# MINIMUM CLASSIFICATION ERROR TRANSFORMATIONS FOR IMPROVING SPEECH RECOGNITION SYSTEMS

Ángel de la Torre, Antonio M. Peinado, Antonio J. Rubio, José C. Segura, Victoria E. Sánchez
e-mail atv@hal.ugr.es
Dpto. de Electrónica y Tecnología de Computadores
Universidad de Granada, 18071 GRANADA (Spain)

## ABSTRACT

Signal representation is an important aspect to be taken into account for pattern classification. Recently, discriminative training methods have been applied to feature extraction for speech recognition. In this paper, we apply the Minimum Classification Error estimation to train the parameters of a feature extractor. This feature extractor is a linear transformation of the original representation space. The new representation of the speech signal makes easier the recognition task and the performance of the different tested recognizers is improved as the experimental results show.

## 1 INTRODUCTION

Feature extraction is very important for the speech recognizer design. An appropriate representation of the patterns simplifies the recognition task, and this could improve the recognizer performance. The most widely used speech signal representations are based on considerations such as speech production models, auditory models, phonetic or acoustic considerations, etc. So, after years of research about this, cepstrum-LPC or bank-of-filter based representations are nowadays very utilized.

However, the use of signal *transformations* for enhancing the most discriminative features is a necessary step. The comparison of two speech signals (a test signal and a reference one) is performed by using a distance measure in the representation space. A linear transformation of the feature space modifies the distance measure and, therefore, the performance of the recognizers. The application of a *liftering window* can be interpreted as a diagonal transformation of the cepstral vectors which represent the speech signal. Juang et al. [1] have studied the cepstrum-LPC representation and the importance of applying a liftering window to enhance the most discriminant components of the cepstral vector. The importance of this operation is shown by two well-known results: first, the performance of speech recognizers is very sensitive to the lifter and second, the ideal lifter depends strongly on the noise conditions [2].

Recently, a feature extractor design based on the Minimum Classification Error (MCE) estimation [3] has been proposed. This method of selecting an adequate representation space is called Discriminative Feature Extraction (DFE). Biem and Katagiri have applied DFE to compute liftering windows [4] and to design a filter bank [5]. Bacchiani and Aikawa have optimized the parameters of a dynamic cepstrum lifter array [6] and Paliwal et al. have proposed the simultaneous training of the feature extractor and the pattern classifier [7]. In all the cases, DFE has been shown to be a powerful tool for error-rate reduction. The feature extractor usually consists of a transformation which is applied to the original feature space. The application of DFE has been studied for different input feature spaces and different restrictions imposed to the transformation.

The DFE strategy presents some problems related to the MCE training of the feature extractor:

- If the models for the speech units are complex, the training process and the estimated classifier depend strongly on the distance measure. This makes necessary the simultaneous reestimation of both, the feature extractor and the classifier.
- This scheme of DFE implies an alteration of the training procedure for the classifier, because it forces the use of the MCE criterion for the classifier training.
- The simultaneous MCE reestimation of classifier and transformation is not advisable for a complex recognizer, due to the fact that obtaining an error-rate minimization modifying the classifier is easier than modifying the feature extractor (because a modification of the feature extractor affects the whole classifier).
- The importance of the initialization is increased as the classifier or the recognition task are more complex because of the complexity of the error function (the MCE algorithm looks for the nearest local minimum).

These problems limit the applicability of DFE, since as the recognition task or the classifier are more complex the improvements are less significative.

In this work we propose a new variant of the DFE strategy in order to avoid these problems. We propose training the feature extractor, which is a linear transformation, by using a very simple classifier, and then performing the training of the definitive classifier using the transformed vectors. So, the procedures for training the classifier are not modified. Due to the simplicity of the classifier used for the DFE training, the error-rate can only be minimized by enhancing the most discriminative features and by including the fewest discriminative ones using the appropriate weights. Then, the new Euclidean distance measure is adapted to the recognition task and conditions. This improves the recognizer because all the training process for the definitive classifier are performed in an optimized feature space from a discriminative point of view.

