

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} =$$

$$(2\pi)^{-\frac{d}{2}} |C|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \mu)^T C^{-1} (\mathbf{y} - \mu) \right\}, \tag{6}$$

which is a Gaussian density with $d \times 1$ mean vector μ and $d \times d$ covariance matrix $C = \sigma^2 \mathbf{I} + WW^T$. Meanwhile, the posterior distribution is determined by

$$p(\mathbf{x}|\mathbf{y}) = (2\pi)^{-\frac{q}{2}} \left| \sigma^{-2} \mathrm{M} \right|^{\frac{1}{2}} \times \tag{7}$$

$$\exp\left\{ -\frac{1}{2} \left[\mathbf{x} - \mathrm{M}^{-1} W^T (\mathbf{y} - \mu) \right]^T \left(\sigma^{-2} \mathrm{M} \right) \left[\mathbf{x} - \mathrm{M}^{-1} W^T (\mathbf{y} - \mu) \right] \right\}$$

It is also a Gaussian density with mean $\mathrm{M}^{-1} W^T (\mathbf{y} - \mu)$ and $q \times q$ covariance matrix $\sigma^2 \mathrm{M}^{-1} = \sigma^2 (\sigma^2 \mathrm{I} + W^T W)^{-1}$.

With these probability functions, we can obtain the *mixture model for PPCA* analogous to GMM. In [7], PCA transformation is performed after generating GMM. We cannot guarantee that the clustering distribution of mixture component is kept the same after transformation. Some useful information is lost accordingly when constructing mixture components. Using mixture of PPCA, we can prevent missing information due to transformation. Our transformation was done prior to establishing mixture component. In case of single mixture PPCA, the probability density model only uses a global linear model to represent data in low dimensional space. In case of mixture of PPCA, we use a mixture of linear models where the distribution of observation vector $\mathbf{y}_i$ given the unobservable latent variable $\mathbf{x}_i$ is defined in (6). The probability density function is expressed via a mixture of M Gaussian densities with mixture weights $\{\omega_j\}$ as follows

$$p(\mathbf{Y}) = \prod_{i=1}^{N} p(\mathbf{y}_i) = \prod_{i=1}^{N} \left[\sum_{j=1}^{M} \omega_j p(\mathbf{y}_i \mid j) \right]. \tag{8}$$

2.3. Relationship between PPCA & GMM

To illustrate the relationship between PPCA and GMM, we express the mixture of PPCA model in a form of

$$p_{\mathrm{PPCA}}(\mathbf{y}_i) = \sum_{j=1}^{M} \omega_j \cdot (2\pi)^{-\frac{d}{2}} \left| \sigma_j^2 \mathbf{I} + W_j W_j^T \right|^{-\frac{1}{2}} \times$$

$$\exp\left\{ -\frac{1}{2} (\mathbf{y}_i - \mu_j)^T (\sigma_j^2 \mathbf{I} + W_j W_j^T)^{-1} (\mathbf{y}_i - \mu_j) \right\} \tag{9}$$

Comparing (9) of PPCA and (1) of GMM, we find the following properties. In GMM, assuming that the features $\mathbf{y}_i$ of different dimensions are uncorrelated, the covariance matrix Σ_j of mixture component j becomes a diagonal matrix. However, FA theory concentrates on modeling feature correlations. In case of using uncorrelated features, the transformation matrix W turns out to be a $d \times q$ zero matrix. The covariance matrix using PPCA becomes $C_j = \sigma_j^2 \mathbf{I}$, which is a diagonal matrix with identical diagonal components. In this case, the implementation procedure of PPCA is similar to that of GMM. This fact can be revealed from the maximum likelihood (ML) estimates of PPCA model parameters as described in Section 3.1. The mixture of PPCA model is then reduced to

$$p_{\mathrm{PPCA}}(\mathbf{y}_i) = \sum_{j=1}^{M} \omega_j \cdot (2\pi)^{-\frac{d}{2}} \left| \sigma_j^2 \mathbf{I} \right|^{-\frac{1}{2}} \times$$

$$\exp\left\{ -\frac{1}{2} (\mathbf{y}_i - \mu_j)^T (\sigma_j^2 \mathbf{I})^{-1} (\mathbf{y}_i - \mu_j) \right\} \tag{10}$$

This shows that PPCA model $p_{\mathrm{PPCA}}(\mathbf{y}_i)$ is equivalent to GMM model $p_{\mathrm{GMM}}(\mathbf{y}_i)$ when the feature components are uncorrelated and $\Sigma_j = \sigma_j^2 \mathbf{I}$. In what follows, we describe the issues of parameter estimation and model selection for PPCA framework.

