Factor Analyzed HMM Topology for Speech Recognition

Chuan-Wei Ting and Jen-Tzung Chien

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan 70101, ROC
{cwting,chien}@chien.csie.ncku.edu.tw

Abstract

This paper presents a new factor analyzed (FA) similarity measure between two Gaussian mixture models (GMMs). An adaptive hidden Markov model (HMM) topology is built to compensate the pronunciation variations in speech recognition. Our idea aims to evaluate whether the variation of a HMM state from new speech data is significant or not and judge if a new state should be generated in the models. Due to the effectiveness of FA data analysis, we measure the GMM similarity by estimating the common factors and specific factors embedded in the HMM means and variances. Similar Gaussian densities are represented by the common factors. Specific factors express the residual of similarity measure. We perform a *composite hypothesis test* due to common factors as well as specific factors. An adaptive HMM topology is accordingly established from continuous collection of training utterances. Experiments show that the proposed FA measure outperforms other measures with comparable size of parameters.

Index Terms: factor analysis, similarity measure, HMM topology, speech recognition

1. Introduction

Factor analysis (FA) is known as a popular data analysis method where the common factors and the specific factors are extracted and accompanied by a factor loading matrix. FA has been developed for data compression [9], signal enhancement [2] and acoustic modeling [3][8]. In [2], FA model was used to build a generalized signal subspace where the noisy signal was projected to a principle subspace with primary clean signal and a complimentary minor subspace which contained delicate speech information. The clean speech was restored from the FA selected subspaces. In [3], FA was employed to explore the underlying trends or variations in acoustic features. The common factors captured similar statistical properties in features while the specific factors expressed the modeling error at each feature. The streamed hidden Markov models (HMMs) were conducted for phone recognition. Nevertheless, not only the variations among features but also the variations due to environmental mismatch are important. In this study, we explore the FA approach to the similarity measure of Gaussian mixture models (GMMs) and apply it to build adaptive HMM topology for speech recognition in presence of noise interference, speaker mismatch or pronunciation variations.

In real-world speech recognition systems, we need to build an adaptive acoustic model to cope with a variety of environmental mismatches. Equivalently, we should construct an adaptive HMM topology to meet all possible variations in the collected speech utterances. The amount of training data also affects the estimation of HMM parameters. In the literature, a maximum likelihood (ML) method [7] was designed to estimate HMM topology by using the successive state splitting. There were two kinds of state splitting for representing the temporal variations and contextual variations

in HMM topology. Pronunciation variations from different speakers were considered in the trained topology. In [6], a sharing scheme was presented for dynamic HMM configuration while the inter-state characteristics were not considered. Typically, these two studies performed the batch learning scheme for HMM. For the situation that the data is continuously observed, the HMM topology should be adaptive to compensate the data variations from new data. The online learning strategy is important to tackle this situation. This idea was similar to the generation of neurons in neural network [4], which was adaptive to fit the nonstationary environments. We recently proposed an incremental learning scheme of HMM topology by performing the hypothesis test of the similarity between new data and the existing HMMs both in state level and in phone level [10]. A new Gaussian component or a new HMM state was produced if the dissimilarity is significant.

In this study, we present the adaptive HMM topology from different perspective. We explore the capability of FA paradigm in measuring the similarity between two GMMs. This capability is adopted to build HMM topology although it is useful in many pattern recognition applications. Traditionally, the Kullback-Leibler divergence (KLD) was used to measure the similarity between two distributions, but the KLD between GMMs is neither analytically tractable nor computationally efficient. In [5], a variational approximation of KLD between GMMs was presented. Interestingly, we cope with this issue from FA perspective. According to FA principle, the common factors are estimated to explore the correlations among mean vectors. We perform the factor analysis of the mean vectors of two GMMs and discover the solution to similarity measure. In an extreme case, we may use only K common factors to reconstruct the invariant mean vectors if two K-component GMMs are almost the same. However, this case is impractical since the pronunciation usually varies for different speakers. Here, we propose a similarity measure as a ratio of the likelihood scores which are calculated from common factors and specific factors. Using this measure, HMM topology adaptively grows up when a new speaking style is observed. The proposed FA based topology is evaluated by the experiments on continuous speech recognition.