In order to evaluate the proposed technique we have developed several isolated-word speaker independent experiments. The recognition results when the proposed transformations are applied are compared to the results obtained by the application of a standard pre-processing technique. The results show the usefulness of the proposed technique to improve the speech recognizer performance.

## 2 DISTANCES AND TRANSFORMATIONS

The *cepstrum-LPC* is a widely used representation for speech recognition. Some authors have proposed the inclusion new features in addition to *cepstrum* such as *energy*, or dynamic features as *delta cepstrum* or *delta energy* [8]. A way to incorporate the different features into the recognition system is the use of a *Multi-feature Weighted Distance Measure* (MWDM) [9],

$$d(\mathbf{x}_t, \mathbf{x}_r) = p_c d_c(\tilde{\mathbf{c}}_t, \tilde{\mathbf{c}}_r) + p_{\Delta c} d_{\Delta c}(\Delta\tilde{\mathbf{c}}_t, \Delta\tilde{\mathbf{c}}_r) + p_E d_E(E_t, E_r) + p_{\Delta E} d_{\Delta E}(\Delta E_t, \Delta E_r) \quad (1)$$

where $\mathbf{x}_t$ and $\mathbf{x}_r$ are the test and reference feature vectors, respectively, $d_c$, $d_{\Delta c}$, $d_E$ and $d_{\Delta E}$ are Euclidean distances for the liftered cepstrum and delta cepstrum vectors, the energy and the delta energy, respectively and $p_c$, $p_{\Delta c}$, $p_E$ and $p_{\Delta E}$ are weights that must be experimentally determined. The liftering is applied to the cepstral coefficients $c(n)$ as,

$$\tilde{c}(n) = c(n)w(n) \quad (n = 1, \ldots, L) \quad (2)$$

where $w(n)$ is the liftering window and $L$ the window length. The delta cepstral coefficients are liftered like the cepstral ones, by using the same liftering window.

Thus, the distance measure obtained after the liftering and weighting is the Euclidean distance between the vectors transformed by a matrix $V$,

$$d(\mathbf{x}_t, \mathbf{x}_r) = \|\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_r\|^2 \quad \text{where} \quad \tilde{\mathbf{x}} = V\mathbf{x} \quad (3)$$

The $V$ matrix is diagonal and its form is,

$$v_{n,p} = 0 \quad \text{if} \quad n \neq p$$

$$v_{n,n} = \begin{cases} \sqrt{p_c} \cdot w(n) & \text{if} \quad n = 1, \ldots, L \\ \sqrt{p_{\Delta c}} \cdot w(n-L) & \text{if} \quad n = L+1, \ldots, 2L \\ \sqrt{p_E} & \text{if} \quad n = 2L+1 \\ \sqrt{p_{\Delta E}} & \text{if} \quad n = 2L+2 \end{cases} \quad (4)$$

The determination of an adequate liftering window and weights (or the $V$ matrix) constitutes an important problem because they determine how the different features are incorporated into the distance measure [1], [2], [8], [9]. The performance of a recognition system could be improved by the application of an adequate transformation to the feature vectors.

## 3 COMPUTING THE MCE TRANSFORMATION

Our goal is to obtain a linear transformation $V$ of the representation space adapted to the recognition task and acquisition conditions in order to improve the recognizer performance. According to the DFE strategy, the $V$ matrix is trained by using the MCE criterion. The elements $v_{n,p}$ are computed iteratively to minimize a cost function $L$ that represents the classification error. At iteration $k$, $v_{n,p}$ is computed by gradient descent of the cost function,

$$v_{n,p}^k = v_{n,p}^{k-1} - \eta \frac{\partial L}{\partial v_{n,p}} \quad (5)$$

