SUBSPACE-BASED SPEAKER-INDEPENDENT VOWEL RECOGNITION

R. Muralishankar and Douglas O'Shaughnessy

INRS-EMT, University of Quebec, Canada.
murali@inrs-emt.uquebec.ca, dougo@inrs-emt.uquebec.ca

ABSTRACT

In this paper, we present a subspace-based approach for speaker-independent vowel recognition. Five vowels (/aa/,/eh/,/iy/,/ow/ and /uw/) from the TIMIT database were considered for the task. The subspaces representing two different vowel classes may have a large common subspace due to speaker variability, noise and coarticulation. We use common principal component (CPC) [1] and its extension i.e., partial-Common principal component (pCPC) to obtain a specific subspace for each vowel which is insensitive to variations. We perform CPC analysis on the covariance matrices of the vowels. pCPC gives q eigenvectors which are common to all vowels and $(p-q)$ vowel specific eigenvectors. For each value of q, vowel specific subspaces are obtained. An input vector from an unknown vowel is classified based on the maximum length of its projection on the specific subspaces. We have choosen 18-dimensional Mel-Frequency Cepstral coefficients as a feature in our recognition task. The specific subspace is treated as a transformation matrix which enhances the vowel-specific information in the feature vector and, inturn, increases signal-to-noise ratio. Recognition experiments were performed on vowels extracted from a multiple speaker set taken from different dialect regions in the TIMIT database. Results for each vowel-specific subspace are presented for different values of q ranging from 1 to 5. The results are encouraging in the context of a speaker-independent framework. Visual Analysis of the vowel basis spectra provides useful and interesting information by highlighting the importance of different frequency regions.

1. INTRODUCTION

Speech recognition plays an increasingly important role in voice web technologies by allowing users to access web sites via telephone using spoken commands. Efficient feature extraction is the key to good performance in speech recognition. An efficient feature should be able to capture the variability in the data caused by a desired source (DS) while suppressing the variability caused by undesirable sources (UDS). In vowel recognition, it is highly desirable to have features which carry linguistic variability of a particular vowel while suppressing the linguistic information of other vowels and speaker variability. In this paper, a scheme to decompose the feature space into subspaces which carry the lingustic variability of a particular vowel is proposed.

Our approach is motivated from subspace-based vowel- consonant segmentation [2] where speech segments are classified as vowels and consonants. In this paper, we would like to identify the unknown vowel segment. The recognition task is accomplished by generating specific subspaces for each vowel that are relatively insensitive to the UDS. This is done by estimating the directions in the feature space where the ratio of the DS variance to UDS variance for a given vowel is high. A conventional feature can then be projected into this subspace to make it relatively insensitive to UDS.

This work is supported by the Canadian government under the NSERC Strategic Partnerships Program.

2. FEATURE TRANSFORMATION

Existing techniques like LPC can model the statistical properties of different vowel sounds. Here, if we consider a particular vowel sound V_i (where $i = 1, 2, ..., k$ and k is the number of vowels in the training set) in the feature vectors as signal (DS) then the complementary $k-1$ vowel sounds V_j (where $j \neq i$) are noise (UDS). We present a linear transformation that aims at finding a subspace of the feature space where the Signal-to-Noise ratio (SNR) is maximum. Such a decomposition can be arrived at by representing V_i and V_j by training vectors obtained from the TIMIT Database. The directions in the feature space where the SNR is maximum can be derived by the partial-Common Principal Component (pCPC) of the covariance matrices of the above vectors. Consider a linear transformation matrix W that maps the original feature vectors x onto $\widehat{x}$:

$$\widehat{x} = W^T x \tag{1}$$

where x is an n-dimensional vector, $\widehat{x}$ is an m-dimensional vector $m \leq n$, and W is an $n \times m$ matrix with m linear independent columns. Let d_i and d_j represent the training vectors containing V_i and V_j respectively, in the original feature space. The covariance matrices for the above training vectors can be written as

$$\begin{aligned} C_i &= E[(d_i - \overline{d_i})(d_i - \overline{d_i})^T] \\ C_j &= E[(d_j - \overline{d_j})(d_j - \overline{d_j})^T] \end{aligned} \tag{2}$$

where $\overline{d_i}$ and $\overline{d_j}$ represent the means of d_i and d_j respectively. We wish to find $W^{(i)}$ to maximize the ratio of the variance of V_i to V_j after transformation. If the density function of the d_i and d_j are assumed to be Normally distributed then their covariance matrices after transformation are given by

$$\begin{aligned} \widehat{C_i} &= W^{(i)T} C_i W^{(i)} \\ \widehat{C_j} &= W^{(i)T} C_j W^{(i)} \end{aligned} \tag{3}$$

