Speech Feature Analysis Using Variational Bayesian PCA

Oh-Wook Kwon*, Kwokleung Chan, Te-Won Lee

***Abstract*—**

In most hidden Markov based automatic speech recognition systems one of the fundamental question is to determine the intrinsic speech feature dimensionality and the number of clusters used in the Gaussian mixture model. We analyzed mel-frequency band energies using a variational Bayesian principal component analysis method to estimate the feature dimensionality as well as the number of Gaussian mixtures by learning a maximum lower bound of the evidence instead of maximizing the likelihood function as used in conventional speech recognition systems. In analyzing the TIMIT speech data set, our method revealed the intrinsic structures of vowels and consonants. The usefulness of this method is demonstrated in the superior classification performance for the most difficult phonemes /b/, /d/ and /g/.

***Index Terms*— Speech analysis, phoneme classification, speech recognition.**

I. Introduction

STANDARD modern speech recognizers are mostly based on mel-scaled cepstral feature vectors and hidden Markov models (HMMs) with continuous observation probability distributions. The feature vectors are obtained by an orthogonal transformation which aims to reduce the dimension and produce coefficients as uncorrelated as possible. The probability distributions are often modeled by mixtures of Gaussian probability distributions with diagonal covariance matrices and are usually trained by the expectation-maximization (EM) algorithm. The number of cepstral features as well as the number of clusters is usually empirically determined and there has been no rigorous way to determine the appropriate feature dimension and the number of mixtures. This is due to the over-fitting problem in maximum likelihood estimation where increasing the cluster number or the dimensionality of the feature vector will always increase the likelihood, in other words, the parameters are often over-fitted to training data when too many Gaussian mixtures are used for acoustic modeling. In this paper, we address this problem by maximizing the evidence of the data and analyzing the intrinsic structure of speech signals that is used to estimate the appropriate feature dimension and the proper number of mixtures with full covariance matrices for each subunit. In contrast with the likelihood that is defined as the probability of the data given the parameters, the evidence is defined as the probability of the data integrated over the distribution of the parameters.

In the conventional setting, a discrete cosine transform (DCT) has been widely used to extract mel-frequency cepstral coefficients (MFCCs). Recently the principal component analysis (PCA) and linear discriminant analysis (LDA) were studied to optimize the transformation and reduce the feature dimension in case multiple frames are used for acoustic modeling [1][2][3]. The transformation replaces finite-impulse response (FIR) filters in the temporal axis. Here, we use mixtures of variational Bayesian principal component analyzers (VBPCA) to analyze mel-frequency band energies and obtain proper transformations. In the VBPCA, parameters are represented by probability distributions and the parameters of the parameter distributions (hyper-parameters) are estimated by maximizing the evidence of observed signals instead of the likelihood in the conventional maximum likelihood estimation paradigm. Hidden variables are introduced so that the evidence is approximated by the lower bound and posterior distributions of the parameters are assumed separable. Prior distribution on the parameters acts as a penalty term and gives an appropriate number of mixtures and the feature dimension and prevents the parameters from over-learning. We analyze the TIMIT speech database and reveal some characteristics for vowels and consonants. We also demonstrate the usefulness of the VBPCA for speech recognition by performing phoneme classification. Fig.1 compares the standard MFCC-based speech recognizer and the assumed system in this paper.

