Towards decomposing the sources of variability in speech

Narendranath Malayath[1], **Hynek Hermansky**[1,2] **and Alexander Kain**[1]
[1]**Oregon Graduate Institute of Science and Technology**, Portland, Oregon, USA.
[2]**International Computer Science Institute**, Berkeley, California, USA.

Abstract

In this paper a method to decompose a conventional feature space (LPC-cepstrum) into subspaces which carry information about the linguistic and speaker variability is presented. Principal component analysis is used to study the correlation between these sub-spaces. Oriented principal component analysis (OPCA) is then used to estimate a sub-space which is relatively speaker-independent. A method to estimate the dimensionality of the speaker independent sub-space is also presented. Original features can now be projected into the speaker independent sub-space to make them less sensitive to speaker variations. Finally the effectiveness of the proposed method in suppressing the speaker dependence is studied by experiments conducted on two different databases.

1 Introduction

Efficient feature extraction is the key to robust speech processing systems. An efficient feature extraction technique should be able to capture the variability in the data caused by a desired source while suppressing the variability caused by undesirable sources. For example, in speech recognition, it is highly desirable to have features which carry mainly linguistic information(LI). Similarly for speaker recognition it is important to have features which carry mainly the speaker specific information(SI). In the case of speech recognition the LI can be considered as the signal and SI as noise. In this paper a scheme to decompose the feature space into sub-spaces which carry information about (i) linguistic variability (relatively independent of speaker) and (ii) speaker variability (relatively independent of linguistic variability) is proposed. This decomposition is achieved by representing LI and SI by appropriate difference vectors. Difference between feature vectors extracted from different phonemes uttered by the same speaker mainly carry the linguistic variability ($\mathbf{d}_l$). Similarly, the speaker variability in the feature space is represented by the difference vectors($\mathbf{d}_s$) between the feature vectors extracted from the same phoneme uttered by different speakers. Such a decomposition can be used to estimate a linguistic sub-space which is relatively less sensitive to speaker variability. This is done by estimating the directions in the feature space where the ratio of the variance caused by the LI to that caused by SI is high. A conventional feature can then be projected into this sub-space to make it relatively insensitive to speaker variability.

2 Decomposition of the feature space

In this section a method to decompose a conventional feature space (defined by LPC-cepstrum) into sub-spaces carrying mainly LI and SI is presented. The initial feature representation $\mathbf{x}$ is the LPC-cepstrum and is considered as a random variable. Figure 1 shows the feature vectors extracted from a segment of speech from two speakers. The rectangular boxes represent feature vectors extracted from a frame of speech data. It is assumed that the segments of speech uttered by the two speakers are linguistically identical(same phonemes) and are perfectly time allingned. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be the cepstral vectors from two different phonemes uttered by the same speaker. The difference vector carrying LI is given by

$$\mathbf{d}_l = \mathbf{x}_2 - \mathbf{x}_1. \qquad (1)$$

By taking the difference between $\mathbf{x}_2$ and $\mathbf{x}_1$ the information which is common to $\mathbf{x}_2$ and $\mathbf{x}_1$ is removed. Hence the static (stationary) speaker characteristics and the channel effects are suppressed. Thus it can be concluded that the difference vector $\mathbf{d}_l$ mainly carry information about the linguistic variability and the variance caused by the dynamic speaker characteristics. Now consider the case where $\mathbf{x}^1$ and $\mathbf{x}^2$ represent the LPC-cepstrum extracted from the same phoneme uttered by two different speakers. Since $\mathbf{x}^1$ and $\mathbf{x}^2$ are features extracted from the speech signal corresponding to the same phoneme their difference will mainly contain SI. The difference vector representing SI is given by

$$\mathbf{d}_s = \mathbf{x}^2 - \mathbf{x}^1. \qquad (2)$$

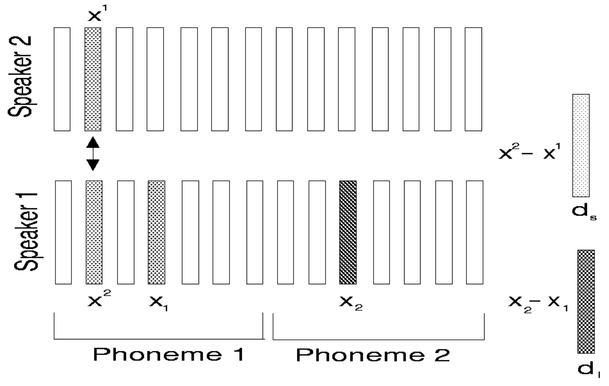

Figure 1: Derivation of the difference vectors.