2. Background Survey

2.1 Factor analysis

FA is popular for data analysis of multivariate observations. Considering a D-dimensional observation $\mathbf{x}$, the data generation using FA is expressed by [9]

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{W}_f \mathbf{f} + \boldsymbol{\varepsilon} \qquad (1)$$

where $\boldsymbol{\mu}$ is mean vector, $\mathbf{f}$ and $\boldsymbol{\varepsilon}$ denote the $M \times 1$ *common factor* vector and $D \times 1$ *specific factor* vector with zero-mean Gaussian distributions $N(0, I)$ and $N(0, \Psi)$, respectively. The covariance matrix of $\mathbf{f}$ is an identity matrix I and that of $\boldsymbol{\varepsilon}$ is a diagonal matrix Ψ. Two sets of

6 – 10 September, Brighton UK

factors are uncorrelated, i.e. $E[\mathbf{f}\boldsymbol{\varepsilon}^T] = 0$. The covariance matrix of $\mathbf{y}$ is obtained by $\Sigma = \mathbf{W}_f \mathbf{W}_f^T + \Psi$. Here, the $D \times M$ matrix $\mathbf{W}_f$ is a *factor loading matrix* with entry (d,m) recording the correlation between feature x_d and common factor f_m. FA extracts the common factors which are used to capture the correlation between features. If the number of common factors is large, the off-diagonal terms of covariance matrix are well represented. The specific factors are used to compensate the incompleteness of representing diagonal terms. The specific factor is named because it only contains the information of feature itself without carrying any information among features. The FA parameters $\{\mathbf{W}_f, \mathbf{f}, \boldsymbol{\varepsilon}\}$ can be estimated by the principal component method, the principal factor analysis method, or the ML method [9]. We present the FA approach to discover the similarity between GMMs for general pattern recognition. This approach is here applied for adaptive acoustic modeling.

2.2 Similarity between two GMMs

To tackle the variations in acoustic models, we need to know the ways of growing or pruning of GMMs from the existing HMMs and the observed training utterances. Practically, we should compare two probability density functions $p_a(\mathbf{x})$ and $p_b(\mathbf{x})$. Let $X = \{\mathbf{x}_t\}$ denote the set of D-dimensional training samples. Assuming two densities are Gaussian, the Kullback-Leibler divergence $\mathrm{KLD}(p_a(\mathbf{x}) \| p_b(\mathbf{x}))$ is calculated by

$$\frac{1}{2}\left\{\log\frac{|\Sigma_b|}{|\Sigma_a|} + \mathrm{tr}[\Sigma_b^{-1}\Sigma_a] - D + (\mu_a - \mu_b)^T \Sigma_b^{-1}(\mu_a - \mu_b)\right\}. \quad (2)$$

where μ_a, μ_b and Σ_a, Σ_b are the mean vectors and covariance matrices of two Gaussian densities, respectively. The closed-form solution exists in using KLD. However, in many applications, the latent variables are used to express the unseen events or activities embedded in the observation data. For example, the Gaussians in a HMM state are latent in acoustic models and are used to describe the clusters of the speech frames aligned to a specific state or sub-segment of a phone. The GMM density function

$$p(\mathbf{x}_t \mid \lambda_i) = \sum_{k=1}^{K_i} \omega_{ik} N(\mathbf{x}_t \mid \mu_{ik}, \Sigma_{ik}), \quad i = a, b \quad (3)$$

has become a central processor in acoustic models. In (3), λ_i denotes the GMM parameters including K_i sets of mixture weights and Gaussian components $\{\omega_{ik}, \mu_{ik}, \Sigma_{ik}\}$. Unfortunately, the KLD between GMMs is not analytically tractable.

