ELSEVIER

Speech Communication 38 (2002) 267–286

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Discriminative feature weighting for HMM-based continuous speech recognizers ☆

Ángel de la Torre *, Antonio M. Peinado, Antonio J. Rubio, José C. Segura, Carmen Benítez

Departamento Electrónica y Tecnología de Computadores, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain

Received 5 July 2000; received in revised form 23 April 2001; accepted 19 June 2001

Abstract

The Discriminative Feature Extraction (DFE) method provides an appropriate formalism for the design of the front-end feature extraction module in pattern classification systems. In the recent years, this formalism has been successfully applied to different speech recognition problems, like classification of vowels, classification of phonemes or isolated word recognition. The DFE formalism can be applied to weight the contribution of the components in the feature vector. This variant of DFE, that we call Discriminative Feature Weighting (DFW), improves the pattern classification systems by enhancing those components more relevant for the discrimination among the different classes. This paper is dedicated to the application of the DFW formalism to Continuous Speech Recognizers (CSR) based on Hidden Markov Models (HMMs). Two different types of HMM-based speech recognizers are considered: recognizers based on Discrete-HMMs (DHMMs) (for which the acoustic evaluation is based on an Euclidean distance measure) and Semi-Continuous-HMMs (SCHMMs) (for which the acoustic evaluation is performed making use of a mixture of multivariated Gaussians). We report how the components can be weighted and how the weights can be discriminatively trained and applied to the speech recognizers. We present recognition results for several continuous speech recognition tasks. The experimental results show the utility of DFW for HMM-based continuous speech recognizers.

Zusammenfassung

Die Methode der diskriminativen Merkmalextraktion (Discriminative Feature Extraction, DFE) stellt einen für den Entwurf des Eingangsmerkmalextraktionsmoduls in Musterklassifizierungssystemen geeigneten Algorithmus zur Verfügung. In vergangenen Jahren wurde dieser Algorithmus erfolgreich auf verschiedene Spracherkennungsprobleme wie die Klassifizierung von Selbstlauten, Phonemklassifizierung, oder die Erkennung isolierter Wörter angewandt. Der DFE Algorithmus kann zur Gewichtung des Beitrags der Komponenten des Merkmalvektors verwendet werden. Diese Variante der DFE, die wir diskriminative Merkmalgewichtung (Discriminative Feature Weighting, DFW) nennen, verbessert Musterklassifizierungssysteme, indem sie jene Komponenten verstärkt, die mehr Relevanz bei der Unterscheidung verschiedener Klassen haben. Diese Veröffentlichung widmet sich der Anwendung der diskriminativen Merkmalgewichtung auf Systeme zur Erkennung kontinuierlicher Sprache, die auf "Hidden Markov Models" (HMM) beruhen. Es werden zwei verschiedene Arten von HMM-basierten Spracherkennungssystemen betrachtet: solche, die

☆ This work has been partially supported by the "Plan Propio de Investigación 1999–2000" of the University of Granada.
* Corresponding author. Tel.: +34-58-243271; fax: +34-58-243230.
E-mail address: atv@ugr.es (Á. de la Torre).

0167-6393/01/$ - see front matter
PII: S0167-6393(01)00068-1

auf diskreten HMMs basieren (für die die akustische Auswertung auf einem Euklidischen Abstandsmaßberuht) und semikontinuierliche HMMs (für die zur akustischen Auswertung eine Mischung von multivarianten Gauß-Verteilungen verwendet wird). Wir berichten, wie die Komponenten gewichtet werden können und wie die Gewichte diskriminativ trainiert und auf Spracherkennungssysteme angewendet werden können. Wir präsentieren Ergebnisse für einige Aufgaben der Erkennung kontinuierlicher Sprache. Die experimentellen Ergebnisse zeigen die Nützlichkeit der diskriminativen Merkmalgewichtung in HMM-basierten Erkennungssystemen für kontinuierliche Sprache.

