FEATURE DECORRELATION METHODS IN SPEECH RECOGNITION. A COMPARATIVE STUDY

Eloi Batlle, Climent Nadeu and José A. R. Fonollosa

Universitat Politècnica de Catalunya*
Dept. of Signal Theory and Communications
Campus Nord UPC, Edifici D5, C/ Jordi Girona, 1-3. 08034 Barcelona, SPAIN.
Tel: +34 93 401 1066; fax: +34 93 401 6447. e-mail: eloi@gps.tsc.upc.es

ABSTRACT

In this paper we study various decorrelation methods for the features used in speech recognition and we compare the performance of each one by running several tests with a speech database. First of all we study the Principal Components Analysis (PCA). PCA extracts the dimensions along which the data vary the most, and thus it allows us to reduce the dimension of the data point without significant loss of performance. The second transform we study is the Discrete Cosine Transform (DCT). As it will be shown, it is an approximation of the PCA analysis. By applying this transform to FBE parameters we obtain the MFCC coefficients. A further step is taken with the Linear Discriminant Analysis (LDA), which, not only reduces the dimensionality of the problem, but also discriminates among classes to reduce the confusion error. The last method we study is Frequency Filtering (FF). This method consists of a linear filtering of the frequency sequence of the log FBE that both decorrelates and equalizes the variance of the coefficients.

1. INTRODUCTION

This work focuses on comparing various methods to decorrelate the set of parameters used in speech recognition.

In speech recognition, filter bank energies (FBE) are widely used because of their clear physical meaning. However, before use them as the parameter set for the Viterbi decoder, it is very useful to decorrelate and to compress them. The decorrelation process allows us to use diagonal covariance matrix, simplifying the system and reducing the number of parameters to train without loss of performance. In the case of the reduction of parameters, we find a trade-off between number of parameters and performance. On one side, the greater the number of parameters the greater the degrees of freedom the model will have to model the acoustic event, but on the other side, as more parameters we take, the system will become more complex and there will be more variance in the estimates of the probabilities.

During the past decade, much effort has been done to find transformations to reduce the number of parameters as well as to decorrelate them at the same time that we increase the performance of the system.

In this paper, we make a comparative study of some of these methods, showing their performance in several situations. We also show the advantage of introducing some sort of discrimination at the same time we decorrelate the feature vectors. As it will be shown in the comparative results, discriminative approaches achieve a significant error rate reduction maintaining the same (or little more) computational cost.

*This work was supported by the Spanish National Research Plan CICYT under grant TIC95-1022-C05-03.

2. BACKGROUND THEORY

In this section we briefly review the theory that involves some the methods used to incorrelate and to reduce the dimension of the speech parameters.

2.1. Filter Bank Analysis

Nowadays, the bank-of-filters front-end processor is used extensively. The sampled speech signal is passed through a bank of Q bandpass filters whose center frequency is distributed uniformly along a logarithmic frequency. This distribution tries to emulate the cochlear location of the hair cell in the human auditory system [8]. Since the aim of this analysis is to quantify the energy of each band of the spectrum of the signal, each of the bandpass signals are usually passed through a nonlinearity.

2.2. Principal Components Analysis

Principal Components Analysis (PCA) is a useful technique often used in statistical analysis of data. PCA is based on the calculation of the major directions of variations of a set of data points in a high dimension space. PCA will extract the direction of the greatest variance, assuming that the less variation of the data, the less information it carries. PCA has some interesting properties that are useful for dimension reduction. Among them, we can highlight that principal components are mutually orthogonal and the largest principal value is the direction of greatest variance. We obtain the projection matrix from the covariance matrix C, defined as

$$C = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)(x_n - \mu)^T \quad (1)$$

where N is the number of frames used in training, x_n is the nth frame, and μ is the mean of the frames

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n \quad (2)$$

The directions of the principal components are the eigenvectors with greatest eigenvalues of the covariance matrix

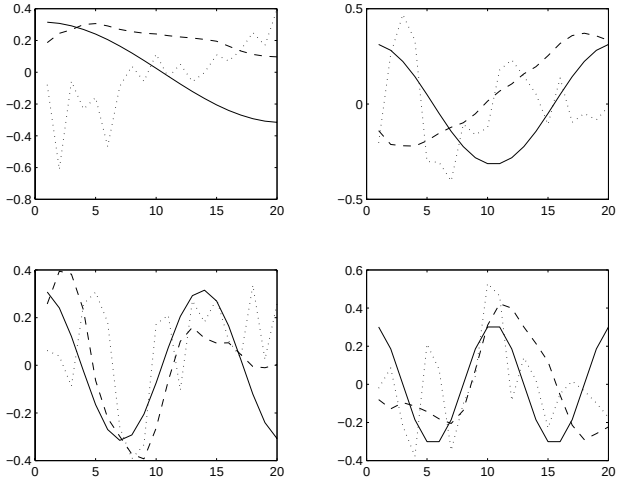

Figure 1. First components of the DCT (solid line), PCA (dashed line) and LDA (dotted line)

