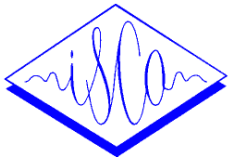

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

A GENERALIZED COMMON VECTOR APPROACH FOR ROBUST SPEAKER INDEPENDENT AUTOMATIC SPEECH RECOGNITION

Der-Jenq Liu and

Department of Electrical and Control Engineering,
National Chiao-Tung University, Hsinchu
william@falcon3.cn.nctu.edu.tw

Chin-Teng Lin

Department of Electrical and Control Engineering,
National Chiao-Tung University, Hsinchu
ctlin@fnn.cn.nctu.edu.tw

ABSTRACT

A new technique is proposed to estimate the robust continuous observation densities of hidden Markov model (HMM) for improving the performance of speaker-independent (SI) automatic speech recognition system. First, a scheme of generalized common vector (GCV), which originated from the common vector approach (CVA), is proposed. The objective of this scheme is to extract a robust speech feature over different speakers. That is, we attempt to obtain a common feature to represent an invariant characteristic over many speakers. Then, based on this scheme, we construct a GCV-based HMM (GCVHMM). An element to extract GCV is integrated into HMM. A re-estimation algorithm for the parameters of GCVHMM is also derived.

1. INTRODUCTION

One of the most important issues of speaker-independent (SI) speech recognition system is the estimation of robust speech model over different speakers. The statistical speech models for each phone unit of the recognition system should be estimated to cover the spectral variations in speech signals caused by intra-speaker differences. The simple way of training a SI recognition system is to use the speech data collected from many speakers to train speech model. It has been reported, however, that the performance such trained models is less effective than that of a speaker-dependent models for individuals. When a speaker-dependent (SD) system trained on speech from a given speaker is tested on other speech data from the same speaker, the error rate may be as low as a half to a third that of a similar SI continuous speech recognition system tested on the same data. In addition, some other adverse conditions, namely, noise, channel interference, microphone transducer distortion also degrade the performance of a SI speech recognition system.

Common vector approach, which was proposed in [1], is to find a common vector representing the invariant characteristic among a set of feature vectors that are linearly independent. The concept of CVA is to estimate a difference subspace that can be used to represents the feature caused by intra-speaker difference, noise, or channel interference. The common vector is obtained by removing the component belonging to the difference subspace. The common vector has been applied in the speaker-independent isolated word recognition.

In a further study [2], the close relationship between the nonzero principal components and the difference subspace, together with the complementary close relationship between the zero principal components and the common vector were shown. That is, by the eigenanalysis of the covariance matrix of all feature vectors in the training data, the eigenvectors corresponding to nonzero eigenvalues constitute the basis of difference subspace, whereas the eigenvectors corresponding to zero eigenvalues constitute the basis of the common subspace. The common vector of X is in the direction of a linear combination of the eigenvectors corresponding to the zero eignevalues of the covariance matrix.

The assumption in CVA that all the vectors in training data should be linearly independent is, however, impractical. The number of collected vectors in training data will be greater than that of the vector's dimension, which results in that the common vector doesn't exist. A modification of CVA to overcome this problem is required. Based on the concept of the CVA and the eigenanalysis of covariance matrix, we propose a new scheme, called the generalized common vector, to make a generalization of CVA.

The idea is that we decompose the feature space into two mutually orthogonal subspaces, one is the difference subspace, and the other is the common vector subspace. We will show that the eigenvalue indicates the degree of feature variation on the corresponding eigenvector. The less value of eignevalue, the more invariant on the corresponding eigenvector it is. Therefore, the difference subspace is constituted by the eigenvectors corresponding to the eigenvalues with greater value, and the common vector subspace is constituted by the ones corresponding to smaller eigenvalues. Based on the scheme proposed, we construct a GCV-based HMM (GCVHMM). An element to extract GCV is integrated into HMM. A re-estimation algorithm for the parameters of GCVHMM is also derived.