Résumé

La méthode DFE (Extraction Discriminante de Paramètres) fournit un formalisme adéquat pour la conception d'un module d'extraction de paramètres pour un système de classification de formes. Au cours des dernières années, cette méthode a été appliquée avec succès à différents problèmes en reconnaissance de la parole tels que la classification de voyelles et de phonèmes ou la reconnaissance de mots isolés. Le formalisme DFE peut être utilisé pour pondérer les contributions des différentes composantes d'un vecteur de paramètres. Cette variante de DFE, que nous appelons DFW (Pondération Discriminante de paramètres), améliore un système de classification de formes en favorisant les composantes assurant la meilleure discrimination interclasses. Cet article est consacré à l'application du formalisme DFW à la reconnaissance de la parole continue par modèles de Markov cachés (HMM). Deux types différents de reconnaisseurs sont étudiés: ceux fondés sur des HMM discrets (utilisant une distance euclidienne) et ceux fondés sur des HMM semi-continus (utilisant des mélanges de gaussiennes). Nous montrons comment les composantes peuvent être pondérées et comment les poids peuvent être appris de façon discriminante. Des résultats expérimentaux sont fournis pour différentes tâches de parole continue. Ces résultats montent l'intérêt du formalisme en reconnaissance de parole continue par HMM.

Keywords: Continuous speech recognition; Discriminative feature extraction; Error-rate; Cost function; Probability density function; Minimum classification error; Hidden Markov model; Discriminative feature weighting; Discriminative weighting by transformation; Partial probability weighting

1. Introduction

Most practical speech recognition systems consist of two modules: the front-end feature extraction module and the back-end classification module (Fig. 1). The classification module is usually statistically designed according to Bayes decision theory and the design of the feature extractor is conventionally based on scientific experience and heuristics. The design of the feature extraction module is a relevant aspect for the performance of the speech recognizer because this module extracts the discriminative information utilized by the classification module to perform the recognition. It is evident that achieving accurate recognition requires a careful design of the feature extraction module (Duda and Hart, 1973).

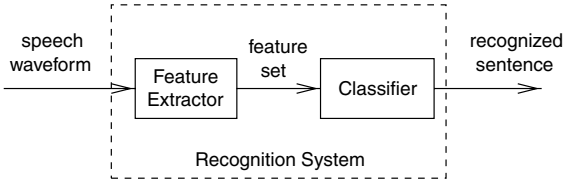

Fig. 1. General scheme of a speech recognition system.

In the early 1990s, design methods based on the discriminative training approach were proposed to adjust the classifier parameters. The Minimum Classification Error/Generalized Probabilistic Descent (MCE/GPD) training method (Juang and Katagiri, 1992) has been successfully applied to speech recognition systems (Chou et al., 1992; McDermott and Katagiri, 1994; Peinado et al., 1995, 1996). More recently, the MCE/GPD method has also been utilized to adjust the parameters of the feature extractor. This approach, called Discriminative Feature Extraction (DFE) (Biem and Katagiri, 1993, 1997), introduced a new

concept in the field of speech recognition: the extension of MCE/GPD method to the feature extractor allows to design it taking into account the main goal of pattern recognition, that is, the recognition accuracy.

DFE has been applied to train several elements in the feature extraction module. For example, Biem and Katagiri have utilized DFE to compute cepstral liftering windows (Biem and Katagiri, 1993, 1997), and to design filter banks (Biem and Katagiri, 1994, 1995, 1997). Bacchiani and Aikawa (1994) have optimized the parameters of a dynamic cepstrum lifter array. In several approaches, the element trained by DFE is a linear transformation of the original feature space (Paliwal et al., 1995; de la Torre et al., 1996a,b, 1997), and Watanabe et al. (1995, 1997) have extended the concept of DFE to a class-dependent feature extractor. The utility of DFE is demonstrated by the results of several speech recognition experiments reported in the bibliography. The application of DFE has improved the accuracy in recognition problems such as classification of vowels, classification of stop consonants, classification of phoneme units or isolated word recognition.

The DFE formalism has been applied to weight the contribution of the different components in the feature vector to the acoustic evaluation. That is the case of the implementations of DFE in which the element discriminatively trained is a diagonal transformation of the feature space (Biem and Katagiri, 1993, 1997; de la Torre et al., 1996a,b). This variant of the DFE method, that will be called Discriminative Feature Weighting (DFW), improves the performance of the speech recognizers by enhancing those components more important for the classification, i.e., the components containing more discriminative information.