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

$$L = \sum_{m=1}^{M} l_m(X_m) \quad (6a)$$

$$l_m(X_m) = \frac{1}{1 + e^{-\alpha d_m(X_m)}} \quad (6b)$$

$$d_m(X_m) = -g_{k(m)} + \frac{1}{\beta} \log \left[ \frac{1}{I-1} \sum_{j \neq k(m)} e^{\beta g_j} \right] \quad (6c)$$

where $g_i = g_i(X_m, \lambda_i)$ are the *discriminant functions* (the recognized class is the one whose discriminant function is the greatest); $\lambda_{k(m)}$ is the correct class for the sequence $X_m$. Thus, $d_m < 0$ (and $l_m \to 0$) if the classification is clearly correct, and $d_m > 0$ (and $l_m \to 1$) if clearly incorrect ($l_m$ is a smooth and derivable classification error function for sequence $X_m$). Factor $\beta$ determines the contribution of the incorrect classes to $d_m$ and $\alpha$ is the transition parameter from correct to incorrect classification. In order to compute the partials $\partial L / \partial v_{n,p}$, it is necessary to know the form of the discriminant functions $g_i$.

The conventional DFE strategy proposes to use the $g_i$ functions of the classifier used for recognition. Using a complex classifier, the cost function is minimized by the adaptation of the distance measure to this classifier. This way of estimating the classifier and the transformation is not optimal because the classifier is trained in the original (not optimized) representation space, and the training process strongly depends on the distance measure (i.e. if a clustering process is performed). Moreover, in the case of a simultaneous reestimation of both the classifier and the transformation, the problems discussed in the introduction lead to a non optimal transformation. We propose another way of computing the $V$ transformation:

- A very simple classifier (with no clustering process for its training) is used for the MCE estimation of the transformation. Using a simple classifier for the transformation estimation, the cost function could only be minimized by applying a transformation which leads to an optimal (from a discriminative point of view) distance measure. Moreover, using a simple classifier the problems related to the MCE estimation of the transformation are minimized. The obtained transformation is not adapted to the classifier and then it could be successfully applied to a more complex classifier.

- The transformation and the definitive classifier are independently trained. Thus, all the training processes for the definitive classifier are performed in the new optimized representation space.

The classifier we propose for the MCE estimation of the transformation models the production of the feature vectors

that belong to each class $\lambda_i$ by one spherical Gaussian probability density function,

$$P(\tilde{\mathbf{x}}|\lambda_i) = \frac{1}{(2\pi\sigma_i^2)^{d/2}} \exp\left(-\frac{1}{2}\frac{\|\tilde{\mathbf{x}} - \tilde{\mathbf{y}}_i\|^2}{\sigma_i^2}\right) \quad (7a)$$

$$\tilde{\mathbf{y}}_i = E_i[\tilde{\mathbf{x}}] \quad \sigma_i^2 = \frac{1}{d} E_i[\|\tilde{\mathbf{x}} - \tilde{\mathbf{y}}_i\|^2] \quad (7b)$$

where $d$ is the dimensionality of the representation space, and $E_i[\cdot]$ means average over all the vectors that belong to the class $\lambda_i$. According to this model, for a given sequence $X_m = \mathbf{x}_1, \ldots, \mathbf{x}_T$, the discriminant functions are constructed as,

$$g_i(X_m|\lambda_i) = \log P(X_m|\lambda_i) = \frac{1}{T}\sum_{t=1}^{T} \log P(\tilde{\mathbf{x}}_t|\lambda_i) \quad (8)$$

From the definition of the cost function, and taking into account that $\tilde{\mathbf{x}} = V\mathbf{x}$, it is possible to compute $\partial L/\partial v_{n,p}$, and this allows the iterative estimation of the transformation by using equation (5). Using this classifier for the MCE algorithm we obtain a transformation of the representation space where the Euclidean distance between vectors that belong to the different classes is maximized. The new representation makes the recognition task easier, independently of the recognizer configuration.