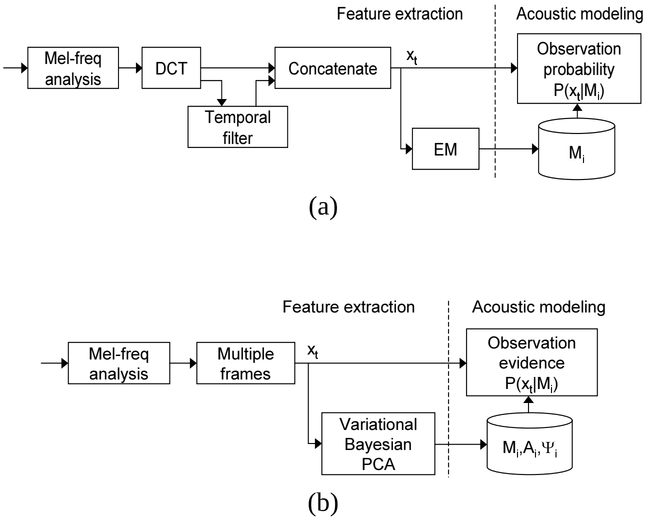


Fig. 1. Feature extraction and acoustic modeling (a) in a standard HMM-based system and (b) in the assumed system in the paper.

II. Variational Bayesian PCA

In the probabilistic PCA (PPCA) [4][5], an N dimensional observed data vector $\mathbf{x}$ is assumed to be generated from in-

*Corresponding author. E-mail: owkwon@ucsd.edu; kwchan@ucsd.edu; tewon@ucsd.edu dependent latent variables $\mathbf{s}$; $\mathbf{x} = \mathbf{A}\mathbf{s} + \nu + \epsilon$ where $\mathbf{s}$ is a zero-mean unit-variance Gaussian random variable, ν is the mean vector and ϵ is a zero-mean Gaussian noise vector with an isotropic covariance $[\mathbf{\Psi}^c]^{-1}$. The PPCA finds $\mathbf{A}$, $\mathbf{\Psi}^c$ and ν which maximize the likelihood of the observed data vectors. The solution can be obtained by eigenvector decomposition of a sample covariance matrix. In factor analysis similar to PCA, a diagonal covariance is assumed for the noise model. Comparison between factor analysis and PCA can be found in [5][6]. However, PPCA does not give the optimal number of independent components because it uses a maximum likelihood criterion, thus leading to the over-fitting problem.

Bayesian PCA [7][8] was proposed to find the intrinsic dimension and the optimal number of clusters in the latent variable model by utilizing prior information on parameters θ. Following the Bayesian framework, Bayesian PCA computes the evidence of the data $P(\mathbf{X})$ instead of the likelihood $P(\mathbf{X}|\theta)$. Direct computation of the log evidence is difficult. We therefore follow the variational method, where the lower bound of the log evidence is obtained by using the Jensen's inequality and then the lower bound is maximized through functional maximization.

Observation data vectors $\mathbf{X} = \{\mathbf{x}_t \in \mathcal{R}^N\}, t = 1, ..., T$, are assumed to be generated from one of C clusters with probability ρ^c. Each cluster is centered at ν^c, has covariance matrix $\mathbf{A}^c\mathbf{A}^{cT}$ and diagonal Gaussian noise with inverse covariance $\mathbf{\Psi}^c = \Psi^c\mathbf{I}$.

$$P(\mathbf{x}_t|\rho^c, \mathbf{A}^c, \nu^c, \mathbf{\Psi}^c) = \sum_{c=1}^{C} \rho^c \int \mathcal{N}(\mathbf{x}_t|\mathbf{A}^c\mathbf{s}_t^c + \nu^c, \mathbf{\Psi}^c) P(\mathbf{s}_t^c) d\mathbf{s}_t^c \tag{1}$$

We follow the notation used in [10] whose learning algorithm with the condition $K = 1$ exactly matches [9]. In each cluster the observation is a linear combination of M independent Gaussian sources s_t^c; $P(s_{tm}^c) = \mathcal{N}(s_{tm}^c|0, 1)$. In Bayesian PCA, the maximum number of sources is restricted to $N - 1$. We also assume zero mean Gaussian density for $\mathbf{A}_n^c$; $P(A_{nm}^c) = \mathcal{N}(A_{nm}^c|0, \alpha_m^c)$. Instead of the likelihood of the data $P(\mathbf{X}|\theta)$, we maximize an approximated lower bound of the log evidence on the data $P(\mathbf{X})$. Assuming the $\mathbf{x}_t$ are independent and identically distributed, introducing posterior probability distribution $Q(\theta)$ and using the Jensen's inequality, log of the evidence is lower bounded by [10]