In this paper, we investigate the application of DFW to Continuous Speech Recognition (CSR) systems. Since the use of continuous speech recognizers based on Hidden Markov Models (HMM) is widely extended (Rabiner and Juang, 1993; Young, 1996), in this paper our interest is focused on the application of DFW formalism to HMM-based CSR systems. Two different types of HMM-based speech recognizers are considered in this paper: Discrete-HMM (DHMM) and Semi-Continuous-HMM (SCHMM) recognizers (Rabiner and Juang, 1993; Huang et al., 1990).

In the case of DHMM-based speech recognizers, the acoustic evaluation is based on an Euclidean distance measure. For this type of recognizer, the components are weighted by applying a diagonal transformation of the feature space that is discriminatively trained according to the DFE formalism. The effect of this transformation is a new representation space where those components more relevant for the discrimination are enhanced (the directions associated to those components are expanded). This transformation can improve the accuracy of the recognizer since the contribution of each component to the distance measure is adjusted by the discriminative training procedure taking into account its discriminative capability.

Nowadays, those speech recognizers making use of a mixture of multivariated Gaussians for the acoustic evaluation (like Continuous-HMM or SCHMM recognizers (Rabiner and Juang, 1993; Young, 1996, 1997)) are utilized more frequently than DHMM-based speech recognizers. For those recognizers, the recognition performance remains invariant to the application of feature space transformations. This behavior (commented in (Young et al., 1997)) takes place because when a transformation is applied, the modification of the partial distance associated to each component is compensated by the modification of the elements of the covariance matrices in the Gaussian probability density functions (see Section 3.2.1). For this kind of recognizers, the contribution of the different components to the acoustic evaluation can be weighted by applying exponents to the partial probabilities corresponding to each component. Under this approach, which will be referred as Partial Probability Weighting (PPW), we propose the application of the DFW formalism to the estimation of these exponents.

In this paper, we investigate the application of DFW method to both categories of HMM-based continuous speech recognizers. The element adjusted by DFW is a diagonal transformation of the representation space (for those recognizers for which the acoustic evaluation is based on an Euclidean distance measure) or a set of PPW

exponents (for those recognizers making use of a mixture of multivariate Gaussians). [1] We report how the DFW method can be applied to both types of recognizers and the improvements that can be achieved by applying this formalism.

The paper is organized into five sections. In Section 2, the DFE formalism is reviewed, and we also discuss which strategy for DFE (or DFW) is appropriate in the context of CSR. In Section 3, we show how the DFW method can be applied to those recognizers based on an Euclidean distance measure (Section 3.1) and those based on a mixture of multivariated Gaussians (Section 3.2). In Section 4, we present experiments to report the effect of applying the DFW formalism to DHMM-based and SCHMM-based continuous speech recognizers, and finally, in Section 5, we conclude with a summary of this work.

2. Discriminative feature extraction for continuous speech recognition

2.1. DFE formalism

The main concept of the DFE method is the discriminative training of the front-end feature extraction module, or a part of it. Even though several criteria could be applied for the discriminative training (like maximum mutual information (Bahl et al., 1986)), DFE is usually based on the MCE/GPD approach (Juang and Katagiri, 1992, 1997), as it is reviewed in this section.

Let $\Phi = \{\phi_1, \phi_2, \ldots, \phi_s\}$ be the set of parameters of the feature extractor to be adjusted by the DFE method. According to MCE/GPD approach, each parameter ϕ_s is iteratively re-estimated in order to minimize a cost function L representing the classification error. At iteration k, each parameter ϕ_s is updated by gradient descent of the cost function,

$$\phi_{s,k} = \phi_{s,k-1} - \eta \frac{\partial L}{\partial \phi_s} \Big|_{\Phi_{k-1}}, \tag{1}$$

where η is a learning factor. Let us suppose that we have a set of training events $\{X_1, \ldots, X_M\}$, and there is a set of classes $\{\lambda_1, \ldots, \lambda_I\}$. The cost function can be constructed as