3. Parameter Estimation & Model Selection

3.1. PPCA Parameter Estimation

In PPCA, the model parameters λ contain the mixture weights $\{\omega_j\}$, mean vectors $\{\mu_j\}$, transformation matrices $\{W_j\}$ and noise variances $\{\sigma_j^2\}$, i.e. $\lambda = \{\omega_j, \mu_j, W_j, \sigma_j^2\}$. We employ the maximum likelihood (ML) estimation to find λ_{ML} from observation data $\mathbf{Y}$

$$\lambda_{\mathrm{ML}} = \arg\max_{\lambda} p(\mathbf{Y}|\lambda). \quad (11)$$

Here, PPCA involves missing data problem due to mixture component label j and latent variable $\mathbf{x}$. Expectation-maximization (EM) algorithm is applied to resolve the problem via two steps. In E-step, we calculate the expectation function $Q(\hat{\lambda}|\lambda)$ of new estimate $\hat{\lambda}$ given current estimate λ. The expectation operation is done with respect to the unobserved variables $\mathbf{x}$ and j

$$Q(\hat{\lambda}|\lambda) = E[\log p(\mathbf{Y}|\mathbf{X},\hat{\lambda})|\mathbf{Y},\lambda] =$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M} p(j|\mathbf{y}_i,\lambda)\, p(\mathbf{x}_i|\mathbf{y}_i,\lambda) \log \omega_j\, p(\mathbf{y}_i|\mathbf{x}_i,\hat{\lambda}) \propto$$

$$\sum_{i=1}^{N}\sum_{j=1}^{M} \gamma_{ij}\, p(\mathbf{x}_i|\mathbf{y}_i,\lambda)\left\{\log\hat{\omega}_j - \frac{d}{2}\log\hat{\sigma}_j^2 - \frac{1}{2\hat{\sigma}_j^2}\left\|\mathbf{y}_i - \hat{W}_j\mathbf{x}_i - \hat{\mu}_j\right\|^2\right\}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M}\gamma_{ij}\left\{\begin{array}{l}\log\hat{\omega}_j - \frac{d}{2}\log\hat{\sigma}_j^2 - \frac{1}{2\hat{\sigma}_j^2}\left\|\mathbf{y}_i - \hat{\mu}_j\right\|^2 - \frac{1}{2}\mathrm{tr}(\overline{\mathbf{x}_{ij}\mathbf{x}_{ij}^T}) \\ + \frac{1}{\hat{\sigma}_j^2}\bar{\mathbf{x}}_{ij}^T\hat{W}_j^T(\mathbf{y}_i - \hat{\mu}_j) - \frac{1}{2\hat{\sigma}_j^2}\mathrm{tr}(\hat{W}_j^T\hat{W}_j\overline{\mathbf{x}_{ij}\mathbf{x}_{ij}^T})\end{array}\right\} \quad (12)$$

where $\gamma_{ij} = p(j|\mathbf{y}_i,\lambda)$ is the posterior probability of mixture component j for generating data $\mathbf{y}_i$ using current estimate λ, $\bar{\mathbf{x}}_{ij} = \mathrm{M}_j^{-1}\hat{W}_j^T(\mathbf{y}_i - \hat{\mu}_j)$, $\overline{\mathbf{x}_{ij}\mathbf{x}_{ij}^T} = \hat{\sigma}_j^2\mathrm{M}_j^{-1} + \bar{\mathbf{x}}_{ij}\bar{\mathbf{x}}_{ij}^T$. As shown in (7), $\bar{\mathbf{x}}_{ij}$ and $\hat{\sigma}_j^2\mathrm{M}_j^{-1}$ are mean vector and covariance matrix of posterior density $p(\mathbf{x}|\mathbf{y})$, respectively. In M-step, we maximize $Q(\hat{\lambda}|\lambda)$ with respect to $\hat{\lambda} = \{\hat{\omega}_j, \hat{\mu}_j, \hat{W}_j, \hat{\sigma}_j^2\}$ and obtain the ML new estimates