A simple measure of the variance or the "scatter" is the determinant of the covariance matrix [4]. Thus the criterion function to be maximized for an i^{th} vowel sound is given by

$$J(W^{(i)}/C_i) = \frac{\left|\widehat{C_i}\right|}{\left|\widehat{C_j}\right|} = \frac{\left|W^{(i)T} C_i W^{(i)}\right|}{\left|W^{(i)T} C_j W^{(i)}\right|} \tag{4}$$

Here, the maximization has to be carried out for all j where $j = 1, 2, .., k$ and $j \neq i$. We get $W^{(i)}$ after maximizing eq. 4 for each of

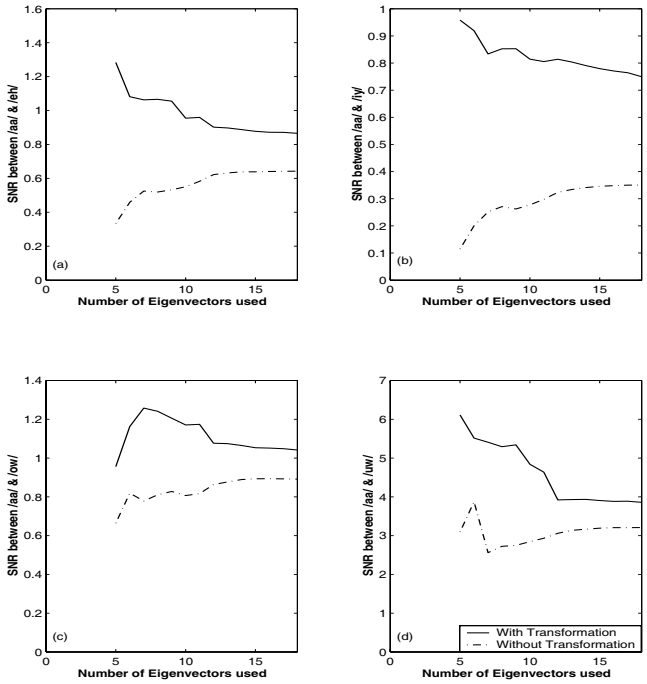

Fig. 1. Variation of SNR with and without transformation as a function of feature dimension. (a) γ_{12} (SNR between /aa/ and /eh/). (b) γ_{13} (SNR between /aa/ and /iy/). (c) γ_{14} (SNR between /aa/ and /ow/). (d) γ_{15} (SNR between /aa/ and /uw/).

the i^{th} vowel sounds and it spans the i^{th} vowel subspace. The above criterion function cannot be solved using Generalized eigenvalue decomposition because the optimization has to be carried out for more than two covariance matrices. To solve for $W^{(i)}$'s, we use the i^{th} specific component obtained from pCPC of C_i where $i = 1, 2, ...k$.

In [5], Malayath et al introduced a SNR measure, defined as the ratio of these variances when original feature vectors are projected onto $w_1^{(i)}$:

$$\gamma_{ij} = \frac{w_1^{(i)T} C_i w_1^{(i)}}{w_1^{(i)T} C_j w_1^{(i)}} \tag{5}$$

If the first m eigenvectors are used, the SNR becomes

$$\gamma_{ij} = \frac{trace\left(W^{(i)T} C_i W^{(i)}\right)}{trace\left(W^{(i)T} C_j W^{(i)}\right)} \tag{6}$$

The SNR of the original feature vectors can be calculated from eq. 6 by making $W^{(i)}$ an identity matrix. Figures 1(a) to 1(d) show the SNR between V_i(/aa/) and V_j (/eh/,/iy/,/ow/ and /uw/) before and after transformation for Mel-frequency cepstral coefficients (MFCC). From the figures, it can be seen that the SNR of the transformed feature vectors is substaintially higher than that of the original feature vectors. Since the specific vowel subspaces are spanned by the eigenvectors after the 4th dimension (here, the first four eigenvectors are common principal components), we have shown SNR starting from the 5th dimension. Also, since the eigenvalues are ordered as a decreasing sequence, the SNR after transformation decreases with increase in dimension.