## 4 EXPERIMENTS AND RESULTS

We have developed several isolated-word speaker independent recognition experiments. The vocabulary is composed of 16 words (10 Spanish digits and 6 keywords) and the data base is composed of 3 repetitions of every word recorded from 40 speakers (20 men and 20 women). The speech signal has been sampled (sample frequency $f_s = 8kHz$) and segmented into frames of 32ms, overlapped 16ms. We have computed 14 cepstral coefficients (from 10 LPC coefficients), 14 delta cepstral ones, energy and delta energy for every frame of speech.

Taking into account the small correlation between these features, we have simplified the problem with the restriction that $V$ is a diagonal matrix. Thus, we have only computed the elements in the main diagonal $v_{n,n} = v_n$. We have obtained two MCE transformations, labelled *MCE-1* and *MCE-2*, from two different initializations. For the second initialization, a higher weight has been applied to the delta cepstral coefficients. The recognition results when the MCE transformations are applied are compared with the results of a standard pre-processing (given by equation (4)), using a raised-sine liftering window [1], and computing the experimental weight as proposed in [9]. The experiments for the standard transformation are labeled as *MWDM*.

Figure 1 shows the $v_n$ weights for the three transformations we have applied. The first 14 weights are applied to cepstrum, the next 14 ones to delta cepstrum, and the last 2 ones, to the energy and delta energy. The weights for delta cepstrum are significatively greater than the weights of the cepstrum because of the differences between their variances (a normalization before computing the MWDM or MCE transformations is necessary).

The resulting liftering windows from MCE transformations are very similar to raised-sine for the lower cepstral coefficients. Important differences can be observed in the

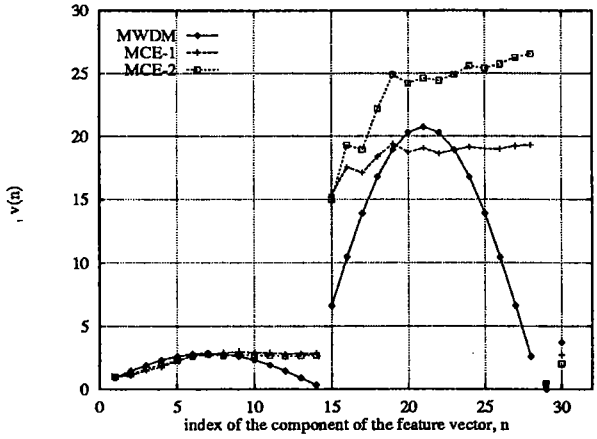


Figure 1: *MWDM, MCE-1 and MCE-2 transformations.*

delta cepstral liftering window. Delta cepstral weights are greater for MCE-2 due to the initialization.

We have developed the recognition experiments using two types of HMM-based speech recognizers: Discrete Hidden Markov Models (*DHMM*) [10] and Multiple Vector Quantization Hidden Markov Models (*MVQHMM*) [11]. These recognizers have been tested for different codebook sizes. Four DHMM recognizers have been implemented, using 64, 128, 256 and 512 centroids. Since there are 16 words in the vocabulary and MVQHMM uses independent codebooks for every class, in this case the experiments have been developed using 4, 8, 16 and 32 centroids per word, in order to compare this experiments to the DHMM ones.

The recognition results for the DHMM system are shown in figure 2 (error-rate versus codebook size). Figure 3 shows the error-rate for the MVQHMM system, for the different transformations.

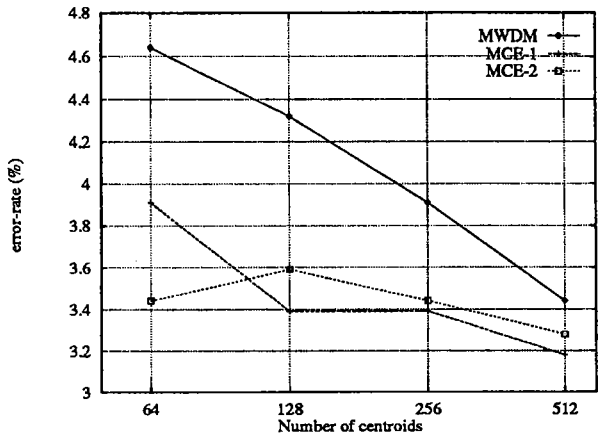


Figure 2: *DHMM recognition error-rate versus codebook size of the VQ codebook, by applying the MWDM, MCE-1 and MCE-2 transformations.*