There are three approximations found in [5]. The first one is to use a single Gaussian approximation $\hat{p}(\mathbf{x}_t)$ to represent the whole GMM. The mean vector and covariance matrix are given by

$$\hat{\mu}_i = \sum_{k=1}^{K_i} \omega_{ik} \mu_{ik} \quad (4)$$

$$\hat{\Sigma}_i = \sum_{k=1}^{K_i} \omega_{ik}[\Sigma_{ik} + (\mu_{ik} - \hat{\mu}_i)(\mu_{ik} - \hat{\mu}_i)^T]. \quad (5)$$

Substituting (4) and (5) into (2), the approximation of KLD between GMMs is determined. The second approximation is to identify the most similar pair of Gaussians among all possible combination pairs from two GMMs. The KLD is approximated by

$$\mathrm{KLD}(p_a(\mathbf{x}) \| p_b(\mathbf{x})) = \min_{k_a, k_b} \mathrm{KLD}(p_{k_a}(\mathbf{x}) \| p_{k_b}(\mathbf{x})) \quad (6)$$

where k_a and k_b are the mixture component indices of $p_a(\mathbf{x})$ and $p_b(\mathbf{x})$, respectively. Such an approximation suffers from high computation due to too many pairs of densities. Also, the mixture weight is not considered. Each Gaussian density is equally treated, and shall lead to poor approximation. The third method is a kind of variational approximation [5] where the Jensen inequality was applied to find the lower bounds of two GMMs by $L_a(b) = E_{p_a(x)}[\log p_b(x)]$ and $L_a(a) = E_{p_a(x)}[\log p_a(x)]$. Using these two bounds, the KLD similarity is written as $\mathrm{KLD}(p_a(\mathbf{x}) \| p_b(\mathbf{x})) = L_a(a) - L_a(b)$. More specifically, the variational approximation is calculated by

$$\sum_{k_a=1}^{K_a} \omega_{ak_a} \log \frac{\sum_{k_a'=1}^{K_a} \omega_{ak_a'} \cdot \exp\{-\mathrm{KLD}(p_a(\mathbf{x}) \| p_{a'}(\mathbf{x}))\}}{\sum_{k_b=1}^{K_b} \omega_{bk_b} \cdot \exp\{-\mathrm{KLD}(p_a(\mathbf{x}) \| p_b(\mathbf{x}))\}}. \quad (7)$$

3. Factor Analyzed HMM Topology

We present a new similarity measure between two GMMs or two HMM states based on the FA perspective. This measure is applied for an adaptive HMM topology. The HMM state parameters including mixture weight, mean vector and covariance matrix are considered.

3.1 Factor analyzed similarity measure

Assuming that two HMM states s_a and s_b have the same number of Gaussians with diagonal covariance matrices. The state parameters are $\lambda_i = \{\omega_{ik}, \mu_{ikd}, \sigma_{ikd}^2\}, i = a, b, 1 \le k \le K$ $1 \le d \le D$. We integrate all parameters of a Gaussian density into a $(2D+1) \times 1$ supervector. As illustrated in Figure 1, we treat each supervector as an observation, and construct a $2K \times (2D+1)$ data matrix Y_λ where each row is formed by a $(2D+1)$-dimensional supervector. There are $2K$ rows due to the parameters of s_a and s_b. To identify similar Gaussians from two HMM states, we estimate the common factors from data matrix Y_λ. To analyze the correlations among Gaussian components, we alternatively view this matrix as $2D+1$ observations. Each observation contains $2K$ features. Similar Gaussians are represented by the same common factor. In an extreme case, if two states are very similar, it means that K common factors are sufficient to reconstruct $Y_\lambda = [\mathbf{y}_{\lambda 1} \cdots \mathbf{y}_{\lambda(2D+1)}]$. Using FA model, each $2K$-dimensional observation vector $\mathbf{y}_{\lambda d}$ is represented by