3. COMMON PRINCIPAL COMPONENTS

The common principal components (CPC)[1] model hypothesizes that multiple datasets share common components, though each dataset has different eigenvalues associated with those components. The CPC hypothesis for k, $p \times p$ covariance matrices, $\Sigma_1, \Sigma_2, ..., \Sigma_k$, is:

$$\Sigma_i = B\Lambda_i B^T, \qquad i = 1, ..., k \tag{7}$$

where B is an orthogonal $p \times p$ matrix, and $\Lambda_i = diag(\lambda_{i1}, ..., \lambda_{ip})$. Note that a component may have a large eigenvalue associated with one dataset, but a small eigenvalue associated with another dataset. Hence there is no canonical ordering of the components by ordering them according to the size of their eigenvalues as in principal component analysis. The common principal components model is equivalent to postulating that the covariance matrices for the datasets are simultaneously diagonalizable by the same orthogonal matrices, i.e., the matrix of common components. The elements of the resulting diagonal matrices contain the respective eigenvalues. Thus:

$$B^T \Sigma_i B = \Lambda_i \tag{8}$$

$i = 1, ..., k$, where B and Λ_i are defined as above. Note that a necessary and sufficient condition for the existence of B is that $\Sigma_1, \Sigma_2, ..., \Sigma_k$ are commutable, that is, $\Sigma_i \Sigma_j = \Sigma_j \Sigma_i$ for all i, j. The sample covariance matrices are modeled as

$$C_i = B\Lambda_i B^T + U_i \tag{9}$$

where C_i is the i^{th} (unbiased) sample covariance matrix and U_i is the i^{th} matrix of error terms. We assume that the original measurements follow a multivariate normal distribution and consequently that $(s_i - 1)C_i$ (s_i is the sample size of the i^{th} dataset) follows a Wishart distribution. Estimating equations are derived by maximizing the likelihood, subject to the constraint of orthogonality on B. Solving the equations, we obtain the maximum likelihood solution for B. The F-G algorithm [3] solves these equations, though without guarantee of global optimality. The estimating equations are for $m, r = 1, ..., p$, $m \neq r$.

$$\beta_m^T \left(\sum_{i=1}^{k} (s_i - 1) \left(\frac{\beta_m^T C_i \beta_m - \beta_r^T C_i \beta_r}{\beta_m^T C_i \beta_m \beta_r^T C_i \beta_r} \right) C_i \right) \beta_r = 0 \tag{10}$$

with $\beta_j^T \beta_j = 1$ and $\beta_j^T \beta_h = 0$ for $j \neq h$, where β_j is the j^{th} column of B. Further, a likelihood ratio statistic is derived to test for the significance of deviations from the model.

3.1. Partial-Commmon Principal Components (pCPC)

Flury [1] extends the CPC model by developing a partial-common principal components model. The partial CPC model hypothesizes that there are only q of p eigenvectors common to all Σ_i. The remaining $(p - q)$ are *specific* to each dataset. That is

$$B_i^T \Sigma_i B_i = \Lambda_i \tag{11}$$

where B_i are orthogonal matrices such that $B_i = [B_1 : B_{2i}]$, B_1 is a $p \times q$ orthonormal matrix of q common eigenvectors, and B_{2i} are $p \times (p - q)$ matrices with $(p - q)$ eigenvectors specific to the i^{th}dataset. Flury indicates that the maximum likelihood equations solving this model are extremely laborious to implement. He recommends instead an approximate solution using the CPC estimates.