$$\log P(\mathbf{X}) = \log \int \prod_t P(\mathbf{x}_t|\theta) P(\theta) d\theta \tag{2}$$

$$\log P(\mathbf{X}) \geq \int Q(\theta) \sum_t \log P(\mathbf{x}_t|\theta) d\theta + \int Q(\theta) \log \frac{P(\theta)}{Q(\theta)} d\theta. \tag{3}$$

The learning algorithm to maximize the lower bound can be found in [9][10]. To compare between models with a different number of clusters, we compute the lower bound of the log evidence given as (the right side in (3))

$$F(\mathbf{X}) = \sum_t \log Z_t + \int Q(\theta) \log \frac{P(\theta)}{Q(\theta)} d\theta \tag{4}$$

where Z_t is the normalization constant for $Q(c_t)$ in [10] and corresponds to c in [9].

For classification, the log evidence of a new test vector $\mathbf{x}_{T+1}$ given a model M_i was approximated as

$$\begin{aligned} \log P(\mathbf{x}_{T+1}|\mathbf{X}, M_i) &= \log \int P(\mathbf{x}_{T+1}|\theta) Q(\theta) d\theta && (5) \\ &\approx \log Z_{T+1}. && (6) \end{aligned}$$

To obtain Z_{T+1}, we estimated $Q(\mathbf{s}_{T+1})$ and computed $Q(c_{T+1})$ over the posterior distribution $Q(\theta)$ fixed after learning. Note that when we evaluate the log evidence of $\mathbf{x}_{T+1}$, the learned $Q(\theta)$ was used to compute Z_{T+1} while the computation of $F(\mathbf{X}, \mathbf{x}_{T+1})$ needs a different $Q'(\theta)$ estimated from $(\mathbf{X}, \mathbf{x}_{T+1})$. The above approximation is valid when the number of samples is moderately enough. For a model learned from a very small number of samples (less than about 10), the log evidence should be computed by integrating likelihood over the learned posterior distribution $Q(\theta)$ as above.

The following priors in the parameters and hyper-parameters are assumed in deriving the lower bound [10]:

$$\begin{aligned} \alpha_m^c &\sim \mathcal{G}(a_0(\alpha_m^c), b_0(\alpha_m^c)) \\ \nu_n^c &\sim \mathcal{N}(\mu_0(\nu_n^c), \Lambda_0(\nu_n^c)) \\ \Psi^c &\sim \mathcal{G}(a_0(\Psi^c), b_0(\Psi^c)) \\ (\rho^1, ..., \rho^C) &\sim \mathcal{D}(d_0(\rho^1), ..., d_0(\rho^C)) \end{aligned} \tag{7}$$

where $\mathcal{G}(.)$, $\mathcal{N}(.)$ and $\mathcal{D}(.)$ are the gamma, normal and Dirichlet distribution, respectively, and μ_0, Λ_0, a_0, b_0 and d_0 are hyper-parameters of the corresponding priors. We used $\mu_0(\nu_n^c) = 0$, $\Lambda_0(\nu_n^c) = 0.001$, $d_0(\rho^c) = 1$, $a_0(\alpha_m^c) = 1.001$, $b_0(\alpha_m^c) = 0.001$, $a_0(\Psi^c) = 1.001$ and varied $b_0(\Psi^c)$ depending on applications. The hyper-parameters were decided so that the resulting priors have a distribution as dispersive (non-informative) as possible. The $b_0(\Psi^c)$ was determined to satisfy both the non-informative property and the assumed noise level. We used the same value of the $b_0(\Psi^c)$ for every phoneme. The hyper-parameter a_0 for a gamma distribution must be larger than unity in order to compute the expectation of $1/X$. We note that $E[X] = b_0/a_0$, $var[X] = b_0/a_0^2$, and $E[1/X] = b_0/(a_0 - 1), a_0 > 1$ for a gamma-distributed random variable X.