Since $\mathbf{x}^1$ and $\mathbf{x}^2$ carry the same linguistic information the difference vector $\mathbf{d}_s$ will mainly carry information about the speaker variability and the difference in the channel and environmental condition captured by the utterances of the two speakers. If (i) the dynamic (non-stationary) speaker characteristics is negligible and (ii) the channel and environmental conditions captured by the speech signals of both the speakers are identical, then the sub-spaces defined by the random vectors $\mathbf{d}_l$ and $\mathbf{d}_s$ capture the linguistic and speaker variability respectively.

3 Subspace based feature extraction

In this section a subspace based feature extraction technique which is expected to yield features which are less sensitive to speaker variations is presented. First a conventional feature extraction technique is used to extract features which contain both linguistic and speaker information. Let this initial feature be represented by $\mathbf{x}$. Then the idea is to extract a set of basis vectors which point to those directions in the feature space where the ratio of variance caused by LI to that caused by SI is maximum. Let these basis vectors be represented by, $\mathbf{e}_{oi}$ i = 1,2 .. k. Now the original feature vectors $\mathbf{x}$ can be projected onto these basis vectors as shown by the following equation.

$$\mathbf{o} = \mathbf{E}_o^T \mathbf{x}, \tag{3}$$

where $\mathbf{E}_o$ is a matrix whose columns are composed of the basis vectors. After the projection, ratio of the variance of $\mathbf{o}$ caused by LI to that caused by SI is maximum.

3.1 A method to derive the basis vectors

In this section a method to derive the basis vectors, $\mathbf{e}_{oi}$, from $\mathbf{d}_l$ and $\mathbf{d}_s$ is developed.

From the random vectors which represent LI and SI, the corresponding covariance matrices can be computed by the following equations.

$$\mathbf{R}_l = E[(\mathbf{d}_l - \overline{\mathbf{d}_l})(\mathbf{d}_l - \overline{\mathbf{d}_l})^T]. \tag{4}$$

$$\mathbf{R_s} = E[(\mathbf{d}_s - \overline{\mathbf{d}_s})(\mathbf{d}_s - \overline{\mathbf{d}_s})^T]. \tag{5}$$

Since the objective is to maximize the variance caused by LI and minimize the variance caused by SI the objective function that we are interested in maximizing can be written as,

$$\frac{Signal}{Noise} = \frac{LI}{SI} = \frac{E\left(\mathbf{d}_l^T \mathbf{e}_i\right)^2}{E\left(\mathbf{d}_s^T \mathbf{e}_i\right)^2} = \frac{\mathbf{e}_i^T \mathbf{R_l} \mathbf{e}_i}{\mathbf{e}_i{}^T \mathbf{R_s} \mathbf{e}_i} = \frac{S_i}{N_i} = \rho_i.$$

In the above equation S_i and N_i are the amounts of signal and noise variance captured by $\mathbf{e}_i$. Note that we are interested in finding the direction $\mathbf{e}_i$ which maximizes the signal-to-noise ratio, ρ_i. Deriving such directions (or projections) is nothing but the solution to the following generalized eigen value problem.

$$\mathbf{R_l} \mathbf{e}_{o_i} = \lambda_{o_i} \mathbf{R_s} \mathbf{e}_{o_i}, \tag{6}$$

The solution to the above stated generalized eigen value problem is called the oriented principal components of the random vector pair ($\mathbf{d}_l$, $\mathbf{d}_s$). They are called oriented due to the fact that the principal component $\mathbf{e}_o$ is steered by the distribution of $\mathbf{d}_s$. It will be oriented towards the direction where $\mathbf{d}_s$ has the minimum variance while maximizing the projection energy of $\mathbf{d}_l$ [1]. From now onwards we refer to the ratio $\frac{\mathbf{e}^T \mathbf{R_l e}}{\mathbf{e}^T \mathbf{R_s e}}$ as the signal-to-noise ratio. Instead of a single basis vector if a set of basis vectors are used (for example the first few oriented principal components) then the signal-to-noise ratio is given by $\frac{trace(\mathbf{E}^T \mathbf{R_l E})}{trace(\mathbf{E}^T \mathbf{R}_s \mathbf{E})}$, where the columns of the matrix $\mathbf{E}$ are composed of a set of oriented principal components. The original SNR can be computed by making $\mathbf{E}$ an identity matrix. Also note that the space spanned by the oriented principal components represents a speaker independent subspace.

