

Speaker-Independent Feature Extraction by Oriented Principal Component Analysis

Narendranath Malayath¹, Hynek Hermansky¹, Alexander Kain¹ and Rolf Carlson²

¹Center for Spoken Language Understanding,
Oregon Graduate Institute of Science and Technology
Portland, Oregon, U. S. A.

²Center for Speech Technology (CTT), KTH,
Box 70014, S-10044 Stockholm, Sweden.

Abstract

In this paper a new method for deriving features which are less sensitive to speaker variations is proposed. The proposed method is based on oriented principal component analysis, a powerful statistical technique. First a simple method to decompose a conventional feature (LPC-cepstrum) into (i) vectors representing linguistic information (independent of speaker) and (ii) vectors representing speaker information is presented. Using these decomposed vectors a sub-space which is relatively independent of speaker variations is derived. A method to optimally 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 demonstrated 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 must be able 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 depend only on the linguistic information(LI). The extracted feature should not depend on the speaker characteristics. Similarly for speaker recognition it is important to have features which depend on the speaker specific information(SI) and should be less sensitive to the linguistic message. In the case of speech recognition the LI can be considered as the signal which is being corrupted by SI(noise) and vice-versa for speaker recognition. In this paper a method for efficiently extracting features which are less sensitive to speaker variability is

proposed. The method is based on estimating the directions in the feature space where the ratio of the variance caused by the LI to that caused by SI is high. The proposed method can be applied on top of any conventional feature extraction technique to reduce the sensitivity of the feature to speaker variability. The theory behind the proposed method is developed in the following section. In Section 3 we experimentally demonstrate the effectiveness of this method in suppressing speaker dependence of LPC-cepstrum. Section 4 derives conclusions from these experiments and summarizes the paper.

2 Subspace based feature extraction

In this section a method to derive linear projections which maximizes the signal-to-noise ratio of features 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}_{o_i} $i = 1, 2 \dots k$. Now the original feature vectors \mathbf{x} can be projected to these basis vectors as shown by the following equation.

$$\mathbf{o} = \mathbf{E}_o^T \mathbf{x}, \quad (1)$$

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}_i caused by LI to that caused by SI is maximum. Hence the columns of the matrix E spans a sub-space which is relatively speaker independent.

2.1 Issues

There are many issues to be addressed in order to derive the basis vectors. The first issue is how to represent the LI and SI. Assume that we have derived a set of vectors \mathbf{d}_l and \mathbf{d}_s such that \mathbf{d}_l represents only LI and \mathbf{d}_s solely represents SI. Once a representation of LI and SI are derived then a method to obtain the basis vectors (\mathbf{e}_{o_i} $i = 1, 2 \dots k$) has to be devised. Another important issue is the estimation of the dimensionality of this speaker independent subspace. This problem directly relates to the optimal number of dimensions required to represent LI. This is dependent on the statistics of the distribution of \mathbf{d}_l and \mathbf{d}_s in the feature space.

2.2 Representation of SI and LI

The initial feature representation \mathbf{x} is assumed to be the LPC-cepstrum and is considered as a random variable. Let \mathbf{x}_1 and \mathbf{x}_2 be the cepstral vectors from two different phonemes uttered by the same speaker. Since \mathbf{x}_1 and \mathbf{x}_2 are features extracted from the speech of the same speaker their difference will contain only LI. The difference vector representing LI is thus given by

$$\mathbf{d}_l = \mathbf{x}_1 - \mathbf{x}_2. \quad (2)$$

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 contain only SI. The difference vector representing SI is given by

$$\mathbf{d}_s = \mathbf{x}^1 - \mathbf{x}^2. \quad (3)$$

Thus the random vectors \mathbf{d}_s and \mathbf{d}_l represent the SI and the LI respectively. It must be also noted that in the feature space \mathbf{d}_l represents the direction of speaker variability and \mathbf{d}_s represent the direction of linguistic variability.