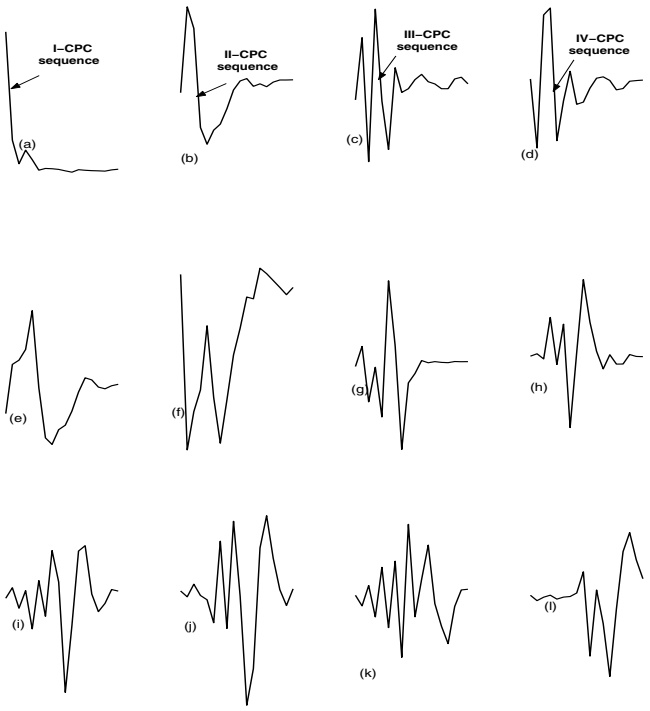

Fig. 2. Vowel subspace sequences. (a)-(d) I-IV CPC sequences for all the vowels. (e)-(l) Specific eigenvector sequences for vowel /aa/.

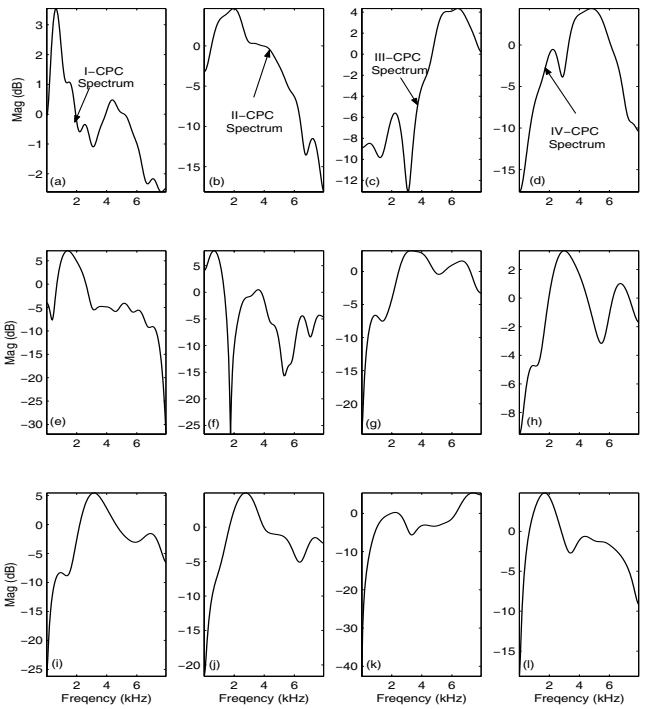

Fig. 3. Frequency responses of the common and specific subspace eigenvectors. (a)-(d) Spectral plots of I-IV CPCs. (e)-(l) Spectral plots of specific eigenvectors of vowel /aa/.

The method involves first obtaining approximate maximum likelihood estimates of the common components, B_1, from the CPC estimates. Then the B_{2i} are obtained by finding B_{2i} that diagonalize C_i subject to B_{2i} being orthogonal to B_1. In our vowel recognition experiment, we derive B_1 and $W^{(i)}(= B_{2i}$, specific eigenvectors for the i^{th} vowel) for each vowel using pCPC. Plugging $W^{(i)}$ into eq. 6 maximizes the SNR for the i^{th} vowel. We have already shown SNR for vowel /aa/ wrt. all other vowels in Fig. 1. Common subspace sequences for all the vowels are shown in Figs. 2(a) to 2(d) while Figs. 2(e) to 2(l) show the specific eigenvector sequence for the vowel /aa/.