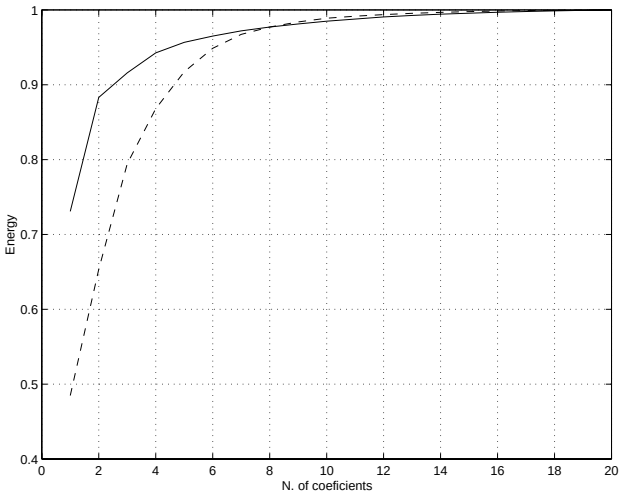

Figure 2. Energy of the signal versus the dimension of the projected space for PCA (solid line) and LDA (dashed line)

C. Since the covariance matrix is known to be real and symmetric, the eigenvectors are guaranteed to be orthogonal.

Examining 1 we can see that PCA analysis is very similar to the KL transform.

2.3. Melcepstrum Analysis

Mel-Frequency Cepstral Analysis (used to obtain the MFCC coefficients) is one the most used analysis technique in speech recognition for its simplicity and good performance. MFCC coefficients are obtained from the log filter-bank amplitudes (the filters are equally spaced along the mel-scale) using the Discrete Cosine Transform (DCT). DCT is defined as

$$c_i = \sqrt{\frac{1}{Q}} \sum_{j=1}^{Q} m_j \cos\left(\frac{\pi i}{Q}\left(j - 0.5\right)\right) \qquad (3)$$

As we can see in Figure 1, DCT is just an approximation of the *optimal* transform, i.e. the PCA transform.

2.4. Linear Discriminant Analysis

In section 2.2. we described an statistical method used to reduce the dimension of the space of interest. Through this section we will show a method that, not only reduces the dimension of the space, but also discriminates among classes. In this approach, the problem of finding a linear transformation is formulated in terms of a problem of minimizing a criterion of class separability function [1].

In LDA analysis, some strong assumptions are made. The first assumption is that all classes share a common within-class covariance (for a global LDA implementation). LDA also assumes a single Gaussian distribution per class, however, the use of multiple mixture densities shows better results [6]. We define the between-class matrix as

$$BSS = \frac{1}{N} \sum_{k=1}^{K} n_k \left(\mu_k - \mu\right)\left(\mu_k - \mu\right)^T \qquad (4)$$

and the within-class matrix as

$$WSS = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{n_k} \left(x_{kn} - \mu_k\right)\left(x_{kn} - \mu_k\right)^T \qquad (5)$$

where N is the total number of training frames, K is the number of the classes to separate, n_k is the number of training patterns of the kth class, x_{kn} is the nth training pattern of the kth class and

$$\mu_k = \frac{1}{n_k} \sum_{n=1}^{n_k} x_{kn} \qquad \mu = \frac{1}{N} \sum_{k=1}^{K} n_k \mu_k \qquad (6)$$

are the mean vector of each class and the global mean vector respectively.

To find the transformation matrix we orthogonalise the WSS matrix by rotating it so the features become independent. By scaling after the rotation, the distribution of WSS in the new space can be made to be the identity matrix. Then, the BSS matrix is projected into this new space and an eigenanalysis can be made to generate another rotation so that WSS and BSS are projected into the same space. It can be shown [6] that this sequence of rotations and scalings is equivalent to find the eigenvectors of the matrix $WSS^{-1}BSS$.

2.5. Frequency Filtering Analysis

In spite of the great performance of the MFCC parameters, they have at least three drawbacks. First of all, they do not have a clear physical meaning, second, they require a linear transformation from the log FBE, and third, in CDHMM with diagonal covariance matrices, applying a shaping window to the cepstral coefficients has an effect only on its length. In order to overcome those disadvantages, we can use the Frequency Filtering Analysis (FF) [7].