$$\mathbf{y}_{\lambda d} = \boldsymbol{\mu}_\lambda + \mathbf{W}_f \mathbf{f}_d + \boldsymbol{\varepsilon}_d \quad (8)$$

where $\boldsymbol{\mu}_\lambda$ is an $2K \times 1$ mean vector calculated from columns of Y_λ and $\{\mathbf{f}_d, \boldsymbol{\varepsilon}_d\}$ are the common factors and specific factors. Therefore, using the FA parameters, we can reconstruct the data matrix Y_f corresponding to common factors. The column vectors are expressed by

$$\mathbf{y}_{fd} = \boldsymbol{\mu}_\lambda + \mathbf{W}_f \mathbf{f}_d. \quad (9)$$

The similarity between HMM states turn out to evaluate the similarity between original data matrix Y_λ and the reconstructed data matrix Y_f. If two states are similar, Y_f shall be close to Y_λ. Moreover, the entries of specific factor should be close to zero. This implies that each common factor sufficiently represents two similar supervectors as shown in Figure 1.
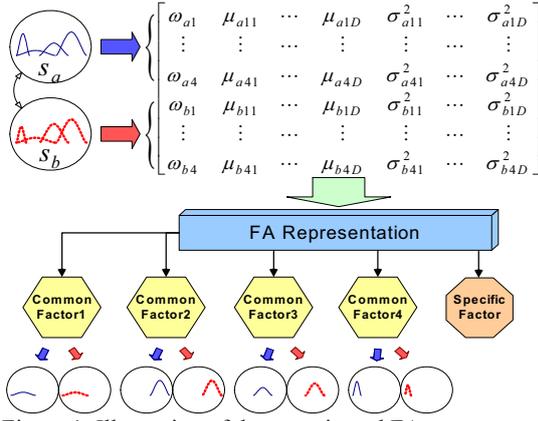
Figure 1: Illustration of data matrix and FA process.

Without loss of generality, we set the number of common factors the same as the number of mixture components, i.e. $M = K$. If the numbers of mixture components in two GMMs are different, we can specify the number of common factors as $M = \max\{K_a, K_b\}$. This kind of setup tends to be reasonable since a new HMM state with less mixture components is less likely different from the existing HMM state with more mixture components. We choose more common factors in this case, and so the new HMM state is harder to be generated. Here, we perform a *composite hypothesis test* to judge the similarity due to the effects of common factors and specific factors. In case of common factors, we define the null and alternative hypotheses by $H_0 : \mathbf{\mu}_\lambda = \mathbf{\mu}_f$ and $H_1 : \mathbf{\mu}_\lambda \neq \mathbf{\mu}_f$, respectively. Notably, we test the similarity of two HMM states by evaluating the equivalence of the mean vectors $\mathbf{\mu}_\lambda$ and $\mathbf{\mu}_f$ of the original matrix Y_λ and the reconstructed matrix Y_f, respectively. The $2K \times 1$ mean vector $\mathbf{\mu}_f$ is calculated from columns of Y_f. Following the Gaussian assumption, the likelihood function under H_0 ($\mathbf{\mu}_\lambda = \mathbf{\mu}_f = \mathbf{\mu}$) is expressed by

$$L_f(H_0) = (2\pi)^{-2K(2D+1)} \left| \Sigma_{H_0} \right|^{-(2D+1)}$$
$$\times \exp\left\{ -\frac{1}{2} \sum_{i=\lambda, f} \sum_{d=1}^{2D+1} (\mathbf{y}_{id} - \mathbf{\mu})^T \Sigma_{H_0}^{-1} (\mathbf{y}_{id} - \mathbf{\mu}) \right\} \quad (10)$$

$$= (2\pi)^{-2K(2D+1)} \left| \Sigma_{H_0} \right|^{-(2D+1)} \exp\{-(2D+1) \cdot \mathrm{tr}[\Sigma_{H_0} \Sigma_{H_0}^{-1}]\}$$

where $\quad \Sigma_{H_0} = [2(2D+1)]^{-1} \sum_{i=\lambda, f} \sum_{d=1}^{2D+1} (\mathbf{y}_{id} - \mathbf{\mu})(\mathbf{y}_{id} - \mathbf{\mu})^T \quad$.