2. COMMON VECTOR APPROACH

In this section, we describe what the common vector is and how to obtain it. Let R^D be $D-$ dimensional

vector space and there be given a set of linearly independent vectors, $X = \{x_m \mid x_m \in R^D,\ 1 \le m \le M_X\}$, where $M_X \le D$. Each x_m can be written as a summation of a common vector x_{com} and a difference vector $x_{m,dif}$:

$$x_m = x_{com} + x_{m,dif},\ 1 \le m \le M_X. \quad (1)$$

The objective of CVA is to find the common vector x_{com} that represents the common properties of the set X, which depends on some criteria with some philosophical sense.

A set X' of difference vectors constituted by subtracting a given reference vector $x_{m'} \in X$ from other vectors in X is defined as $X' = \{x'_m \mid x'_m = x_m - x_{m'},\ m \ne m'\}$. The vector subspace spanned by the set X' is called difference subspace, denoted as $S_{X'}$. It can be shown that all the vectors $x'_m \in S_{X'}$ are also linearly independent, and thus the rank of $S_{X'}$ is $M_X - 1$. Let $Z = \{z_1, z_2, ..., z_{M_X-1}\}$ be an orthonormal set of $S_{X'}$, where Z can be obtained through Gram-Schmidt orthogonalization method in linear algebra.

To obtain a unique solution for the common vector, one may make an assumption that the vector $x_{m,dif}$ is the projection of x_m onto the difference subspace $S_{X'}$, i.e.,

$$x_{m,dif} = \sum_{l=1}^{M_X-1} < x_m, z_l > z_l, \quad (2)$$

where $< x_m, z_l >$ is the inner product of x_m and z_l. Then the common vector x_{com} is obtained by subtracting projections of any vector x_m on the difference subspace $S_{X'}$ from this same vector, i.e., $x_{com} = \tilde{x}_m = x_m - x_{m,dif}$. It can be seen that x_{com} is orthogonal to $x_{m,dif}$; that is, each x_m can be decomposed into two components so that both of them are mutually orthogonal and x_{com} is orthogonal to $S_{X'}$. It has been also shown that the common vector x_{com} is independent from index m', that is, $x_{com} = \tilde{x}_{m''} = \tilde{x}_{m'''}$, with $m'' \ne m'''$. Furthermore, it is also independent from the selected reference vector $x_{m'}$. Therefore, x_{com} is unique for X, so it represents the common property among X.

3. GENERALIZATION OF COMMON VECTOR

In this section, we attempt to make a generalization of CVA based on the insight into the meaning of all the eigenvalues of covariance matrix Φ_X of X.

3.1 Eigenanalysis of covariance matrix

The covariance matrix of X is defined as $\Phi_X = \sum_{m=1}^{M_X} (x_m - x_{avg})(x_m - x_{avg})^T \Big/ M_X$. The pairs of one eigenvalue and its corresponding eigenvector of Φ_X, denoted as λ_l and v_l respectively, satisfy the following equation: $\Phi v_l = \lambda_l v_l$, $1 \le l \le D$. Since Φ is a symmetric matrix, $\Phi_X^T = \Phi_X$, and nonnegative definite, $v^T \Phi_X v \ge 0$, $v \in R^D$, all the eigenvalues are real and nonnegative, and the corresponding eigenvectors are mutually orthogonal.

The variance of projection of $(x_m - x_{avg})$ on v_l, denoted as σ_l^2, is defined by

$$\sigma_l^2 = \sum_{m=1}^{M_X} < x_m - x_{avg}, v_l >^2 \Big/ M_X = v_l^T \Phi_X v_{l=1} = \lambda_l, \quad (3)$$

subject to $\langle v_l, v_l \rangle = 1$. That is, eignevalue λ_l can be interpreted as the variance of projection of $(x_m - x_{avg})$ on its corresponding eigenvector v_l. The less value of eignevalue, the more invariant of the projection of $(x_m - x_{avg})$ on the corresponding eigenvector it is. Thus, the eigenvalue can be used as a measure of the variance on its corresponding eigenvector. It is, therefore, reasonable that the common vector of X is in the direction of a linear combination of the eigenvectors corresponding to the zero eignevalues of the covariance matrix [2].