2.3 A method to derive the basis vectors

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

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

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

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

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,

$$\frac{Signal}{Noise} = \frac{LI}{SI} = \frac{E(\mathbf{d}_l^T \mathbf{e})^2}{E(\mathbf{d}_s^T \mathbf{e})^2} = \frac{\mathbf{e}^T \mathbf{R}_l \mathbf{e}}{\mathbf{e}^T \mathbf{R}_s \mathbf{e}}. \quad (6)$$

In the above equation we are interested in finding the direction \mathbf{e} which maximizes the signal-to-noise ratio. Deriving such directions (or projections) is nothing but the solution to the following generalized eigen value problem.

$$\mathbf{R}_l \mathbf{e}_{o_i} = \lambda_{o_i} \mathbf{R}_s \mathbf{e}_{o_i} = \frac{S_i}{N_i}, \quad (7)$$

where S_i and N_i are the amounts of signal and noise variance captured by \mathbf{e}_{o_i} . 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 $\frac{\mathbf{e}^T \mathbf{R}_l \mathbf{e}}{\mathbf{e}^T \mathbf{R}_s \mathbf{e}}$ as the signal-to-noise ratio. If \mathbf{e} represent a set of basis vectors (for example the first few oriented principal components) then the signal-to-noise ratio is given by $\frac{trace(\mathbf{E}^T \mathbf{R}_l \mathbf{E})}{trace(\mathbf{E}^T \mathbf{R}_s \mathbf{E})}$. The original SNR can be computed by making \mathbf{E} an identity matrix. Also note the the space spanned by the basis vectors represents a speaker independent subspace.

2.4 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. 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 λ_{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.

$$p = \max_j \sum_{i=1}^j \lambda_{o_i} S_i. \quad (8)$$

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.

3 Experiments

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

3.1 VOICE database

This data base 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 set of metronomes. This made sure that the same sentences spoken by different speakers were almost perfectly time alligned. LPC cepstrum(10th order LPC represented by 15 cepstral coefficients) was extracted from these sentences. The difference vectors corresponding LI nd 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 the correlation matrix between the principal components extracted from \mathbf{R}_l and \mathbf{R}_s was computed by the following equation,

$$\mathbf{R} = \mathbf{E}_l^t \mathbf{E}_s, \tag{9}$$

where \mathbf{E}_l and \mathbf{E}_s are matrices whose columns are the principal component vectors. If \mathbf{R} is an identity matrix it means that the \mathbf{e}_l and \mathbf{e}_s are identical which in turn suggests that the variability introduced by the LI and SI are so similar that they cannot be separated using a linear projection technique. Figure 1 shows the mesh plot of the correlation matrix \mathbf{R} . Note that there is a significant number of off-diagonal elements with high values. 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 2 shows the SNR of the projected and the original features. The dotted 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 produces is significantly higher than that of the original cepstral feature. It can also be noted that the best SNR is obtained while using the first basis vector and as more and more basis vectors are used the SNR deteriorates. Then the optimum number of basis vectors were estimated using equation 8. The dotted line in Figure 3 shows the variation of the weighted SNR as

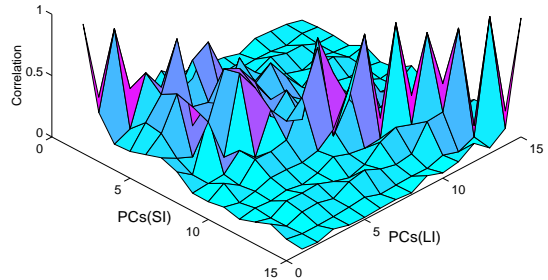


Figure 1: The correlation between the principal components extracted from the vectors representing LI and SI. High values of off-diagonal elements show that the LI and the SI are not oriented in the same directions in the feature space