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

where $l_m(X_m)$ is the cost function for the event X_m. This function should verify that if X_m is correctly classified then $l_m \to 0$, and if it is incorrectly classified, $l_m \to 1$. Usually, the cost function l_m is defined as a sigmoid function of an error measure $d_m(X_m)$,

$$l_m(X_m) = \frac{1}{1 + \exp[-\alpha d_m(X_m)]}, \tag{3}$$

where α is the transition parameter from correct to incorrect classification. The error measure $d_m(X_m)$ can be defined as

$$d_m(X_m) = -g_{k(m)} + \frac{1}{\beta} \log \left[\frac{1}{I-1} \sum_{j \neq k(m)} \exp(\beta g_j) \right], \tag{4}$$

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) and $\lambda_{k(m)}$ is the correct class for the event X_m. This definition of the error measure makes d_m negative if the classification of the input utterance is clearly correct and positive if clearly incorrect. The parameter β determines the contribution of the incorrect classes to d_m.

Making use of these definitions, the re-estimation of the set of parameters Φ is possible by using Eq. (1), where the partials $\partial L/\partial \phi_s$ can be written as follows:

[1] In the first case, the transformation is an element in the feature extraction module and the discriminative training of the transformation could be considered as a particular case of DFE. However, the PPW exponents are applied during the evaluation of the probabilities of the HMM states generating the feature vectors (the weights are not applied in the feature extraction module), and therefore, the discriminative training of the PPW exponents cannot be formally considered as a particular case of DFE. In spite of it, the underlying idea in both cases is the discriminative training of some parameters to be applied during the recognition process in order to weight the contribution of the components in the feature vector to the acoustic evaluation.

$$\frac{\partial L}{\partial \phi_s} = \frac{1}{M} \sum_{m=1}^{M} \frac{\partial l_m}{\partial \phi_s}, \tag{5a}$$

$$\frac{\partial l_m}{\partial \phi_s} = \frac{\partial l_m}{\partial d_m} \sum_{i=1}^{I} \frac{\partial d_m}{\partial g_i} \frac{\partial g_i}{\partial \phi_s}, \tag{5b}$$

$$\frac{\partial l_m}{\partial d_m} = \alpha l_m (1 - l_m), \tag{5c}$$

$$\frac{\partial d_m}{\partial g_i} = \begin{cases} -1 & \text{if} \quad i = k(m), \\ \dfrac{\exp(\beta g_i)}{\sum_{j \neq k(m)} \exp(\beta g_j)} & \text{if} \quad i \neq k(m) \end{cases} \tag{5d}$$

and the partials $\partial g_i / \partial \phi_s$ can be obtained from the definition of the classifier utilized for the discriminative training of the feature extractor.

2.2. Selection of a strategy for DFE

In order to perform the adjustment of the parameters Φ in the context of DFE, it is necessary to define the discriminant functions $g_i(X_m, \lambda_i)$. The most natural criterion for selecting the discriminant functions is to define them from the classifier utilized for the recognition process. This way of selecting the discriminant functions allows to perform jointly the discriminative training of the parameters of both, the feature extractor and the classifier, as proposed in (Biem and Katagiri, 1993, 1997).

When the parameters of the feature extractor and the classifier are simultaneously adjusted, the DFE method can be considered as a simple extension of the MCE/GPD method to the feature extraction module. However, some differences between the discriminative training of the classifier and the feature extractor must be considered. Most of the parameters of the classification module are class-dependent, and their adjustment only affects locally to the recognition process. During the discriminative training of them, some recognition errors can be corrected by the adjustment of some parameters while the rest of the classifier is not modified. In contrast to it, the parameters of the feature extraction module are shared by all classes and a modification of them can affect to the whole recognition system. Some authors have illustrated the importance of this conceptual difference. For example, in (Biem et al., 1997) a neural network implements both, the classifier and the feature extractor, and the MCE/GPD method is applied to train it. In this case, the parameters of the classifier can be randomly initialized, but the performance is very sensible to the initialization of the parameters associated to the feature extraction module.