By using FF not only we decorrelate the parameters but in addition we approximately equalize the variance of the cepstral coefficients up to a given quefrency index. As it was shown by experimental results, a simple low-order FIR

filter suffices to improve significantly the performance of the recognition system. Choosing Q, the number of channels to use, appropriately, we can obtain the filter that equalizes the variance by a least-squares modeling.

After a few tests over several databases, it was seen that a filter that performs very well in a wide range of environments and values of Q is

$$h(z) = z - z^{-1} \tag{7}$$

Although this filter performs quite well in several environments, it is possible to obtain another filter adapted for each situation. Once we have decided the working environment for the recognition system, we can estimate the parameters of the new filter in order to obtain the optimum filter in terms of variance equalization for that environment.

3. EXPERIMENTAL RESULTS

The experiments described in this section were carried out using the TIMIT database [2]. The recognizer was used as a phonetic classifier without any grammar. The front end for the feature extraction was a filter-bank analysis, computed from a fast Fourier Transform (FFT), followed by a subsequent decorrelation and in some cases a discriminative process (PCA, DCT, LDA or FF).

All experiments are carried out using a window size of 20ms with a displacement of 10ms. The HMMs used have a left-to-right topology and three emitting states. The 48 classes (associated to HMMs) are the phone set defined in [5].

3.1. Results without temporal information

The first experiment tries to show the performance of each parameter transformation, without taking into account any other information.

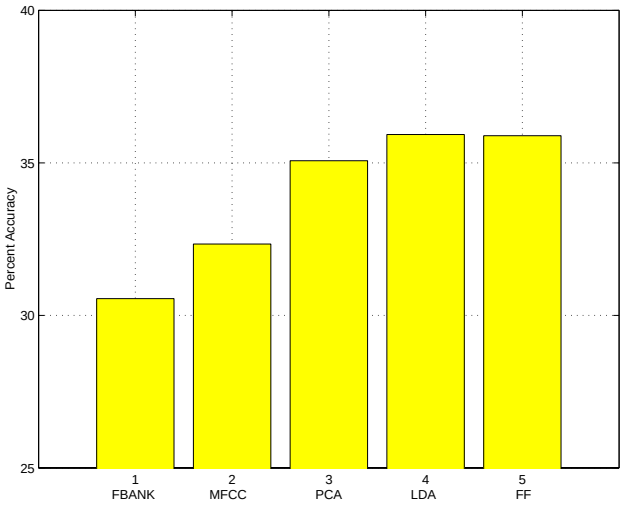

Figure 3. Accuracy using the projected coefficients without extra information

In figure 3 we can see the results using vectors of 16 coefficients. As it is shown, using FBANK parameters as the data set for the recognizer result in a very poor performance of the system. This is, in part, due to the correlation of its parameters. We can also see the reduction in accuracy produced by the approximation of the PCA analysis by the Discrete Cosine Transform (PCA vs MFCC).

From this first experiment we can conclude that the two discriminative approaches (FF and LDA) clearly outperform the non-discriminative techniques (MFCC and PCA).

3.2. Results using derivatives and energy information

One of the drawbacks of the HMM systems are the lack of the *history* information of the system, that is, the independence assumption between two frames of speech data. This is a consequence of the use of first order HMMs. With the purpose of bypass this drawback, we can add temporal information directly to the system as a part of the data set.

In the second block of the experiments we carried out, we use the temporal information in the form of derivatives of the frames (first and second order).

The results are divided into non-discriminative and discriminative techniques, in order to see, not only the effect of the decorrelation of the parameters, but also the discriminative effect of some of the approaches.

In figure 4 are shown the results for the non-discriminative techniques, i.e., MFCC and PCA.

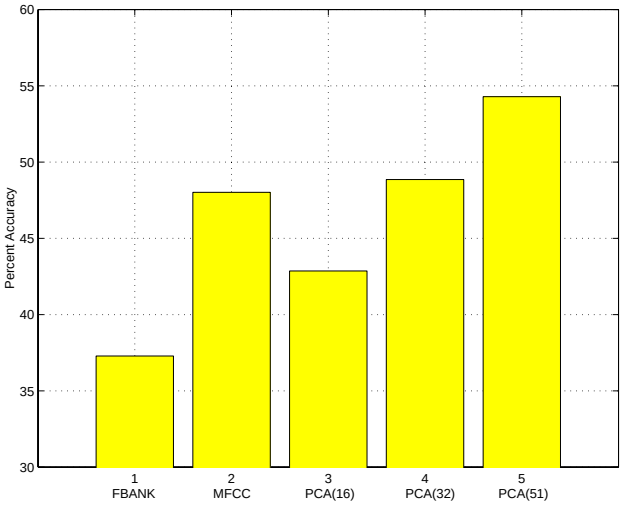

Figure 4. Accuracy using the projected coefficients, first and second derivatives and energy for non-discriminative techniques

