# A DFE-BASED ALGORITHM FOR FEATURE SELECTION IN SPEECH RECOGNITION

*Ángel de la Torre, Antonio M. Peinado, Antonio J. Rubio, Victoria Sánchez*

Dpto. de Electrónica y Tecnología de Computadores
Universidad de Granada, 18071 GRANADA (Spain)
e-mail atv@hal.ugr.es tel: 34-58-243271 fax 34-58-243230

## ABSTRACT

The algorithms for the reduction of the number of features without degrading the performance of pattern recognition systems play an important role in real applications.

In this work a new algorithm for feature selection is proposed. This algorithm is based on the Discriminative Feature Extraction (DFE) technique and has been applied to speech recognition. The experimental results show that the recognition systems accept important reductions of the number of features without a degradation of the performance. For the representation used in our experiments, the recognition error-rate is not significantly increased when the number of components in the feature vector is reduced from 42 to 20.

## 1. INTRODUCTION

Signal representation is a crucial issue in the design of speech recognizers. The components of the feature vector determine the information used by the recognizer for classification. Usually, the application of a *feature space transformation* becomes a necessary step for obtaining an appropriate representation. In order to improve the system performance, the effect of the transformation must be the enhancement of the most discriminative features.

In the framework of speech recognition, the performance of the recognizers can be improved by the application of appropriate transformations. For example, the application of a *liftering window* can be interpreted as a diagonal linear transformation of the initial cepstral representation which enhances certain cepstral coefficients. Juang et al. [1] and Junqua et al. [2] have studied how the liftering window affects the recognizer performance. Recently a new method known as *Discriminative Feature Extraction* (DFE) has been proposed for computing feature space transformations. Biem and Katagiri have applied this method to the computation of liftering windows [3] and the design of filter banks [4]. Paliwal [5] proposed the simultaneous discriminative reestimation of both the transformation and the classifier. Torre et al. [6] [7] have proposed the estimation of the DFE transformation in a pre-training stage by using a simple classifier in order to obtain a proper algorithm convergence.

The DFE technique is a useful tool for including new features. When a new component is included into the feature vector, the DFE-trained transformation determines its contribution to the distance measure. If the new component is relevant for the classification, a large weight is applied in order to improve the recognizer performance. Otherwise, the assigned weight is smaller, so that the performance is not decreased.

Another aspect to be considered is the number of features. The inclusion of new components into the feature vector has two main effects:

- The recognition problem becomes more complex from a computational point of view, and the procedures for recognition could be impractical for some applications.
- The distance measure can be degraded, because non discriminant features could mask the contribution of the discriminant ones to the distance measure. In this case, the effect of adding new features would be an error-rate increment.

The second problem can be avoided by the DFE technique, because in the case of non discriminant new components, they are included with a low weight into the distance measure. However, the feature selection is very important for real applications, for which the computational load is the main restriction [8].

In this work an algorithm for feature selection is proposed. The algorithm, based on the DFE method, determines which is the least relevant component in the feature vector. The number of features can be reduced by iteratively removing the least discriminative component. In this way, the number of features can be reduced as much as necessary with a good performance/number of features trade-off. The proposed method is evaluated for a speaker independent isolated-word recognition task. The experimental results show the usefulness of the DFE-based feature selection algorithm.

## 2. DISCRIMINATIVE FEATURE EXTRACTION

The *Feature Extractor* (which will be assumed to be a linear transformation $V$) processes the input vector $\mathbf{x}$ and gives to the classifier a transformed vector $\mathbf{y} = V\mathbf{x}$. The basic idea of the DFE method is the computation of the $V$ transformation by using the *Minimum Classification Error* (MCE) criterion [9]. The elements of the transformation $v_{n,p}$ are iteratively trained by a gradient descent procedure in order to minimize a cost function $L$ which represents the classification error,

$$v_{n,p}^{k} = v_{n,p}^{k-1} - \eta \frac{\partial L}{\partial v_{n,p}} \tag{1}$$