3.2 Estimation of the dimensionality of the speaker independent subspace

In this section a method to estimate the dimensionality of the speaker independent subspace is presented. This estimation can be made solely depending on the signal-to-noise ratio. But it must be noted that a direction with high SNR can be one where both the signal variance and noise variance is low (but their ratio being large). While estimating the dimensionality of the speaker independent subspace it is desirable to deemphasize such direction. This can be achieved by weighting the SNR ρ_{o_i}, with S_i. The following equation shows how this weighted signal to noise ratio can be used to estimate the dimensionality of the subspace.

$$p \quad = \quad \arg\max_{j} \sum_{i=1}^{j} \rho_{o_i} S_i . \tag{7}$$

Thus from the above equation, p is the dimensionality of the subspace which maximizes the SNR and at the same time captures most of the signal variance.

4 Experiments

In this section the results obtained in applying the proposed method to two different databases are presented.

4.1 VOICE database

This database consists of 15 sentences uttered by four male and four female speakers. While recording the sentences the speakers were asked to speak in synchrony with a metronome. This made sure that the same sentences spoken by different speakers were almost perfectly time allingned. LPC-cepstrum (10th order LPC represented by 15 cepstral coefficients) was extracted from these sentences. The difference vectors corresponding LI and SI were computed as described in Section 2.2 (equations 2 and 3). While the natural time alignment between the sentences spoken by different speakers was exploited to compute the $\mathbf{d}_s$, the phonetic labeling was used to compute $\mathbf{d}_l$. The covariance matrix corresponding to these vectors are given by $\mathbf{R}_l$ and $\mathbf{R}_s$. In order to compare the statistics of the distribution of $\mathbf{d}_l$ and $\mathbf{d}_s$, the eigen vectors corresponding to $\mathbf{R}_l$ and $\mathbf{R}_s$ were compared using the following equation,

$$C_i \quad = \quad {\mathbf{e}_{l i}}^T \mathbf{e}_{s i}, \tag{8}$$

where $\mathbf{e}_{l i}$ and $\mathbf{e}_{s i}$ are the i_{th} eigen vector derived from $\mathbf{R}_l$ and $\mathbf{R}_s$ respectively. The correlation C_i reflects the similarity of the distribution of $\mathbf{d}_l$ and $\mathbf{d}_s$. For example, if for all i, C_i is unity, then it means that the variability introduced by the LI and SI are so similar that they cannot be separated using a linear projection. Figure 2 shows the correlation C_i for all the eigen vectors. The first two eigen vector corresponding to LI and SI are highly correlated. This suggests that the major amounts of variance caused by LI and SI are oriented in the same direction in the feature space. It can also be noted that 3_{rd} through the 7_{th} eigen vectors of $\mathbf{R}_l$ and $\mathbf{R}_s$ are relatively uncorrelated. This suggests that the statistics of the distribution of $\mathbf{d}_l$ and $\mathbf{d}_s$ are essentially different and thus a set of basis vectors can be derived to improve the SNR. The set of basis vectors, $\mathbf{e}_{o_i}$ were then computed from $\mathbf{R}_l$ and $\mathbf{R}_s$ using equation 7. The original feature vectors were then projected into the space spanned by the basis vectors using the equation 1. Figure 3 shows the SNR of the projected and the original features. The doted line shows the variation of the SNR of the LPC-cepstrum as a function of the number of cepstral coefficients. The solid line indicates the SNR of the transformed feature. It can be observed that the SNR of the transformed feature is significantly higher than that of the original cepstral feature. It can also be noted that the highest SNR is obtained while using the first basis vector and as more and more basis vectors were used the SNR deteriorates. The optimum number of basis vectors were then estimated using equation 8. The dotted line in Figure 4 shows the variation of the weighted SNR as a function of the number of basis vectors used. From the figure it is clear that the weighted SNR is maximum when the first four basis vectors were used. This suggests that the optimum number of principal components to be used in order to extract the LI is around four.

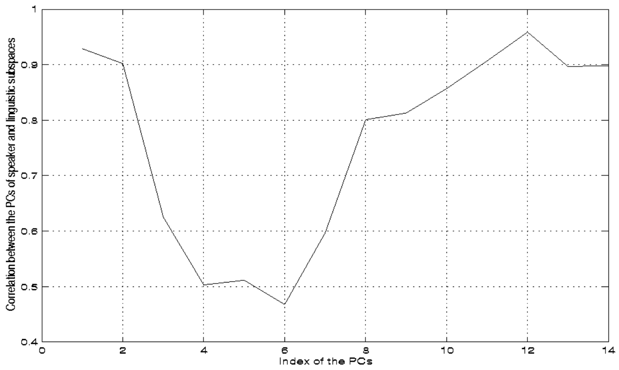

Figure 2: The correlation between the principal components extracted from the vectors representing LI and SI.