Similarly, the likelihood function under H_1 is obtained by

$$L_f(H_1) = (2\pi)^{-2K(2D+1)} \left| \Sigma_{H_1} \right|^{-(2D+1)}$$
$$\times \exp\left\{ -\frac{1}{2} \sum_{i=\lambda, f} \sum_{d=1}^{(2D+1)} (\mathbf{y}_{id} - \mathbf{\mu}_i)^T \Sigma_{H_1}^{-1} (\mathbf{y}_{id} - \mathbf{\mu}_i) \right\} \quad (11)$$

$$= (2\pi)^{-2K(2D+1)} \left| \Sigma_{H_1} \right|^{-(2D+1)} \exp\{-(2D+1) \cdot \mathrm{tr}[\Sigma_{H_1} \Sigma_{H_1}^{-1}]\}$$

where $\Sigma_{H_1} = (\Sigma_\lambda + \Sigma_f)/2$, Σ_λ and Σ_f denote the $2K \times 2K$ covariance matrices estimated from $\{\mathbf{y}_{\lambda d}\}_{d=1}^{2D+1}$ and $\{\mathbf{y}_{fd}\}_{d=1}^{2D+1}$, respectively. The *test statistic* is computed as a likelihood ratio

$$q_f = \frac{L_f(H_0)}{L_f(H_1)} = \frac{\left| \Sigma_{H_1} \right|^{2D+1}}{\left| \Sigma_{H_0} \right|^{2D+1}} \,. \quad (12)$$

The negative logarithm of test statistic $-2\log q_f$ asymptotically approaches to a chi-square distribution with $2K$ degrees of freedom [1]. Given a significance level α, two HMM states are judged to be significantly different if $-2\log q_f \leq \chi^2_{2K,\alpha}$. For this case, the variations are sufficient to justify the generation of a new HMM state from the existing acoustic models.

Additionally, if two GMMs are similar, the common factors $\mathbf{f}_d$ are sufficient to recover the original data matrix Y_λ. The resulting specific factors $\mathbf{\varepsilon}_d$ shall be very small. Due to this property, we conduct a hypothesis test to evaluate the difference between the values in specific factor vector $\mathbf{\varepsilon}_d$ and the value 0 rather than evaluating their similarity. The null and alternative hypotheses are defined as $H_0 : \mathbf{\mu}_\varepsilon = 0$ and $H_1 : \mathbf{\mu}_\varepsilon \neq 0$, respectively, where $\mathbf{\mu}_\varepsilon$ is the mean vector calculated from the residual signals $\{\mathbf{\varepsilon}_d\}_{d=1}^{2D+1}$. The likelihood functions under H_0 and H_1 are the same as (11) and (10), respectively, except that reconstructed samples $\{\mathbf{y}_{rd}\}_{d=1}^{2D+1}$ of specific factors are used rather than using those of common factors. Hence, the test statistic is defined as $q_r = |\Sigma_{H_0}|^{2D+1} \big/ |\Sigma_{H_1}|^{2D+1}$ and can be shown to asymptotically approach to χ^2_{2K}. From these two test statistics, we construct a composite hypothesis test as $H_0 : \mathbf{\mu}_\lambda = \mathbf{\mu}_f, \mathbf{\mu}_\varepsilon = 0 \quad$ and $\quad H_1 : \mathbf{\mu}_\lambda \neq \mathbf{\mu}_f, \mathbf{\mu}_\varepsilon \neq 0 \quad$. The negative logarithm of the test statistic $q = q_f q_r$ of the composite hypothesis shall asymptotically approach to a chi-square distribution χ^2_{4K} with $4K$ degrees of freedom.