where $\eta$ is the convergence coefficient. Let $\{X_1, \ldots, X_M\}$ be the set of training sequences and $\{\lambda_1, \ldots, \lambda_I\}$ the set of classes; the cost function can be defined as,

$$L = \sum_{m=1}^{M} l_m(X_m) \tag{2a}$$

$$l_m(X_m) = \frac{1}{1 + e^{-\alpha d_m(X_m)}} \quad (2b)$$

$$d_m(X_m) = -g_{k(m)} + \frac{1}{\beta} \log \left[ \frac{1}{I-1} \sum_{j \neq k(m)} e^{\beta g_j} \right] \quad (2c)$$

where $g_i = g_i(X_m, \lambda_i)$ are the *discriminant functions* (the recognized class is the one whose discriminant function is the largest one) and $\lambda_{k(m)}$ is the correct class for the considered sequence $X_m$. This way, $l_m \rightarrow 0$ for a clearly correct classification and $l_m \rightarrow 1$ for an incorrect classification ($l_m$ is a derivable and smoothed error function for $X_m$).

In order to compute $\partial L / \partial v_{n,p}$ it is necessary to know the discriminant functions, which are given by the definition of the classifier. The DFE technique provides a transformation of the feature space which improves the recognizer performance.

## 3. DFE-BASED FEATURE SELECTION

The DFE technique is a useful tool for the inclusion of new features. But a large number of features implies an increment of the computational complexity which could be unacceptable for real applications. For this reason, in order to exploit the advantages of the DFE method, a criterion for feature selection is necessary. The feature selection is possible without a degradation of the system performance by removing the least discriminative features.

Thus, our problem is the search of the least discriminative features. This way, it is possible to reduce the number of features as much as necessary with a good performance/number of features trade-off, by successively removing those features.

The DFE technique provides a mechanism that allows this search. Let us suppose a classifier whose discriminant functions can be written as,

$$g_i(X_m, \lambda_i) = \sum_{n=1}^{d} w_n h_{i,n}(X_m, \lambda_i) \quad (3)$$

where $d$ is the number of components and $h_{i,n}$ is a partial discriminant function for class $i$ which only includes information about the $n$-th component of the feature vector. A reduced cost function $\mathcal{L}_n$ can be defined (for every component $n = 1, \ldots, d$) similarly to the cost function $L$ defined in equations (2), but removing the contribution of the $n$-th component (by setting $w_n = 0$ in equation (3)). Therefore, the component to be removed (for Minimum Classification Error) is the one whose suppression produces the least increment of the cost function, that is, the $m$-th component which verifies,

$$\mathcal{L}_m < \mathcal{L}_n \quad \forall\ n \neq m \quad (4)$$

This method allows a one-by-one elimination of the least relevant components by using the DFE/MCE criterion. After the selection of the most discriminative ones, it is possible to estimate a DFE transformation for the reduced representation space in order to apply adequate weights to the remaining features. Finally, after the selection and the application of the transformation, the recognition system (as complex as necessary) can be trained by using the reduced and transformed feature vectors.

## 4. EXPERIMENTAL RESULTS

### 4.1. Recognition task and signal representation

The presented technique has been applied to a speaker independent isolated word recognition task (16 Spanish words vocabulary). Each feature vector is obtained from a 32ms speech frame. The initial vectors are composed of 20 cepstral coefficients (from LPC coefficients), 20 delta cepstral, the energy and the delta energy, which amounts to 42 components.

### 4.2. Feature selection

In order to obtain an adequate representation, it is necessary to apply a *liftering window* to the cepstral and delta cepstral coefficients [1] and to weight the different types of components [10]. Four different methods for the feature selection have been applied, in order to compare the approach we propose:

(a) Using a *raised-sine* liftering window [1] (**RS**). The highest order cepstral and delta cepstral coefficients are removed.

(b) Using a *statistically-weighted* liftering window [11] (**SW**). The highest order cepstral and delta cepstral coefficients are removed.