Here we can see another advantage of the PCA analysis. With PCA, it is possible to reduce the dimension of a space even if it is non-homogeneous. By non-homogeneous we mean that a vector is composed by the FBANK parameters, its first and second derivatives, the energy of the speech frame and its derivatives. We calculated the PCA projection space for this global vector and we find the directions of maximum information. Then we can reduce the dimension of the vector maintaining the error rate. The results shown in figure 4 are for MFCC with 51 coefficients (16 for the MFCC, 16 for each of its derivatives, 1 for the energy and 1 more for each derivative), and for PCA with 16, 32 and 51 coefficients. As we can see, PCA with 32 coefficients outperforms MFCC with 51 coefficients.

In figure 5 we can see the accuracy of the discriminative approaches discussed in this paper.

LDA has the same ability we explained for PCA to deal with non-homogeneous vectors of mixed parameters and its derivatives. Again we tested the technique with full dimension and with a reduced space. For LDA we can see that we can reduce the dimension of the projected space more than we did for PCA and still outperform MFCC analysis.

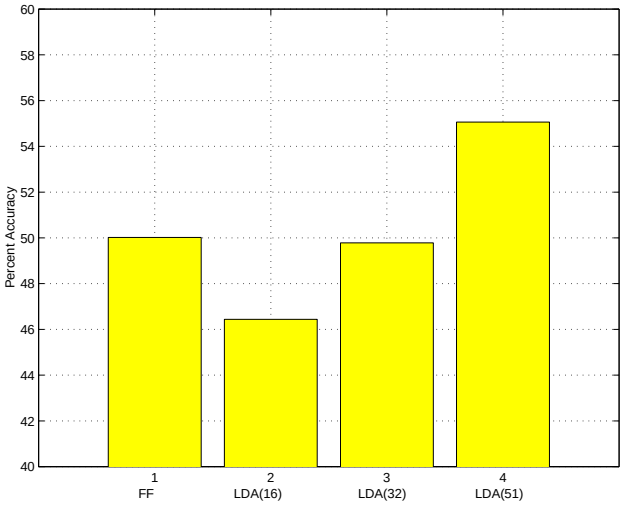

Figure 5. Accuracy using the projected coefficients, first and second derivatives and energy for discriminative techniques

4. CONCLUSIONS

In this paper we presented several techniques used in speech recognition to decorrelate the feature vector obtained from a filter-bank analysis. It was also shown that some of these approaches, not only decorrelates the parameters, but also perform some sort of decorrelation among classes, and thus reducing the error rate.

We also shown some of the advantages of using PCA and LDA analysis instead of DCT, like the possibility of using non-homogeneous vectors for the reduction of the working dimension.

Finally, we also demonstrated the power of the frequency-filtering analysis, which offers a significant reduction of the error rate while maintaining a low computational cost.

REFERENCES

[1] Duda, R.O., Hart, P.E.. *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York 1973.

[2] Fisher, W.M., Doddington, G.R.. "The DARPA Speech Recognition Research Database: Specification and Status".*Proc. DARPA Speech Recognition Workshop*, pp. 93-99, Palo Alto 1986.

[3] Hunt, M.J., Richardson, S.M., Bateman, D.C., Piau, A.. "An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination".*Proc. International Conference on Acoustics, Speech and Signal Processing*, pp. 881-884, Toronto 1991.

[4] Johnson, R.A., Wichern, D.W. . *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey 1992.

[5] Lee, K.F.. *Automatic Speech Recognition. The Development of the SPHINX System*, Kluwer Academic Publishers, Norwell, 1989.

[6] Leggetter, C.J.. *Improved Acoustic Modelling for HMMs Using Linear Transformations*, PhD Thesis, University of Cambridge, 1995.

[7] Nadeu, C., Mariño, J.B., Hernando, J., Nogueiras, A.. "Frequency and Time Filtering of Filter-Bank Energies for HMM speech recognition".*Proc. International Conference on Spoken Language Processing*, pp. 430-433, Philadelphia 1996.

[8] Ruggero, M.A., *The Mammalian Auditory Pathway: Neurophysiology*, chapter Physiology and Coding of Sound in the Auditory Nerve, Springer-Verlag, New York 1992.