Different strategies for the discriminative training of the feature extractor have been compared in (Paliwal et al., 1995; de la Torre et al., 1996b). These comparative experiments suggest that, for a simple classification problem (involving a small number of classes) and when the discriminant functions are simple, the best recognition results could be achieved in the case of simultaneous discriminative training of both, the classifier and the feature extractor. However, when the discriminant functions or the recognition problem are complex, the different nature of both modules makes this strategy less effective. In this case, when both modules are simultaneously trained the evolution of the feature extractor parameters is small compared to the classifier parameters. This occurs because a modification of the feature extractor affects the whole recognizer, while a modification of the classifier parameters can resolve local training errors (and reduce the cost function) without a modification of the rest of the classifier. As a result, the reduction of the cost function is easier by modifying the classifier parameters than modifying the feature extractor parameters, and the feature extractor is not properly trained. In addition, in some situations, the simultaneous discriminative training makes also the classifier be improperly trained, providing a greater training cost function and worse recognition performance than in the case of performing only the discriminative training of the classifier (de la Torre et al., 1996b). Therefore, for complex recognition problems, the comparative experiments show that the independent training of the feature extractor and the classifier provides better results than the simultaneous discriminative training.

Since the MCE/GPD method only guarantees to find a local minimum, for a complex cost function (because of the high number of classes or the complexity of the discriminant functions) the

parameters of the feature extractor only support a small adjustment, and obtaining a globally optimal set of parameters for the feature extractor becomes difficult. In order to achieve a proper adjustment of the feature extractor parameters, in a previous work (de la Torre et al., 1996b) we suggest (in addition to the independent training of the feature extraction module and the classification module) the use of a simplified classifier (and hence simplified discriminant functions) for the DFE process. This strategy simplifies the cost function to be minimized, and the feature extractor obtained in this case approaches the globally optimal solution better than in the case of using a complex cost function for the DFE method. After the discriminative training of the feature extractor, the classifier can be independently trained by means of a training procedure based on either maximum likelihood criterion or a discriminative criterion.

2.3. DFE for continuous speech recognition

Usually, current CSR systems are designed to deal with large vocabularies and the recognition is performed using acoustic units smaller than words, such as phoneme-like units (PLUs), diphonemes, context-dependent phonemes, etc. (Lee, 1990; Lee et al., 1990). In order to deal with a small number of classes (and hence, simplify the cost function) for the discriminative training of the feature extractor each class roughly corresponds to a context-independent phoneme. This does not limit the application of the feature extractor to a recognition system based on context-independent PLUs: since the feature extractor is globally optimized in order to improve the discrimination among the different PLUs, its application could increase the accuracy of the recognizers even in the case of one based on context-dependent PLUs.

In order to perform the discriminative training of the feature extractor, the sentences in the training data-base must be segmented into phonemes. Thus, for the DFE procedure, every speech event X_m corresponds to a sequence of feature vectors associated by segmentation to a certain phonetic class $\lambda_{k(m)}$.

The optimal conditions for the DFE procedure are achieved for an accurate segmentation of the training data-base. But commonly, an a priori accurate segmentation is not available for the training procedures of CSR systems, and algorithms for an automatic segmentation and labelling are necessary. In this work, the segmentation is performed by the Viterbi Beam Search (VBS) algorithm (Rabiner and Juang, 1993), which guarantees the optimal segmentation of the sentence, given a recognition system and the phonetic transcription of the sentence.

The objective of the DFE method is to improve the accuracy of the recognizer. By using the improved recognizer, a better segmentation could be obtained. And this better segmentation could lead to a better solution for the feature extractor, which could lead again to a better segmentation, etc. This suggest a segmental procedure to perform the DFE method (like that depicted in Fig. 2) similar to the segmental GPD training procedures described in (Chou et al., 1992; Juang et al., 1997). This way, at every segmental iteration, a new segmentation of the training data-base is obtained (via the VBS algorithm), the feature extractor is updated to this segmentation and the recognizer is updated to the obtained feature extractor. Several comments must be considered with respect to the proposed segmental procedure:

- The probabilistic descent theorem guarantees the monotonic minimization of the cost function L for a small enough value of the learning factor η (see Eq. (1)) during the DFE iterations. But the convergence of the segmental procedure is not guaranteed since a new segmentation could increase the value of the cost function. This fact could make difficult the selection of a convergence criterion for this segmental procedure. In the experiments, we have observed that the cost function is roughly minimized with the segmental iterations. This point is discussed in Section 4.2.1.
- Because of the use of the phonetic transcription to perform the segmentation, in the case of an accurate enough recognizer, the obtained segmentation tends to be very accurate. In this case, the differences among segmentations obtained at different segmental iterations tends to be irrele-