4.2 TIMIT database

In the previous section we observed that by using the first four oriented principal components the LI can be efficiently represented. In this section we attempt to evaluate the generalization capability of this method. By generalization capability we mean the performance of these basis functions on any dataset other than the one from which it was extracted. A training and a test set were identified from TIMIT. Each of these sets contain ten phonetically balanced sentences uttered by 100 speakers. Since the sentences spoken by different speakers were not allingned, DTW paths were used to derive the difference vectors, $\mathbf{d}_s$. The proposed method was then used to extract the basis vectors from the training set as well as the test set. Figure 5 shows the performance of the basis vectors on both the training and the test data. From the figure it is clear that the basis functions derived from the training data set performs almost as good as those derived from the test data. This shows that the proposed method is capable of finding the directions in the feature space which separates the SI from the LI irrespective of the type of data (provided that the training set is sufficiently large). The optimum number of basis vectors was then estimated using equation 8. The solid line in Figure 4 shows the variation of the weighted SNR. From the figure it is evident that the linguistic information can be efficiently represented by approximately four basis vectors. This observation is consistent with the earlier work reported by Hermansky [2, 3].

5 Conclusions

Results indicate that the proposed method to represent the LI and SI by appropriate difference vectors is effective in identifying subspaces corresponding to the LI and SI. Once these subspaces are identified then the oriented principal component analysis can be used as a tool to suppress the variance in undesired directions (noise or SI) and to enhance variance in the desired direction (signal or LI). We also observed that the optimal number of oriented principal components is four for the simultaneous enhancement of LI and suppression of SI. It was also demonstrated that the proposed method has strong generalization capability. i.e., the basis functions derived from a sufficiently large amount of data can be used to enhance the signal-to-noise ratio of any new set of data.

References

[1] K.I. Diamantaras and S. Y. Kung, *Principal Component Neural Networks - Theory and Applications*, John Wiely & Sons, first edition, 1996.

[2] Hynek Hermansky, "An efficient speaker-independent automatic speech recognition by simulation of some properties of human auditory perception," in *Proc. of ICASSP*, Dallas, 1987, pp. 1159–1162.

[3] Hynek Hermansky, "Perceptual linear predictive(plp) analysis of speech," *JASA*, vol. 87, no. 4, pp. 1738–1752, 1990.

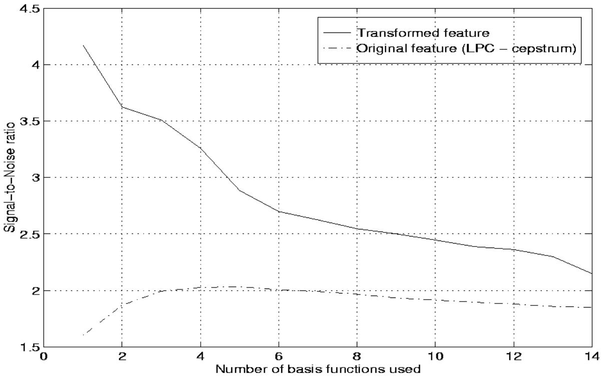

Figure 3: Demonstration of the improvement in signal-to-noise ratio due to application of the basis vectors.

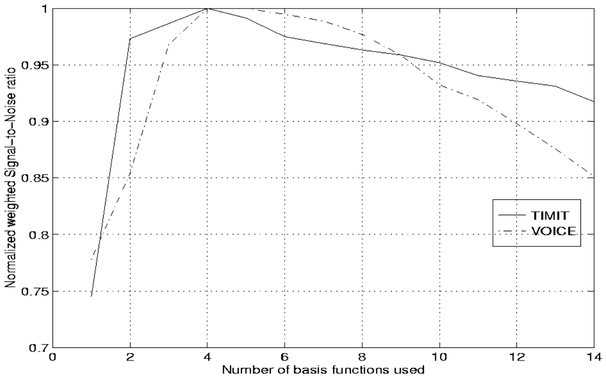

Figure 4: The weighted SNR is exhibiting a peak at around four indicating the optimal number of basis vectors to be used for representing the LI.

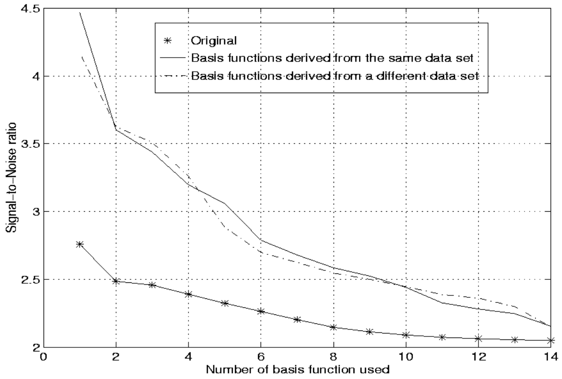

Figure 5: Comparison of the performance of the basis vectors when they were used on the training and test data.