The input signals were normalized to have zero mean and unit variance in the preprocessing stage. If the number of samples in a cluster was less than a threshold times the number of samples divided by the number of the current clusters, or the maximum vector norm of the columns in $\mathbf{A}^c$ was less than a second threshold, the cluster was removed. In this work, we used 0.01 and 1×10^{-10} for the thresholds, respectively. The lower the thresholds, the longer iteration it takes for the cluster with a small weight to be removed. Once the cluster is removed, the lower bound increases.

To create a new cluster, we selected a cluster randomly with the probability proportional to the contribution to the likelihood. The selected cluster was split into two clusters whose mean vectors are displaced in the positive and negative direction of the column vector with the maximum vector norm of $\mathbf{A}^c$ and then $\mathbf{A}^c$'s were randomly perturbed around 0.9 times the original matrix by adding 0.1 times square root of the maximum vector norm times a random matrix whose elements were generated by a unit-variance Gaussian random number generator. If the new cluster increased the lower bound of the log evidence, it was included in the cluster set. Initially, the cluster set included a predefined number of clusters same as the maximum number of clusters used for maximum likelihood clustering. If the cluster set was not changed in the predefined number of clustering epochs, the learning procedure was terminated.

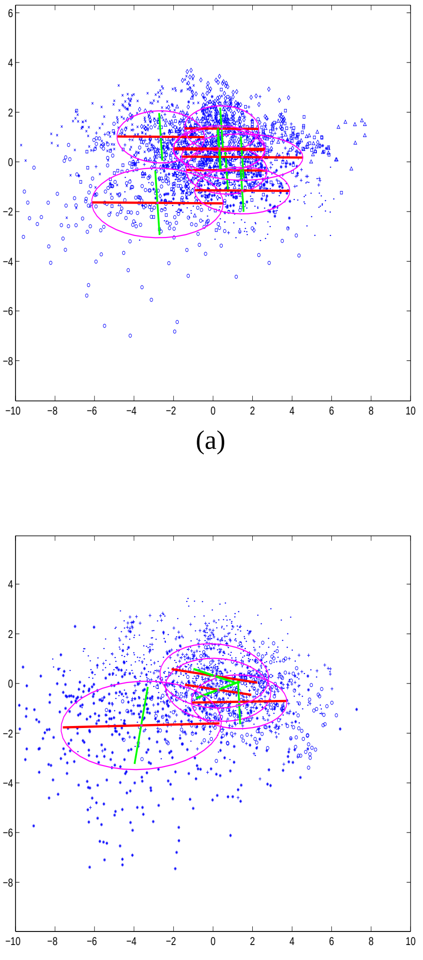
The final dimension of the clusters depends on the hyper-parameter $b_0(\Psi^c)$ of the noise inverse covariance. A larger value yields a smaller dimension. We control these hyper-parameters to obtain reduced dimensions. With a small value in a noninformative prior case, the dimension is usually the same as the original dimension minus one because there are many samples for each cluster.

III. Experimental Results

We used the TIMIT speech database in the experiment and extracted 22 frequency band energies from the label information of the training set by using the HTK [11]. The frame shift was 10 ms and the window was 20 ms. We used the 48 phoneme set and 10,000 frames of speech data for each phoneme were used to reduce the computation in the learning algorithm.