$$\hat{\omega}_j = \left(\sum_{i=1}^{N}\sum_{j=1}^{M}\gamma_{ij}\right)^{-1}\left(\sum_{i=1}^{N}\gamma_{ij}\right) \quad (13)$$

$$\hat{\mu}_j = \left(\sum_{i=1}^{N}\gamma_{ij}\right)^{-1}\left(\sum_{i=1}^{N}\gamma_{ij}\left[\hat{W}_j^T\bar{\mathbf{x}}_{ij} - \mathbf{y}_i\right]\right) \quad (14)$$

$$\hat{W}_j = \left(\sum_{i=1}^{N}\gamma_{ij}\overline{\mathbf{x}_{ij}\mathbf{x}_{ij}^T}\right)^{-1}\left(\sum_{i=1}^{N}\gamma_{ij}(\mathbf{y}_i - \hat{\mu}_j)\bar{\mathbf{x}}_{ij}^T\right) \quad (15)$$

$$\hat{\sigma}_j^2 = \left(d\cdot\sum_{i=1}^{N}\gamma_{ij}\right)^{-1}\left(\sum_{i=1}^{N}\gamma_{ij}\left[\begin{array}{l}\left\|\mathbf{y}_i - \hat{\mu}_j\right\|^2 - 2\bar{\mathbf{x}}_{ij}^T\hat{W}_j^T(\mathbf{y}_i - \hat{\mu}_j) \\ + \mathrm{tr}(\hat{W}_j^T\hat{W}_j\overline{\mathbf{x}_{ij}\mathbf{x}_{ij}^T})\end{array}\right]\right) \quad (16)$$

EM algorithm guarantees that the likelihood of new estimates is higher than that of current estimates. The ML estimates are determined when the convergence condition is met.

3.2. Model Selection Criteria

Conventionally, the optimal model is constructed according to the likelihood measure. But, the overlarge model is prone to be found. To penalize overlarge model, it is popular to apply model selection criteria and perform model regularization. In the literature, there are many model selection criteria proposed to estimate the size of model parameters. Here, the parameter number m is related to the number of mixture component M and the feature dimension q. AIC [1], BIC [13], MDL [11] were well established and served as universal model selection criteria for different models. Also, we developed the prediction information criterion (PIC) for model selection of HMM [3]. In general, the derived formulas of AIC, BIC and MDL have the form of

$$C(\hat{\lambda}(m), m) = -\log p(\mathbf{Y}|\hat{\lambda}(m)) + \mathrm{P}(m), \quad (17)$$

where $\hat{\lambda}(m)$ is the ML estimate of λ with totally m parameters. Then, we identify the "best" model with parameter number m selecting between $m_{\min}$ and $m_{\max}$ via

$$\hat{m} = \arg\min_{m_{\min}\le m\le m_{\max}} C(\hat{\lambda}(m), m). \quad (18)$$

In (17), the criterion is consisted of a log likelihood function $\log p(\mathbf{Y}|\hat{\lambda}(m))$ and a penalty $\mathrm{P}(m)$. Penalty $\mathrm{P}(m)$ is proportional to the parameter number m.