4. FEATURE TRANSFORMATION AS FILTERBANK

If we consider the transformation as a filter bank as was done in [2], we get the relative importance of frequency bands for each vowel. Now consider a $N \times N$ transformation matrix $W^{(i)}$ such that

$$\widehat{x} = W^{(i)} x \tag{12}$$

Let $\omega_l^{(i)}$ be the l^{th} column of $W^{(i)}$. The l^{th} component of $\widehat{x}$ is the inner product of x with $\omega_l^{(i)}$. That is,

$$\widehat{x}_l = \omega_l^{(i)T} x = \sum_{m=0}^{N-1} x(m)\omega_{ml}^{(i)} \tag{13}$$

where $\omega_{ml}^{(i)}$ is the m^{th} component of $\omega_l^{(i)}$. This summation can be interpreted as filtering of $x(n)$ advanced by $N-1$ samples [6]:

$$\widehat{x'}_l(n) = \sum_{m=0}^{N-1} x(n+N-1-m)h_l^{(i)}(m) \tag{14}$$

where $h_l^{(i)}(m) = \omega_{N-1-m,l}^{(i)}$ are the impulse response coefficients of the filter $H_l^{(i)}(z)$. The sequences $\widehat{x}_l(n)$ are obtained by downsampling the sequences $\widehat{x'}_l(n)$ by N. In the transform domain, the convolution in eq. 14 can be written as

$$\widehat{X_l}(e^{j\omega}) = X(e^{j\omega})H_l^{(i)}(e^{j\omega}) \tag{15}$$

where $H_l^{(i)}(e^{j\omega})$ is the l^{th} filter frequency response for the i^{th} vowel. We know that the LPC-Cepstrum is the inverse Fourier transform of the logarithm of the all-pole LPC spectrum. Thus, when the input features $x(n)$ are the LPC-cepstral vectors, $X(e^{j\omega})$ represents the log spectrum. Therefore, according to eq. 15, the transformation process can be seen as a multiplication of the log spectrum by the frequency response of the filters corresponding to the largest eigenvalues which indicates the relative importance of different frequency bands for the i^{th} vowel. Spectral plots for common principal components are shown in Figs. 3(a) to 3(d) while Figs. 3(e) to 3(l) show the specific component spectral plots for the vowel /aa/. We have used LPC-Cepstrum for subspace spectral plotting and MFCC as feature for recognition.

5. SUBSPACE-BASED RECOGNITION

The specific subspaces for all the vowels are obtained using pCPC. The test signal is divided into overlapping frames and the feature vector x_l corresponding to the l^{th} frame is obtained using MFCC. We obtain the recognition result as follows. The length of the projection $\widehat{x}_i$ on the vowel subspace $W^{(i)}$ is used as a similarity measure between the input vector x and the class i. The input vector is

Table 1. Recognition performance (in %) of specific vowel subspaces derived using pCPC.

Specific subspace for each vowel obtained with										
	$q=1$		$q=2$		$q=3$		$q=4$		$q=5$	
Vowels	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
/aa/	72.75	65.11	64.67	58.13	64.67	55.81	64.67	58.91	61.37	52.71
/eh/	55.99	47.16	58.05	57.54	58.80	53.77	56.92	54.71	57.11	56.60
/iy/	90.05	88.40	65.94	56.15	66.07	61.23	59.67	55.43	58.58	55.07
/ow/	56.69	62.96	56.69	51.85	56.33	58.02	57.39	67.90	59.85	62.96
/uw/	71.60	55.55	62.96	51.85	59.25	51.85	55.55	55.55	58.02	44.44

then classified according to the maximal similarity value:

$$\underset{i=1,...,k}{argmax} \ \|\widehat{x}_i\|^2 = \underset{i=1,...,k}{argmax} \ \left\|W^{(i)T}x\right\|^2 \qquad (16)$$

We tested with the training dataset to determine the number of eigenvectors of $W^{(i)}$ in combination (for a particular value of q) that would give higher recognition rate. We found that the number of eigenvectors of $W^{(i)}$ in combination are different for different vowel specific subspaces. It also varies with the value of q (number of common components). This suggests that the subspace dimensions are different for different vowel sounds and also that a common subspace associated with each vowel is also different. We also found that the recognition performance of the testing dataset follows the same pattern as that of training.