We analyzed real speech signals using the TIMIT database. Fig. 2 shows the clustering results when the EM algorithm and the variational Bayesian PCA were used for the phoneme /aa/. The number of samples in a cluster was limited to 250 and 500 in each case to give a clear look, which makes the two sample sets seem not the same. The sample data and the obtained components were projected into a space spanned by the principal components in the single cluster case. In the figures, the ellipses denote the contour of unit variance, two lines in a cluster denote the two major components and the cross point of the lines is the mean of the cluster. Note that the principal components in the single cluster case are orthogonal. In the EM algorithm, we used 8 Gaussian mixtures with diagonal covariance matrices. The resulting Gaussians partitioned the signal space into small regions. Comparatively, the Bayesian PCA with 8 initial clusters resulted in 4 Gaussians with full covariance matrices. The intrinsic dimension was 20, 18, 7, and 4 and the cluster probability was with $\rho^c = 0.56, 0.37, 0.05, 0.02$ for each cluster respectively.

Fig. 3 shows the number of clusters and the minimum/maximum/average dimension of the intrinsic components for the 48 phonemes in the TIMIT database. The bar denotes the number of clusters, the lower/upper end point of the thick line denotes the minimum/maximum dimension, and the square on the thick line denotes the average dimension. We set the hyper-parameter of the inverse noise covariance to $b_0(\Psi^c) = 0.5$ in the experiment. The average dimension was computed by arithmetic mean of the dimension of every cluster. The results showed that vowels were clustered into 2 to 4 Gaussians whereas most consonants had 1 or 2 clusters. The average dimension of the latent signals was 18.1. When all speech signals were used, the Bayesian PCA yielded 7 clusters and the average dimension of 19.7 (the ALL case in the figure). The results can be exploited in determining the number of Gaussians in the conventional continuous HMM-based systems. The average dimension of the latent signals for each phoneme showed a small variance.



(a)

(b)

Fig. 2. Clustering results for the phoneme /aa/ with (a) the EM algorithm with diagonal covariance matrices and (b) the VBPCA. Due to the additional flexibility in adapting to the data density the VBPCA captures the density structure more accurately.

Using one of the most confusing phoneme pairs /b/, /d/ and /g/, we performed phoneme classification experiments to test the usefulness of the method. We extracted the speech data for the three phonemes from the independent core test set of the TIMIT database. Table I compares the frame accuracy of the conventional EM algorithm and the VBPCA. The 'D/F' denotes 'diagonal/full' covariance matrix and 'M' denotes the number of clusters for each phoneme. In the EM cases, the frequency band energies were transformed to 12 MFCCs. In the VBPCA, the /b/, /d/ and /g/ phonemes had all 1 cluster and the intrinsic dimension was 18, 18, 16, respectively. Columns of the $\mathbf{A}^c$ with the norm 0.01 times less than the maximum norm were set to zero to reduce the dimension. The VBPCA outperformed the EM algorithm with 16 Gaussian distributions and with a single full covariance Gaussian (F12). With diagonal covariance Gaussians increased to 32, we obtained worse accuracy. The VBPCA produced the higher accuracy than using a single Gaussian with the full covariance matrix without dimension reduction (F22), which implies that it effectively learns covari-