3.2 Scheme of GCV

Consider a set $X = \{x_m \mid x_m \in R^D,\ 1 \le m \le M_X\}$ of training data, which are not linearly independent. We attempt to find a vector called generalized common vector, x_{gcom} to represent the invariant characteristic. We first decompose the vector space R^D into two mutually orthogonal vector subspaces, one is the difference subspace, S_{dif}, and the other is the common vector subspace, S_{gcom} with rank L. All the vectors in X can be expressed as $x_m = \tilde{x}_m + x_{m,dif}$, where $x_{m,dif} \in S_{dif}$, and $\tilde{x}_m \in S_{gcom}$. It is intuitional to let x_{gcom}

be the average of all $\widetilde{x}_m$ because such an x_{gcom} minimizing the following criterion : $C = \sum_{m=1}^{M_X} \left\| \widetilde{x}_m - x_{gcom} \right\|^2 / M_X$, where $\|\cdot\|$ is the operation of 2 norm.

Next, we want to find what S_{dif} and S_{gcom} are. Suppose $\{\upsilon_1, \upsilon_2, ..., \upsilon_L\}$ is a basis of S_{gcom}. Thus, $\widetilde{x}_m = \sum_{l=1}^{L} \langle x, \upsilon_l \rangle \upsilon_l$, and $x_{gcom} = \sum_{l=1}^{L} \langle x_{avg}, \upsilon_l \rangle \upsilon_l$. Therefore, C can be rewritten as $C = \sum_{m=1}^{M_X} \sum_{l=1}^{L} \langle x - x_{avg}, \upsilon_l \rangle^2 / M_X = \sum_{l=1}^{L} \upsilon_l{}^T \Phi_X \upsilon_l$. To minimize C , we add the constraints $\sum_{l=1}^{L} \kappa(1 - \upsilon_l{}^T \upsilon_l)$, which insure that $\|\upsilon_l\| = 1$, into C to form the Lagrange equation, $C_L = \sum_{l=1}^{L} \upsilon_l{}^T \Phi_X \upsilon_l + \sum_{l=1}^{L} \kappa(1 - \upsilon_l{}^T \upsilon_l)$. By taking partial derivation of C_L with respect to υ_l and setting the result to zero, we obtain that $\Phi_X \upsilon_l = \kappa_l \upsilon_l$. That is, υ_l is an eigenvector of Φ_X and its corresponding eigenvalue, κ_l , is one of the L smallest eigenvalues of Φ_X .

In summary, to obtain the generalized common vector, we make an eigenanalysis of the covariance matrix of X . Then the generalized common vector can be expressed by the L eigenvectors, each of which corresponds to one of the L smallest eigenvalues of Φ_X .

4. GENERALIZED COMMON-VECTOR-BASED HMM (GCVHMM)

In this section, we construct a framework, called generalized common-vector-based HMM for robust speaker-independent speech recognition. The main concept of applying GCV to this task is that any feature vector of one specific spoken word can be decomposed into two mutually orthogonal components, difference vector and common vector. The difference vector represents the spectral variation containing in speech signal of the word, whereas the common vector represents the invariant feature of the word. To obtain a robust speech recognition engine, we integrate the GCV into a HMM to form a so-called GCVHMM. In this section, we shall also derive a method to estimate the parameters of GCVHMM, including the parameters regarding to GCV. The method of estimation is based on Baum-Welch method [3], which iteratively re-estimates the parameters of HMM.

4.1 Structure of GCVHMM

In this study, a N state, left-to-right continuous observation density HMM, denoted as Ω , is considered. The initial probability for state i is denoted by $\delta_i = P(\theta_0 = i),\ 1 \le i \le N$, and the transition probability from state i to state j by $a_{ij} = P(\theta_t = j \mid \theta_{t-1} = i),\ 1 \le i, j \le N$ for $i,\ j = 1, \cdots, N$. Denote $\delta = \{\delta_i\}_{i=1}^{N}$, and $A = \{a_{i,j}\}_{i,j=1}^{N}$. The observation density in state i , $b_i(o_t) = P(o_t \mid \theta_t = i)$, which is assumed to be a mixture of Gaussians, is then given as

$$b_i(o_t) = \sum_{k=1}^{M} c_{i,k} b_{i,k}(o_t), \qquad (4)$$

where $b_{i,k}(o_t)$ is a Gaussian distribution, M is the mixture number, and $c_{i,k}$ is the probability of mixture k in state i .