3.3. PPCA Model Selection

In this paper, we present a new model selection criterion for PPCA framework. This criterion is feasible to determine model size, or equivalently the mixture number M as well as the feature dimension q. Hopefully, the selected model is as compact as possible. We apply the Bayesian comparison theory to derive the criterion from observation data $\mathbf{Y}$ [8]. In Bayesian formulation, the models are compared via the posterior probability $p(\mathcal{M}|\mathbf{Y})$ assigned to model $\mathcal{M}$

$$p(\mathcal{M}|\mathbf{Y}) \propto p(\mathbf{Y}|\mathcal{M})\, p(\mathcal{M}). \quad (19)$$

Prior probability $p(\mathcal{M})$ means how plausible the alternative models were before the data are collected. Given the observation data $\mathbf{Y}$, we can infer the parameters λ for model $\mathcal{M}$ through the posterior probability

$$p(\lambda|\mathbf{Y},\mathcal{M}) = \frac{p(\mathbf{Y}|\lambda,\mathcal{M})\, p(\lambda|\mathcal{M})}{p(\mathbf{Y}|\mathcal{M})} \quad (20)$$

In (19) and (20), the term $p(\mathbf{Y}|\mathcal{M})$ is so-called *evidence* or *prediction density* [3]. The objective of Bayesian comparison is to select the "best" model by maximizing the posterior probability $p(\mathcal{M}|\mathbf{Y})$ in (19). Assume that there is no strong reason to assign the prior distribution $p(\mathcal{M})$ to alternative models, our aim turns out to maximize the evidence $p(\mathbf{Y}|\mathcal{M})$. To simplify the expressions, we omit the notation $\mathcal{M}$ and derive the evidence for mixture of PPCA as given in (9).

Hence, we can represent the evidence

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} = \int \exp\{g(\mathbf{x})\}\, d\mathbf{x}. \quad (21)$$

In case of single mixture component, $g(\mathbf{x})$ is given by

$$-\frac{(d+q)}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left\|\mathbf{y} - W\mathbf{x} - \mu\right\|^2 - \frac{1}{2}\left\|\mathbf{x}\right\|^2.$$

Using the second-order Taylor approximation, the function $g(\mathbf{x})$ is expressed by

$$g(\mathbf{x}) \cong g(\tilde{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \tilde{\mathbf{x}})^T g''(\tilde{\mathbf{x}})(\mathbf{x} - \tilde{\mathbf{x}}). \quad (22)$$

From 1st order differential $g'(\mathbf{x}) = \sigma^{-2}W^T(\mathbf{y} - W\mathbf{x} - \mu) - \mathbf{x}$, we obtain the Fisher information matrix

$$g''(\mathbf{x}) = -\sigma^{-2}W^TW - \mathbf{I} \quad (23)$$

and the evidence

$$p(\mathbf{y}) \cong \exp\{g(\tilde{\mathbf{x}})\} \cdot \int \exp\left\{\frac{1}{2}(\mathbf{x}-\tilde{\mathbf{x}})^T g''(\tilde{\mathbf{x}})(\mathbf{x}-\tilde{\mathbf{x}})\right\} d\mathbf{x} \quad (24)$$
$$= \exp\{g(\tilde{\mathbf{x}})\} \cdot (2\pi)^{-\frac{q}{2}} \left|-g''(\tilde{\mathbf{x}})\right|^{-\frac{1}{2}}$$

In (24), we adopt the property of integrating a Gaussian density $N(\mathbf{x}|\tilde{\mathbf{x}}, -g''(\tilde{\mathbf{x}})^{-1})$ to be one. Finally, the logarithm of evidence $\log p(\mathbf{y})$ is derived for PPCA model selection

$$\log p(\mathbf{y}|\mathbf{x}) + \log p(\tilde{\mathbf{x}}) + \frac{q}{2}\log(2\pi) - \frac{1}{2}\log\left|\frac{1}{\sigma^2}W^T W + \mathbf{I}\right|. \quad (25)$$

The mixture component number M is a parameter of $p(\mathbf{y}|\mathbf{x})$. In the experiments, we fix M and use this model selection criterion to determine the dimension q.