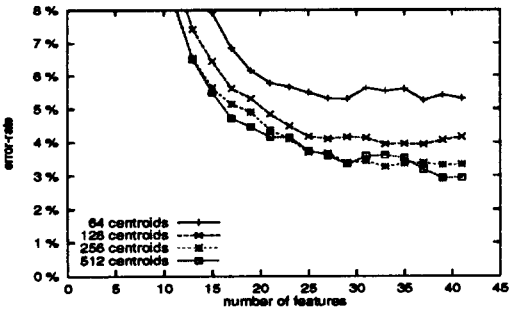
(c) The *minimum variance* features are removed (**MV-FS**). In this method, in order to reduce the number of features: (1) a DFE transformation is estimated; (2) the lowest variance features after the transformation (the ones whose contribution to the distance measure is less important) are removed; (3) a new DFE transformation is computed in the reduced feature space.

(d) DFE-based feature selection (**DFE-FS**). The features to be removed are selected by using the DFE-based criterion presented in the previous section.

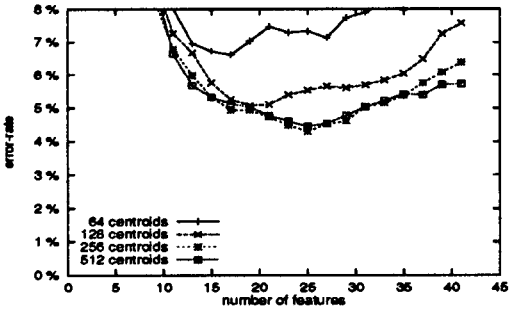
### 4.3. Recognition results

The four methods for reducing the number of features have been applied to two different variants of HMM-based speech recognition systems: Discrete Hidden Markov Model (DHMM) [12] and Multiple VQ Hidden Markov Model (MVQHMM) [13][14]. The DHMM system has been implemented for codebook sizes of 64, 128, 256 and 512 centroids, and the MVQHMM one, for codebook sizes of 4, 8, 16 and 32 centroids per class. Figures 1 and 2 represent the error-rate versus the number of features for both DHMM and MVQHMM recognition systems.

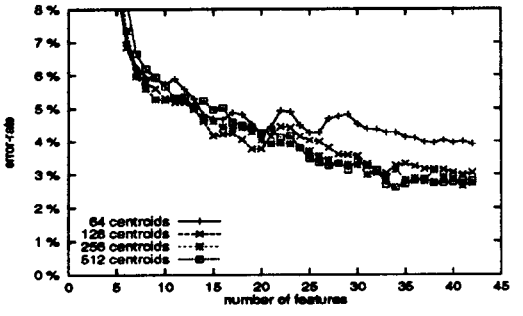
The recognition results show that the performance of the systems strongly depends on the representation of the speech signal. The DFE-based representations (**MV-FS** and **DFE-FS**, figures 1c, 1d, 2c and 2d) improve the performance with respect to the raised-sine and statistically-weighted liftering windows (**RS** and **SW**, figures 1a, 1b, 2a and 2b). The DFE-based feature selection (figures 1d and 2d) provides the best performance/number of features trade-off for both DHMM and MVQHMM recognition systems, and for the different codebook sizes. In the case of a DFE-based feature selection, the number of features can be reduced from 42 to 20 without a significant degradation of the system performance (the reduction plots remain almost flat between both numbers of features).