3.2 Adaptive HMM topology

Generally, new data collection contains unknown speaking variations that the existing HMMs do not cover. The HMM topology should be adapted to compensate these variations for speech recognition. To capture the characteristics of new coming data and preserve the goodness of existing HMMs, we present the learning algorithm for an adaptive HMM topology. Starting from an existing set of HMMs, the first step in learning procedure is to train the HMMs from new data collection and conduct the evaluation of similarities between original models and new models. This work focuses on the evaluation of similarity between two GMMs or HMM states. The evaluation was done in phone level. If the measure shows that two states in different HMMs are dissimilar, a new Markov state is produced. If two GMMs are judged to be similar, we will further check the similarity between two Gaussians in state level by using the hypothesis test approach introduced in [10].

In the case of a three-state HMM, we will produce $2^3 = 8$ possible topologies as shown in Figure 2 through conducting the similarity test of HMM states. In this figure, dark node shows the merged state from the original state and the new state. The white node denotes the original state either $s_{a,\bullet}$ or $s_{b,\bullet}$. We model different kinds of variations for an HMM. The similarity between states is measured by using the KLD or the proposed FA measure in the experiments. After determining the HMM topology, Bayesian learning strategy is used to incrementally update the HMM parameters and their hyperparameters. Readers may refer [10] for details.
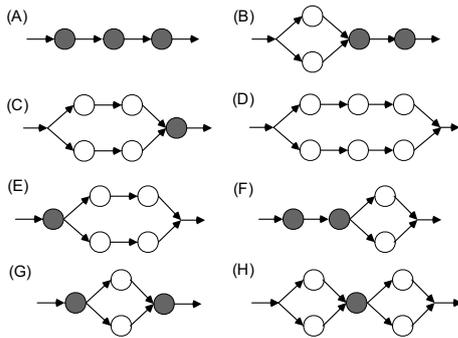
Figure 2: Eight topologies in a three-state HMM.

4. Experiments

4.1 Experimental Setup

In the experiments, we evaluated the proposed FA similarity for continuous speech recognition using Wall Street Journal (WSJ) corpus and reported the word accuracy on using WSJ0. There were 7239 training utterances from 84 speakers (SI84). The test set consisted of 330 utterances from 8 speakers. We used the 5K non-verbalized punctuation, closed vocabulary and trained the trigram language model. The 39 dimensional MFCC feature vectors were extracted. We constructed 41 models (39 monophone models, one silence model, and one short pause model) in initial monophone HMM using TIMIT corpus. We then obtained the standard triphone HMM [11] using WSJ. In the baseline system, each triphone was represented by a three-state HMM with 8 or 16 mixture components. Significance level α is set to be 0.15 in phone level and 0.05 in state level.

To assess the performance of adaptive HMM topology, we divided the training utterances into four clusters by K-means algorithm, and performed four epochs of topology learning. In the procedure, we first trained 84 speaker-dependent HMMs with only one Gaussian in HMM states. We then constructed the speaker-dependent supervectors containing all HMM mean vectors. Having 84 supervectors, K-means clustering was performed to obtain four gender-dependent speaker clusters for adaptive learning. The numbers of speakers in two male clusters and two female clusters are 12, 31, 24, and 17, respectively. In this work, we fixed the order of leaning: the first male cluster, the first female cluster, the second male cluster, and the second female cluster. The hyperparameters were updated and the HMM topology was adapted to meet new pronunciation variations due to speakers and genders.