4. Experiments

We applied the proposed PPCA for text-independent speaker identification (SID). The mixture of PPCA model was adopted to establish speaker models. We also developed the model selection criterion based on Bayesian comparison theory in which the evidence was maximized to find the "best" model. Such criterion was feasible to determine the model size. The common factors with lower dimension $q < d$ were identified. In the experiments, we used TCC300 Mandarin speech database for SID evaluation [4]. All utterances were recorded in 16 bits at 16 kHz sample rate. The original feature vector was consisted of 12 MFCC's, 12 delta MFCC's, one log energy and one delta log energy, i.e. $d = 26$. We sampled 20 utterances from 20 speakers for testing. Each speaker provided training data with 200 seconds. The ML model parameters were trained for each speaker. We used VQ approach to establish the initial population of mixture components for GMM and PPCA. The initial transformation matrix W_j was estimated by performing PCA using mixture dependent data. In Table 1, we compare SID accuracies for different dimensions q. The number of mixture components is set to be eight, $M = 8$. When model selection criterion is applied, SID accuracy 82.5% was achieved. This result is obtained when $q = 22$. This result is even better than 79.8% when using larger model $q = 24$.

Table 1: SID recognition rates (%) using PPCA

$q = 18$	$q = 20$	$q = 22$	$q = 24$	$q = 26$
70	76	**82.5**	79.8	85

5. Conclusions

We have presented a new mixture of PPCA framework for speaker identification. PPCA model was referred as a special case of FA model when the covariance matrix of noise model was assumed to be an isotropic matrix. Under this assumption, we were able to efficiently find the factor loading transformation matrix via eigen-decomposition realization instead of applying gradient decent algorithm. Different from conventional PCA, PPCA model had the strict definitions of probability densities so that we could derive PPCA ML parameters via EM algorithm. The underlying concept of PPCA was to establish the relation between original features and their corresponding common factors. The correlations of different feature components were properly characterized. We also presented the PPCA model selection criterion. The proposed criterion was originated from Bayesian comparison viewpoint. We derived the Fisher information matrix and constructed model selection criterion, which was feasible to determine feature dimension and mixture component number for model regularization. In general, the proposed PPCA model selection had the benefits of compact modeling and efficient computation of model parameters. When features were assumed to be uncorrelated, PPCA was reduced to GMM. The experiments on speaker identification showed that PPCA model selection obtained desirable performance. In the future, we will further explore the MAP PPCA parameter estimates for speaker adaptation. More experiments will be conducted to investigate the memory requirement and computational cost for PPCA and GMM models.

6. References

[1] H. Akaike, "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716-723, 1974.

[2] M. W. Campbell, K. T. Assaleh, and C. C. Broun, "Speaker recognition with polynomial classifiers", *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 4, pp. 205-212, 2002.

[3] J.-T. Chien and S. Furui, "Predictive hidden Markov model selection for decision tree state tying", *EUROSPEECH*, pp. 2701-2704, 2003.

[4] J.-T. Chien and C.-H. Huang, "Bayesian learning of speech duration models", *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 558-567, 2003.

[5] S. Furui, "Recent advances in speaker recognition", *Pattern Recognition Letters*, vol. 18, pp. 859-872, 1997.

[6] J. He, L. Liu and P. Gunther, "A discriminative training algorithm for VQ-based speaker identification", *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 353-356, 1999.

[7] H.-C. Kim, D. Kim, S.-Y. Bang, "Extensions of LDA by PCA mixture model and class-wise features", *Pattern Recognition* vol. 36, pp. 1095-1105, 2003.

[8] D. J. C. Mackay, "Bayesian interpolation", *Neural Computation*, vol. 4, pp. 405-447, 1992.

[9] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", *Speech Communication*, vol. 17, pp. 91-108, 1995.

[10] D. A. Reynolds, R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72-83, 1995.

[11] J. Rissanen, "A universal prior for integers and estimation by minimum description length", *The Annals of Statistics*, vol. 11, no. 2, pp. 416-431, 1983.

[12] M. Roch and R. R. Hurtig, "The integral decode: a smoothing technique for robust HMM-based speaker recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 315-324, 2002.

[13] G. Schwarz, "Estimating the dimension of a model", *The Annals of Statistics*, vol. 6, no. 2, pp. 461-464, 1978.

[14] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers", *Neural Computation*, vol. 11, pp. 443-482, 1999.