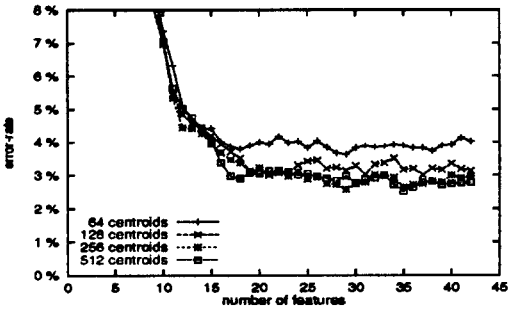

(a) **RS**: raised-sine liftering window




(b) **SW**: statistically-weighted liftering window

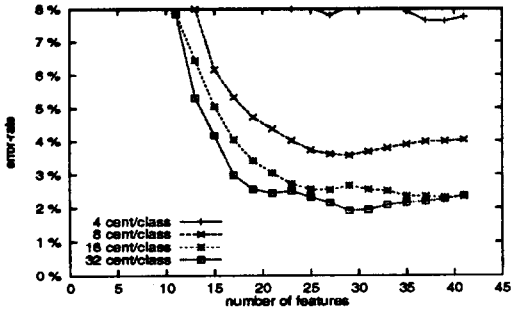



(c) **MV-FS**: selection based on minimum variance

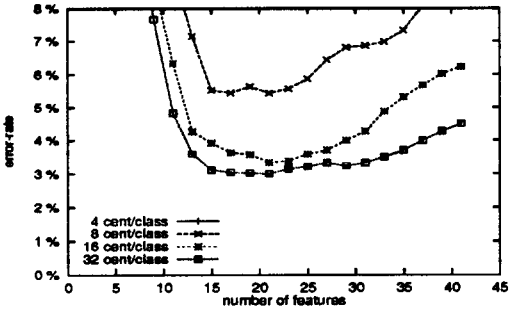



(d) **DFE-FS**: selection based on DFE

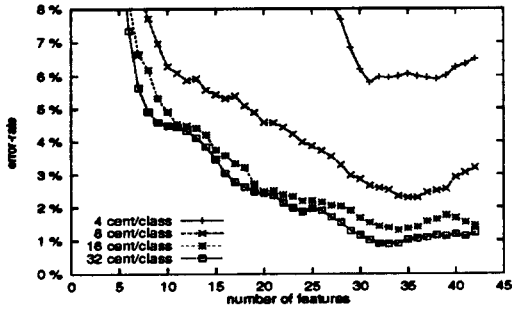
Figure 1: ***DHMM recognition results: error-rate versus number of features in the representation space***



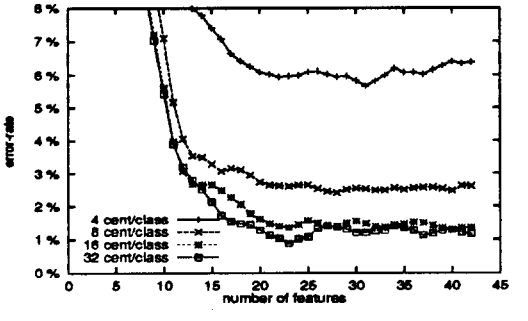

(a) **RS**: raised-sine liftering window




(b) **SW**: statistically-weighted liftering window




(c) **MV-FS**: selection based on minimum variance




(d) **DFE-FS**: selection based on DFE

Figure 2: ***MVQHMM recognition results: error-rate versus number of features in the representation space***

### 4.4. The role of the components in the feature vector

The role of a component in the feature vector is determined by two facts: (a) whether the selection procedure removes it or not and (b) its standard deviation after the DFE procedure (if it is not removed).

Figure 3 represents a selection score provided by the DFE-based selection algorithm. The first 20 components are the cepstral coefficients, the next 20 are the delta-cepstral ones, and the last two are the energy and delta energy coefficients. The lowest score features are removed first. For example, if 25% of the components are removed, the remaining components are the ones above the 25% line. As it can be observed, the high order cepstral coefficients are the first to be removed. The low order cepstral coefficients are the most important for a very reduced representation.

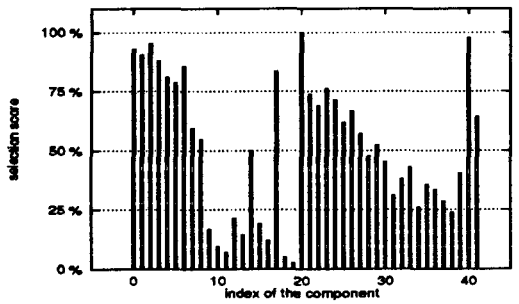



Figure 3: *Selection score provided by the DFE-FS algorithm*

The standard deviation of the components (after the application of a DFE transformation) is represented in Figure 4. In this case, the DFE transformation has been computed when no component is removed. The standard deviation represent the average contribution of each component to the distance measure used by the classifier.