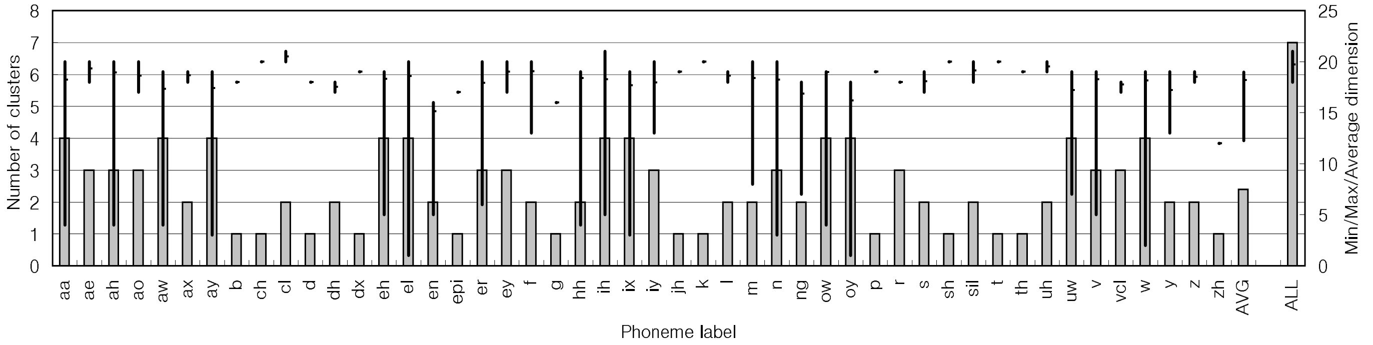


Fig. 3. The number of identified clusters and the minimum/maximum/average dimension for each phoneme. The AVG denotes the average statistics over all phonemes and the ALL denotes the case in which whole speech signals were input to the algorithm.

TABLE I

AVERAGE FRAME ACCURACY (%) FOR /B/, /D/ AND /G/ PHONEME CLASSIFICATION USING THE EM ALGORITHM AND THE VBPCA WITH VARYING COVARIANCE KIND AND NUMBER OF MIXTURES

Algorithm	EM							VBPCA
Cov(dim,M)	D(12,1)	D(12,2)	D(12,4)	D(12,8)	D(12,16)	F(12,1)	F(22,1)	F(*,*)
Frame Acc.	34.0	48.7	49.8	51.9	51.9	49.5	58.7	59.6

ance parameters by avoiding over-fitting.

IV. CONCLUSIONS

We applied VBPCA to determine an appropriate feature dimension and the number of mixtures for each subunit in a standard HMM-based system with continuous observation probability distributions. By using VBPCA, we were able to use different transformations with different feature dimension for each phoneme. Hyper-parameters and the number of mixtures were obtained by maximizing a lower bound of the evidence. The feature analysis results showed that the appropriate number of clusters with full covariance matrices is 2 to 4 for vowels and 1 to 2 for consonants. With a minimally informative prior, the feature dimension of each phoneme was shown to be about 18, which is a little larger than the standard dimension of the MFCC feature. Phoneme classification experiments with confusing phonemes showed that the VBPCA yielded frame accuracy higher than the standard MFCC feature and the single Gaussian with a full covariance matrix.

Future research will extend this paradigm for multi-frame input signals in order to optimize the temporal filters in feature extraction and perform speaker independent phoneme recognition over the complete phoneme set by embedding the models into a standard HMM-based system.

REFERENCES

[1] S.S. Kajarekar, B. Yegnanarayana, H. Hermansky, “A study of two dimensional linear discriminants for ASR,” Proc. ICASSP, 2001.

[2] T. Fukuda, M. Takigawa, T. Nitta, “Peripheral features for HMM-based speech recognition,” Proc. ICASSP, 2001.

[3] R. Gemello, D. Albesano, L. Moisa, R.D. Mori, “Integration of fixed and multiple resolution analysis in a speech recognition system,” Proc. ICASSP, 2001.

[4] M.E. Tipping, C.M. Bishop, “Probabilistic principal component analysis,” J. Royal Statistical Society, Series B, 61, part 3, pp. 611-622, 1999.

[5] M.E. Tipping, C.M. Bishop, “Mixtures of probabilistic principal component analysers,” Neural computation, 1998.

[6] S. Roweis, Z. Ghahramani, “A unifying review of linear Gaussian models,” Neural computation 11, 1999.

[7] C.M. Bishop, “Variational principal components,” Proc. ICANN, 1999.

[8] C.M. Bishop, “Bayesian PCA,” Proc. NIPS, 1998.

[9] Z. Ghahramani and M.J. Beal, “Variational inference for Bayesian mixtures of factor analysers,” Proc. NIPS, 1999.

[10] K. Chan, T.-W. Lee, T. Sejnowski, “Variational learning of clusters of undercomplete nonsymmetric independent components,” Journal of Machine Learning Research, vol. 3, pp. 99–114, 2002.

[11] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, *The HTK Book,* Microsoft Corp., 2000.