For each mixture of each state, there is a mechanism of the GCV. Let $\widetilde{o}_{t,i,k}$, which plays the same role as $\widetilde{x}_m$, be extracted by $V_{i,k} o_t$, where $V_{i,k} = [v_{i,k,1},\ ...,\ v_{i,k,L}\ ]^T$. In fact, each $v_{i,k,l}$ is an eigenvector corresponding to one of the L smallest eigenvalues of the covariance matrix, $R_{i,k}$. Assume that $\widetilde{o}_{t,i,k}$ is a random variable of Gaussian distribution with mean $\eta_{i,k}$ that, in fact, plays the same role as x_{gcom} . Or $\eta_{i,k}$ can be expressed as $V_{i,k} \mu_{i,k}$, where $\mu_{i,k}$, which play the same role as x_{avg} , is the average of o_t at mixture k in state i . Therefore,

$$b_{i,k}(o_t) = \frac{\left|\Lambda_{i,k}{}^{-1}\right|}{\sqrt{(2\pi)^L}} \exp\{-\frac{1}{2} z_{t,i,k}{}^T \Lambda_{i,k}{}^{-1} z_{t,i,k}\}, \qquad (5)$$

where $z_{t,i,k} = \widetilde{o}_{t,i,k} - \eta_{i,k}$, and $\Lambda_{i,k}$ is assumed to be diagonal, i.e., $\Lambda_{i,k} = diag[\sigma_{i,k,1}, ..., \sigma_{i,k,L}]$. Thus $\left|\Lambda^{-1}{}_{i,k}\right| = \prod_{l=1}^{L} \sigma^{-1}{}_{i,k,l}$.

Denote $B = \{b_i(\cdot)\}_{i=1}^{N}$ and $\Omega = \{\delta, A, B\}$.

4.1 Re-estimation algorithm

In general, estimation of parameters for HMM is based on the Baum-Welch algorithm [3], [4] (or equivalently the EM (expectation maximization) algorithm). The EM algorithm is a two-step iterative procedure. In the first step, called the expectation step (E step), we compute the

auxiliary function for the equation $Q(\Omega,\Omega') = \sum_{\Theta}\sum_{K} P(O,\Theta,K \mid \Omega)\cdot\log(P(O,\Theta,K \mid \Omega'))$.

where the observation sequence $O = (o_1,\ o_2,\ \cdots,\ o_T)$, the unobserved state sequence $\Theta = (\theta_0,\ \theta_1,\ \cdots,\ \theta_T)$, and the unobserved mixture component sequence $K = (k_1,\ k_2,\ \cdots,\ k_T)$.

In the second step, called the maximization step (M step), we find the value of Ω' that maximizes $Q(\Omega,\Omega')$, i.e.,

$$\overline{\Omega} = \arg\max_{\Omega'} Q(\Omega,\Omega') . \qquad (6)$$

It has been shown that if $Q(\Omega,\Omega') \geq Q(\Omega,\Omega)$, then $P(O \mid \Omega') \geq P(O \mid \Omega)$ [3]. Therefore, iteratively applying the E and M steps of equations guarantees monotonic increase in the likelihood. The iterations are continued until the increase in the likelihood is less than some predetermined threshold. The re-estimates of parameters are listed bellow [5]:

$$\overline{\delta}_i = P(O,\theta_0 = i \mid \Omega)/P(O \mid \Omega),$$

$$\overline{a}_{i,j} = \sum_{t=1}^{T} P(O,\theta_{t-1} = i,\theta_t = j \mid \Omega) \Big/ \sum_{t=1}^{T} P(O,\theta_{t-1} = i \mid \Omega),$$

$$c_{i,k} = \sum_{t=1}^{T} P(O,\theta_t = i,k_t = k \mid \Omega) \Big/ \sum_{t=1}^{T} P(O,\theta_t = i \mid \Omega),$$