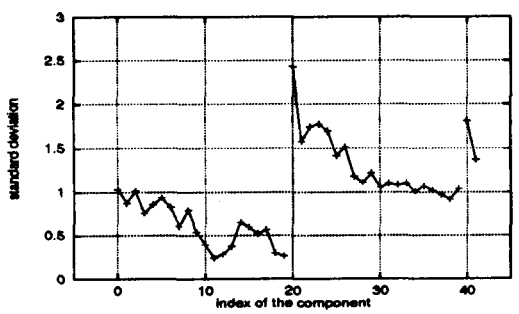



Figure 4: *Standard deviation of the transformed components*

## 5. CONCLUSIONS

The feature selection methods are very important for real applications. This work presents a new method for feature selection, based on the Discriminative Feature Extraction (DFE) technique. In the proposed method the least discriminative components of the feature vector are removed by using the Minimum Classification Error (MCE) criterion. After the feature selection, a DFE transformation of the reduced space is computed. The proposed feature selection method has shown a good behavior for a speaker-independent isolated word recognition task. From an initial representation space with 42 components, the number of features can be reduced to 20 without a significant increment of the error-rate.

## 6. REFERENCES


[1] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Trans. on ASSP*, vol. 35, pp. 947–954, July 1987.

[2] J. Junqua and H. Wakita, "A Comparative Study of Cepstral Lifters And Distance Measures for All Pole Models of Speech in Noise," in *Proc. of ICASSP-89*, pp. 476–479, 1989.

[3] A. Biem and S. Katagiri, "Feature extraction based on Minimum Classification Error/Generalized Probabilistic Descent method," in *Proc. of ICASSP '93*, vol. 2, pp. 275–278, 1993.

[4] A. Biem and S. Katagiri, "Filter bank design based on discriminative feature extraction," in *Proc. of ICASSP '94*, vol. 1, pp. 485–488, 1994.

[5] K. K. Paliwal, M. Bacchiani, and Y. Sagisaka, "Minimum Classification Error training algorithm for feature extractor and pattern classifier in speech recognition," in *Proc. of EUROSPEECH '95*, vol. 1, pp. 541–544, 1995.

[6] A. de la Torre, A. M. Peinado, A. J. Rubio, J. C. Segura, and V. E. Sánchez, "Minimum Classification Error Transformations for Improving Speech Recognition Systems," in *EUSIPCO-96*, 1996.

[7] A. de la Torre, A. M. Peinado, A. J. Rubio, V. E. Sánchez, and J. E. Díaz, "An application of Minimum Classification Error to feature space transformations for Speech Recognition," *Speech Communication*, In Press.

[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[9] B. Juang and S. Katagiri, "Discriminative Learning for Minimum Error Classification," *IEEE Trans. on Signal Processing*, vol. 40, pp. 3043–3054, Dec. 1992.

[10] A. Peinado, P. Ramesh, and D. Roe, "On the Use of Energy Information for Speech Recognition Using HMM," in *Proceedings of EUSIPCO-90*, vol. 2, (Barcelona), pp. 1243–1246, Sept. 1990.

[11] Y. Tohkura, "A weighted cepstral distance measure for speech recognition," *IEEE Trans. on ASSP*, vol. 35, no. 10, pp. 1414–1422, Oct. 1987.

[12] L. R. Rabiner, J. G. Wilpon, and F. K. Soong, "High Performance Connected Digit Recognition Using Hidden Markov Models," *IEEE Trans. on ASSP*, vol. 37, pp. 1214–1225, Aug. 1989.

[13] J. Segura, A. Rubio, A. Peinado, P. García, and R. Román, "Multiple VQ Hidden Markov Modelling for Speech Recognition," *Speech Communication*, vol. 14, pp. 163–170, April 1994.

[14] A. Peinado, J. Segura, A. Rubio, P. Garcia, and J. Pérez, "Discriminative codebook design using Multiple Vector Quantization in HMM-based speech recognizers," *IEEE Trans on Speech and Audio Processing*, vol. 4, pp. 88–95, Mar. 1996.