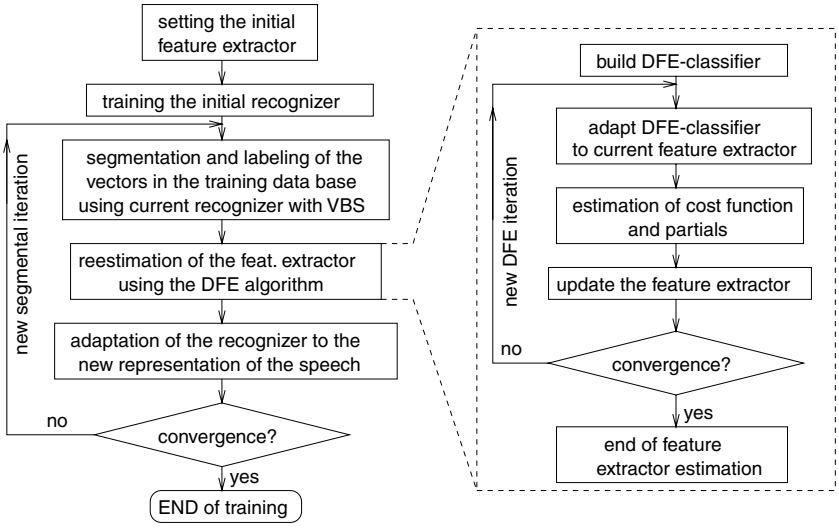

Fig. 2. Segmental DFE algorithm for CSR.

vant, and hence the initial segmentation could be considered accurate enough and all the DFE training procedure be performed using the first segmentation.

- Due to the coarticulation phenomenon, the sequences of vectors X_m associated to each class λ_i present an important variability, specially for the frames at the beginning and at the end of each sequence. This suggests that operations such as applying a temporal window to each sequence of vectors or the substitution of the sequence of vectors by the average vector (or using a temporal window to obtain the average vector) could be interesting in order to make easier the discriminative training of the feature extractor. Some experiments have been performed to clarify this point (see Section 4.2.2).

3. Discriminative feature weighting in HMM-based speech recognizers

All the considerations about the DFE formalism discussed in Section 2 can be directly applied to DFW. This is evident when the components are weighted by means of a diagonal transformation, since in this case DFW is a particular case of DFE. If the components are weighted by means of the PPW exponents, there are also class-independent parameters that are discriminatively trained in order to weight the contribution of the components and therefore the considerations in the last section can also be applied.

3.1. DFW for recognizers based on an Euclidean distance measure

Some of the speech recognition systems are based on an Euclidean distance measure. That is the case of DHMM systems (Rabiner and Juang, 1993) or Multiple Vector Quantization HMM (MVQHMM) systems (Segura et al., 1994). For those recognizers, the acoustic evaluation is performed by Vector Quantization (VQ): each feature vector is substituted by the discrete symbol verifying that the prototype vector associated to this symbol is the nearest one to the feature vector using the Euclidean distance measure.

The application of transformations to the feature space takes a special importance for the recognizers based on an Euclidean distance measure. For those recognizers, the effect of a linear transformation (described by a matrix) is an expansion or contraction of certain directions in the feature space (if the matrix is not diagonal, a rotation is also applied) which modifies the Euclidean

distance measure defined for the original representation space. As a consequence of it, the effect of the transformation is to weight the contribution of the different components of the feature vector to the acoustic evaluation.

In order to optimize the acoustic evaluation (and hence to achieve an accurate recognition) the transformation should enhance those components carrying more discriminative information. The relevance of transformations for recognizers based on an Euclidean distance measure is well known. For example, for those speech recognizers using feature vectors based on the cepstral coefficients, the performance is very sensible to the applied liftering window (Juang et al., 1987; Tohkura, 1987; Junqua and Wakita, 1989; de la Torre et al., 1996b).

The DFE method can be applied to the estimation of a linear transformation of the feature space, as proposed in (Paliwal et al., 1995; de la Torre et al., 1996a,b). For a N-dimensional feature space, the transformation is described by an $N \times N$ squared matrix. If the correlation among the different components in the feature vector is small, the transformation can be restricted to be diagonal (de la Torre et al., 1996b), and then, the effect of the transformation is a simple weighting of the contribution of each component to the Euclidean distance measure (and this approach derives to a DFW). Usually, the correlation among the components in the feature vectors is small (for example, for those representations based on cepstral coefficients) and the use of a diagonal transformation can be considered.