4.2 Recognition Results

Table 1 shows the word accuracies of standard HMM and adaptive HMM topologies using KLD and FA similarity measures, which are denoted by AHMM-KLD and AHMM-FA, respectively. The approximation of KLD between two GMMs was performed according to (2), (4) and (5). In standard HMM, different learning epochs were considered by using different amount of training data for ML batch training with fixed topology. The proposed AHMM-FA conducted the incremental training with adaptive topology. Both AHMM methods started from the baseline system with 8 mixture components (denoted by HMM-8). The models in single epoch were trained with eight components in HMM states. For comparison, we also carried out the standard HMM with 16 mixture components (denoted by HMM-16). The model size of AHMM-KLD was comparable with that of AHMM-FA. AHMM-KLD and AHMM-FA outperformed HMM-8 in the second, third and fourth learning epochs. AHMM-FA obtained additional improvement compared to

HMM-16 even in the case that the model size of AHMM-FA was smaller than HMM-16. In the fourth epoch, HMM-FA achieved 10.6% and 9.6% word error rate reductions over HMM-8 and HMM-16, respectively. Additionally, we implemented the case (D) of Figure 2 for each learning epoch. After four learning epochs, we obtained a HMM with four times of model parameters of HMM-8. This HMM had a word accuracy of 76.31%. From the experimental analysis, we found that AHMM-KLD more likely detected high similarity between two GMMs since only one Gaussian was approximated in KLD. This phenomenon resulted in the circumstance that the increased parameters of AHMM-KLD mostly came from newly added mixture components rather than newly added states.

Table 1: Word accuracies (%) at different epochs

	Epoch1	Epoch2	Epoch3	Epoch4
HMM-8	53.69	70.8	85.5	93.33
HMM-16	56.75	71.68	85.67	93.41
AHMM-KLD	53.69	71.23	85.69	93.61
AHMM-FA	53.69	72.17	85.75	94.04

5. Conclusions

This paper presented a factor analyzed similarity measure between two GMMs for adaptive HMM topology learning. The computation of this measurement was efficient because there was no need of calculating the KLD of each pair of Gaussian densities. We applied this measure for adaptive HMM topology learning and determined whether a new GMM or HMM state should be added in the current HMM structure. The experiments on WSJ corpus showed that the results using the proposed FA similarity measure performed better than KLD similarity measure. In the future, we are integrating the temporal variations into the HMM topology learning and extending the proposed method for similarity measure in HMM level. We will also apply FA measure for other pattern recognition applications.

6. References

[1] T.W. Anderson, *Introduction to Multivariate Statistical Analysis*, 2nd ed. New York: Wiley, 1984.

[2] J.-T. Chien and C.-W. Ting, "Factor analyzed subspace modeling and selection", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 239-248, 2008.

[3] J.-T. Chien and C.-W. Ting, "Acoustic factor analysis for streamed hidden Markov modeling", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 5, July 2009.

[4] S. Furao and O. Hasegawa, "An incremental network for on-line unsupervised classification and topology learning", *Neural Network*, vol. 19, pp. 90-106, 2006.

[5] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models", in *Proc. of ICASSP*, pp. 317-320, 2007.

[6] M.-Y. Hwang and X. Huang, "Dynamically configurable acoustic models for speech recognition", in *Proc. of ICASSP*, pp. 669-672, 1998.

[7] M. Ostendorf and H. Singer, "HMM topology design using maximum likelihood successive state splitting", *Computer Speech and Language*, vol. 11, pp. 17-41, 1997.

[8] A.-V. I. Rosti and M. J. F. Gales, "Factor analysed hidden Markov models for speech recognition", *Computer Speech and Language*, vol. 18, no. 2, pp. 181-200, 2004.

[9] M. S. Srivastava, *Methods of Multivariate Statistics*, John Wiley & Sons, 2002.

[10] C.-W. Ting, K.-Y. Lee, and J.-T. Chien, "Adaptive HMM topology for speech recognition", in *Proc. of INTERSPEECH*, pp.1237-1240, 2008.

[11] K. Vertanen, "Baseline WSJ acoustic models for HTK and SPHINX: training recipes and recognition experiments", *Technical Report*, Cavendish Laboratory, 2006.