The recognition results suggest the following comments:

1. Both MCE transformations lead to significative improvements of the recognizer performance with respect to MWDM, for both types of recognizers, DHMM and MVQHMM.
2. The improvements are observed for all the codebook

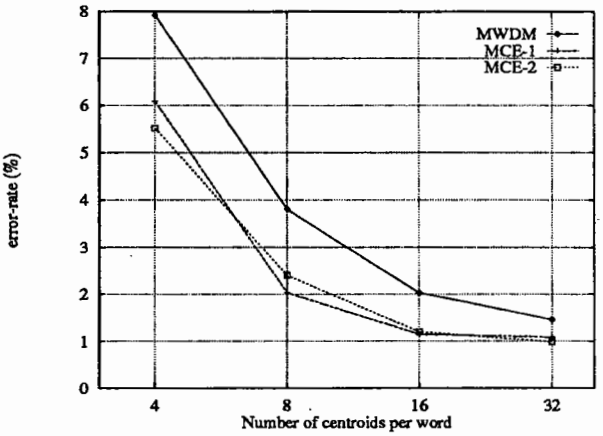


Figure 3: *MVQHMM recognition error-rate versus codebook size of the VQ codebook by applying the MWDM, MCE-1 and MCE-2 transformations.*

sizes. Even though the MCE transformations are computed using a very simple classifier, they lead to improvements independently of the complexity of the definitive classifier.

3. For the MVQHMM recognizer, the classification score is a composition of a quantization score (provided by the codebooks of each class) and a generation score (provided by the HMM's)[11]. The MCE representation (where the distance between the vectors belonging to different classes is maximized) improves directly the quantization score. For this reason, the improvements are more important for the MVQHMM recognizer.

4. Since the transformation is applied at the beginning of the classifier training procedure (in contrast to conventional DFE), all the classifier training steps are performed in a more discriminative space. I.e. obtained clustering is better adapted for the discrimination than the one obtained using the MWDM transformation. Thus, the performance is improved by application of the MCE transformations even in the case that the quantization score is not used (as in DHMMs).

5. The MWDM used as reference is the result of a complex process of selection: a wide set of liftering windows has been tested and the estimation of the MWDM weights has required a lot of recognition experiments. The estimation of the MCE transformation does not imply a high computational cost because of the use of a simple classifier for its training.

6. The proposed technique does not increase the computational cost of the recognition systems, and does not modify the training process, since the MCE transformation is computed in a pre-training stage.

7. The obtained MCE transformations suggest that using the same liftering window for cepstrum and delta cepstrum is not optimal for speech recognition.

## 5 CONCLUSIONS

The MCE algorithm has been applied to compute transformations of the feature space for improving the speech recognizers. We have proposed a new variant of DFE that consists of computing a MCE-estimated transformation in a pre-training stage by using a very simple classifier. Independent training of both the transformation and the definitive classifier presents some advantages with respect to conventional DFE. This method has been successfully applied to speaker-independent isolated word recognition. The application of the MCE transformations improves the performance of the recognizers with respect to standard pre-processing techniques, for all the tested configurations.

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

[4] 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.

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

[6] M. Bacchiani and K. Aikawa, "Optimization of time-frequency masking filters using the Minimum Classification Error criterion," in *Proc. of ICASSP '94*, vol. 2, pp. 197–200, 1994.

[7] 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.

[8] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," *IEEE Trans. on ASSP*, vol. 34, pp. 52–59, Feb. 1986.

[9] 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.

[10] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in *Proceedings of the IEEE*, vol. 77, pp. 257–285, Feb. 1989.

[11] 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.