3.1.1. Single Gaussian DFW for the estimation of transformations

For those recognizers based on an Euclidean distance measure, we apply the discriminative training formalism to the estimation of a linear diagonal transformation of the representation space. This transformation is described by an $N \times N$ diagonal matrix V, where N is the size of the feature vectors. The parameters to be discriminatively trained are the elements in the main diagonal, $v_{p,p}$ $(1 \leqslant p \leqslant N)$.

In order to simplify the discriminant functions (according to the discussion in Section 2.2) each phoneme class λ_i is modeled by a single spherical Gaussian probability density function (pdf),

$$p(\tilde{\boldsymbol{x}} \mid \lambda_i) = \frac{1}{(2\pi\sigma_i^2)^{N/2}} \exp\left(-\frac{1}{2} \frac{\|\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{y}}_i\|^2}{\sigma_i^2} \right), \tag{6a}$$

$$\tilde{\boldsymbol{y}}_i = \frac{1}{J_i} \sum_{j=1}^{J_i} \tilde{\boldsymbol{x}}_{j,i}, \quad \sigma_i^2 = \frac{1}{J_i N} \sum_{j=1}^{J_i} \|\tilde{\boldsymbol{x}}_{j,i} - \tilde{\boldsymbol{y}}_i\|^2, \tag{6b}$$

where $\tilde{\boldsymbol{x}} = V\boldsymbol{x}$ is a transformed vector, $\|\tilde{\boldsymbol{x}}_{j,i} - \tilde{\boldsymbol{y}}_i\|^2$ is the squared Euclidean distance measure between $\tilde{\boldsymbol{x}}_{j,i}$ and $\tilde{\boldsymbol{y}}_i$, $\tilde{\boldsymbol{x}}_{j,i}$ $(j = 1, \ldots, J_i)$ are the transformed training vectors belonging to the class λ_i, and $\tilde{\boldsymbol{y}}_i$ and σ_i^2 are, respectively, the mean and the average variance of the vectors belonging to the class λ_i. During all the discriminative training procedure, an averaged variance is utilized instead of a covariance matrix in order to force the Gaussians to be spherical and to perform the discriminative training based on an Euclidean distance measure (instead of a Mahalanobis one).

According to this model, for a given input sequence of vectors $X_m = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ belonging to the class $\lambda_{k(m)}$, the discriminant functions can be defined as

$$\begin{aligned} g_i(X_m, \lambda_i) &= \log p(X_m \mid \lambda_i) = \frac{1}{T} \sum_{t=1}^{T} \log p(\tilde{\boldsymbol{x}}_t \mid \lambda_i) \\ &= -\frac{N}{2} \log(2\pi\sigma_i^2) - \frac{1}{2T\sigma_i^2} \sum_{t=1}^{T} \|\tilde{\boldsymbol{x}}_{j,i} - \tilde{\boldsymbol{y}}_i\|^2 \\ &= -\frac{N}{2} \log(2\pi\sigma_i^2) - \frac{1}{2T\sigma_i^2} \\ &\quad \times \sum_{t=1}^{T} \sum_{p=1}^{N} \left[v_{p,p}(x_t(p) - y_i(p)) \right]^2, \end{aligned} \tag{7}$$

where $x_t(p)$ and $y_i(p)$ are, respectively, the pth component of the vectors $\boldsymbol{x}_t$ and $\boldsymbol{y}_i$. From the definition of the discriminant functions the partials $\partial g_i / \partial v_{p,p}$ can be obtained,

$$\frac{\partial g_i}{\partial v_{p,p}} = -\frac{1}{T\sigma_i^2} \sum_{t=1}^{T} v_{p,p}[x_t(p) - y_i(p)]^2. \tag{8}$$

Finally, from the last equation and using Eqs. (5a)–(5d) the partials $\partial L / \partial v_{p,p}$ are derived, and the re-estimation formula (Eq. (1)) can be applied, where the parameters ϕ_s to be re-estimated are the

elements $v_{p,p}$ of the transformation matrix V. This way, using a single spherical Gaussian pdf to represent each class λ_i, the discriminative training formalism provides a new representation of the speech frames, for which the Euclidean distance between vectors belonging to different classes is maximized, and this makes the recognition process easier. This discriminative procedure will be referred as Single Gaussian Discriminative Weighting by Transformation (SGDWT).