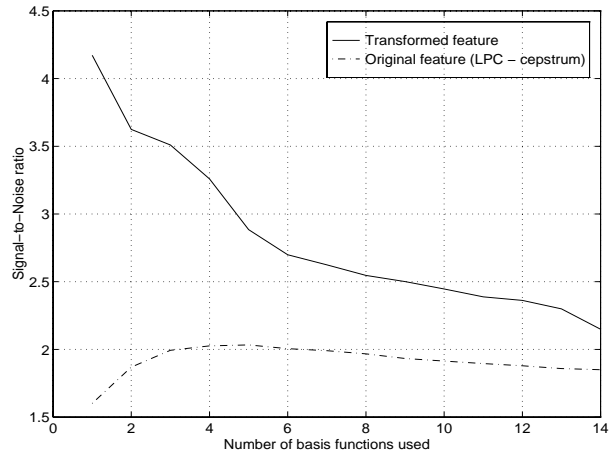


Figure 2: Demonstration of the improvement in signal-to-noise ratio due to application of the basis vectors.

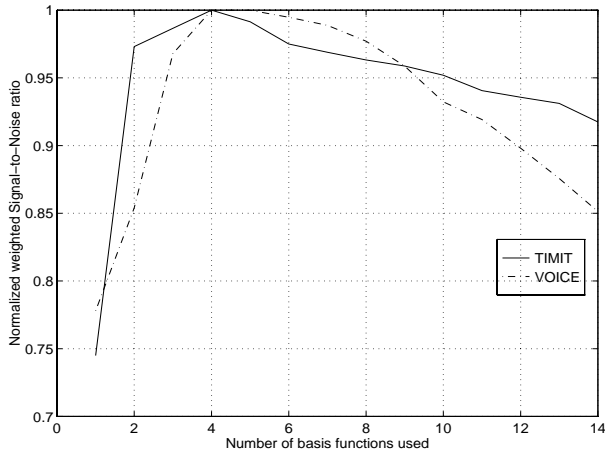


Figure 3: The weighted signal-to-noise is exhibiting a peak at around four indicating the optimal number of basis vectors to be used for representing the LI.

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 are used. This suggests that the optimum number of principal components to be used in order to extract the LI is around four.

3.2 TIMIT database

In the previous section we observed that by using the first four oriented principal components the LI can be separated from the SI. 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 has been extracted. A training and a test set was identified from TIMIT. Each of these sets contain ten phonetically balanced sentences uttered by 100 speakers. The proposed method was used to extract the basis vectors from the training set as well as the test set. Figure 4 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 3 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.

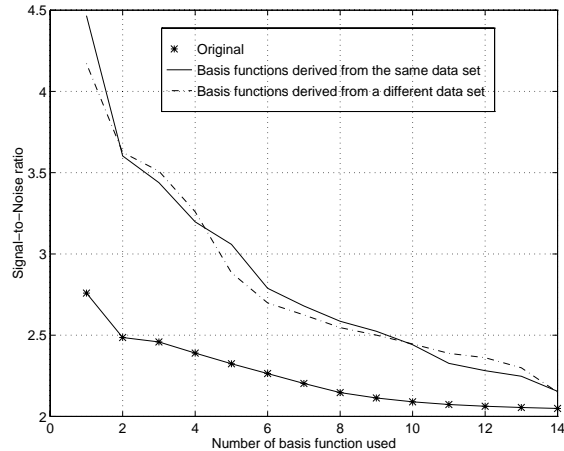


Figure 4: Comparison of the performance of the optimal basis functions when they were used on the training and test data

4 Summary and Conclusions

In this paper a new based on oriented principal component analysis was developed to derive features which are less sensitive to speaker variations. The proposed method can be used to suppress the speaker variability of any arbitrary set of features. The various steps involved can be summarized as

1. Extraction of a conventional feature (LPC-Cepstrum).
2. Representation of linguistic information and speaker information by difference vectors.
3. Estimation of a set of optimal basis vectors from the difference vectors.
4. Estimation of the number of basis vectors to be used by maximizing the weighted SNR.
5. Transformation of the original feature set by projecting it into the space spanned by the basis vectors.

Results indicate that the the proposed method to represent the SI and LI by appropriate difference vectors is effective in identifying subspaces corresponding to the SI and LI. Once these subspaces are identified then the oriented principal component analysis can be used as a powerful 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 Wiley & Sons, first edition, 1996.