6. RESULTS AND DISCUSSION

Vowel recognition experiments were conducted on the TIMIT data base. This 8 major dialect regions of the United States. The speech signals are stored in two major sets in the TIMIT database - "train" and "test", which are to be used for training and testing purposes, respectively. The speech data in each set are further separated into 8 subsets, dr1 to dr8, according the speakers' dialect regions. We have selected 5 vowels /aa/,/eh/,/iy/,/ow/ and /uw/ from the continuous sentences in the train set, dr3. dr3 contains 72 and 26 speakers for training and test utterances, respectively.

Feature vectors were obtained for each frame of a vowel from the train and test sets. The duration of each frame of speech was 30 ms, with an overlap of 20 ms between successive frames. Each frame of speech was Hamming windowed and processed to yield a 18-dimensional feature vector. For obtaing MFCC, the Mel-scale was simulated using a set of 18 triangular filters. For LPC-Cepstrum, an 18th order LPC analysis was performed after preemphasis with $a = 0.95$. A covariance matrix was generated for each of the above vowels from dr3. pCPC was performed on the vowel covariance matrices. Here, for every vowel, q (number of common principal components for vowels) was varied from 1 to 5 and for each value of q, specific subspaces were obtained. Feature vectors were obtained for each frame of a test vowel.

Table 1 gives the speaker-independent vowel recognition performance of our approach for q varying from 1 to 5. From the table, we can see recognition performance of vowels /aa/, /iy/ and /uw/ is better with $q = 1$. This means that these vowels have less overlap between the other vowels. This is quite valid from the fact that these vowels are placed at three corners of the *classic vowel triangle*. For vowels /eh/ and /ow/ better performance is obtained when specific subspaces are obtained by setting $q > 1$. This shows that these vowels have comparatively more than one CPC. The results that were shown in Table 1 are from the dr3 train and test set. We have also tested our algorithm with train and test datasets of the above vowels extracted from the TIMIT database, across all the dialect regions (dr1 to dr8). The recognition performances were almost similar to that one presented in Table 1. This connotes that there is no drastic downfall in recognition performance with respect to the train and test sets extracted from different dialect regions and also with different speakers. Table 2 shows the confusions with our vowel recognition scheme. The average recognition performance of our scheme is 71.72%.

Table 2. Confusion matrix for Vowel recognition using Vowel specific subspaces.

	/aa/	/eh/	/iy/	/ow/	/uw/
/aa/	84	8	5	20	12
/eh/	35	122	21	19	15
/iy/	8	23	244	1	0
/ow/	5	11	0	55	10
/uw/	5	1	1	5	15

7. CONCLUSION

We have proposed a new subspace based approach for speaker independent vowel recognition. Our subspace based approach uses pCPC to generate vowel specific subspaces. We have shown that these vowel specific subspaces improve SNR of a feature vector. The filter bank interpretation of the feature transformation throws light on the relative significance of different frequency bands for a particular vowel. We have shown speaker independent vowel recognition peformance in Table 1 and also similar performances are noticed for the test datasets across all the dialect regions.

8. REFERENCES

[1] B. Flury, "*Common Principal Components and related multivariate models*," John Wiley and Sons, Inc., 1988.

[2] R. Muralishankar, A. Vijaya Krishna and A. G. Ramakrishnan, "Subspace based vowel-consonant segmentation," *IEEE Workshop on Statistical Signal Processing*, pp. 589-592, 2003.

[3] B. Flury and W. Gautschi, "An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form," *SIAM J. Sci. Stat. Comput.*, vol. 7, no. 1, pp. 169-184, 1986.

[4] R. Duda and P. Hart, "*Pattern Classification and Scene Analysis,*" John Wiley and Sons, Inc., 1973.

[5] N. Malayath, H. Hermansky and A. Kain, "Towards decomposing the source of variability in speech," in *Proc. EUROSPEECH'97*, Rhodes, Greece, 1997.

[6] A. Makur, "BOT's based on nonuniform filter banks," *IEEE Trans. Signal Proc.*, vol. 44, no. 8, pp. 1971-1981, 1996.