$$\overline{\mu}_{i,k} = \sum_{t=1}^{T} P(O,\theta_t = i,k_t = k \mid \Omega)\cdot o_t \Big/ \sum_{t=1}^{T} P(O,\theta_t = i,k_t = k \mid \Omega),$$

$$\overline{\sigma}_{i,k,l} = \sum_{t=1}^{T} P(O,\theta_t = i,k_t = k \mid \Omega)\cdot(z'_{t,i,k,l})^2 \Big/ \sum_{t=1}^{T} P(O,\theta_t = i,k_t = k \mid \Omega)\cdot,$$

$$\overline{V}_{i,k} = [\overline{v}_{i,k,1},\ldots,\overline{v}_{i,k,L}]^T,$$

where

$$R_{i,k}\overline{v}_{i,k,l} = \varepsilon_{i,k,l}\overline{v}_{i,k,l},$$

and

$$R_{i,k} = \sum_{t=1}^{T} P(O,\theta_t = i,k_t = k \mid \Omega)\cdot(o_t - \mu_{i,k})\cdot(o_t - \mu_{i,k})^T.$$

5. EXPERIMENT RESULT AND DISCUSSION

In this section, we shall perform experiments to compare the performance of GCVHMM and conventional HMM. In our experiments, we use a speech database from 20 persons including 10 males and 10 females. The sampling rate of the speech signal in the database is 8kHz. Each one speaks 10 times of each Mandarin digit. We use the spoken digits by 16 persons as the training data, and the rest as the testing data. The features used in this test is 12th order cepstral coefficients.

In Mandarin digits, there are two confusable subsets, one includes 0, 1, and 7; the other includes 6 and 9. The performances of GCVHMM for digits 6 and 9 are shown in Table I. In the experiments, we set the parameter L is 11, that is, we truncate one dimension in the feature space. The performances of conventional HMM for the two digits are shown in Table II. From the result in Tables I and II, we can see that the recognition rates of GCVHMM for digit 6 are better than those of conventional HMM. The performance degradation in conventional HMM caused by the effect of confusion between digits 6 and 9 are enhanced by GCVHMM for SI recognition.

Table I. The recognition rates of GCVHMM for Mandarin digits 6 and 9.

State No.	2	3	4	5	6
Digit 6	92.237	93.151	93.607	90.868	95.434
Digit 9	97.018	97.018	98.165	97.706	98.624

Table II. The recognition rates of conventional HMM for Mandarin digits 6 and 9.

State No.	2	3	4	5	6
Digit 6	82.192	80.137	86.758	92.237	93.151
Digit 9	97.018	95.642	95.642	94.725	95.413

6. CONCLUSIONS

In this study, we first propose a scheme of the generalized common vector that originated from the common vector approach. Then, based on the proposed scheme, we propose a new acoustic model called GCVHMM for improving the recognition robustness in acoustic level in SI recognition. The scheme of generalized common vector in GCVHMM is used to extract the invariant characteristic among different speakers. We perform experiments on Mandarin digits for GCVHMM and conventional HMM. The results show that the performance degradation in conventional HMM caused by the effect of confusion are enhanced by GCVHMM for SI recognition.

7. REFERENCES

[1] M. Bilginer Gülmezoğlu, Vakif Dzhafarov, Mustafa Keskin, and Ataiay Barkana, "A novel approach to isolated word recognition," *IEEE Trains. Speech and Audio Processing*, vol. 7, No. 6, pp. 620-628, Nov. 1999.

[2] M. Bilginer Gülmezoğlu, Vakif Dzhafarov, and Ataiay Barkana, " The vommon vector approach and its relation to principal component analysis," *IEEE Trains. Speech and Audio Processing*, vol. 9, no. 6, pp. 655-662, Nov. 2001.

[3] Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc.*, vol.39, pp. 1-38, 1977.

[4] B. H. Juang, "Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT&T Tech. J.*, vol. 64, no. 6, pp. 1235-1249, 1985.

[5] D. J. Liu, "A Study on Continuous Speaker-Independent Mandarin Digits Recognition based on GCVHMM and FAR Algorithm," Ph.D. dissertation, Dept. Elect. Con. Eng., National Chiao Tung Universuty, Hsinchu, 2002.