3.1.2. Applying temporal windows to the single Gaussian DWT

The sequences of vectors X_m associated by the automatic segmentation algorithm to a certain class λ_i present an important variability at the beginning and at the end due to the coarticulation phenomenon. Also, the determination of the exact limit between two consecutive phonemes becomes difficult to the segmentation procedure due to the continuity of the speech. In order to minimize the effect of the continuity and the coarticulation, during the discriminative training of the transformation V a temporal window can be applied to the sequence of vectors in the discriminant functions,

$$g_i(X_m,\lambda_i) = -\frac{N}{2}\log(2\pi\sigma_i^2) - \frac{1}{2T\sigma_i^2}\sum_{t=1}^{T} W(T,t)\|\tilde{\boldsymbol{x}}_{j,i}-\tilde{\boldsymbol{y}}_i\|^2, \quad (9)$$

where $W(T,t)$ is a temporal window with length T that reduces the weight for those vectors near the limits of the sequence and enhances those vectors near the center of the sequence.

3.2. DFW for recognizers based on a mixture of Gaussians

During the last years, the use of recognizers based on mixtures of Gaussians has been widely extended, since these recognizers provide significantly better performance than those based on a discrete VQ (Young et al., 1997; Rabiner and Juang, 1993; Rubio et al., 1997). In a discrete VQ-based recognizer (like a DHMM one) each input vector is substituted by the discrete symbol verifying that the associated prototype vector is the nearest one to the input vector, using an Euclidean distance measure. In contrast to it, for a mixture-of-Gaussians based recognizer (like a SCHMM or a Continuous-HMM recognizer (Huang and Jack, 1989, 1990; Rabiner and Juang, 1993)), the acoustic evaluation is performed making use of a set of Gaussian pdfs. In this case, the probability of the vector $\boldsymbol{x}$ being generated by each Gaussian pdf, $p(\boldsymbol{x}\,|\,G_k)$, is evaluated, and this set of probabilities is utilized to perform the recognition process.

3.2.1. Invariance of the mixture-of-Gaussians based recognizers to transformations of the feature space

As discussed previously, a recognizer for which the acoustic evaluation is based on an Euclidean distance measure can be optimized by the application of a transformation of the feature space. If the transformation enhances those components in the feature vector carrying more discriminative information, the performance of the recognizer can be improved.

However, the situation is very different for those recognizers based on a mixture of Gaussians. Let G_k be a Gaussian pdf and $\boldsymbol{y}_k$ and Σ_k, respectively, the mean vector and the covariance matrix describing G_k. The probability of an input vector $\boldsymbol{x}$ being generated by the Gaussian pdf G_k is

$$p(\boldsymbol{x}\,|\,G_k) = \frac{1}{(2\pi)^{N/2}\sqrt{|\Sigma_k|}} \times \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{y}_k)^{\mathrm{T}}\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{y}_k)\right), \quad (10)$$

where N is the size of the feature vectors, $|\Sigma_k|$ is the determinant of the covariance matrix and Σ_k^{-1} is the inverse matrix of Σ_k. If a linear transformation V is applied to the feature space, the input vector, the mean vector and the covariance matrix are transformed as follows:

$$\tilde{\boldsymbol{x}} = V\boldsymbol{x}, \quad (11a)$$

$$\tilde{\boldsymbol{y}}_k = V\boldsymbol{y}_k, \quad (11b)$$

$$\tilde{\Sigma}_k = V\Sigma V^{\mathrm{T}} \quad (11c)$$

and therefore, the probability $p(\boldsymbol{x}\,|\,G_k)$ is transformed as follows:

$$\begin{aligned} p(\tilde{\boldsymbol{x}}\,|\,\tilde{G}_k) &= \frac{1}{(2\pi)^{N/2}\sqrt{|\tilde{\Sigma}_k|}} \\ &\quad \times \exp\left(-\frac{1}{2}(\tilde{\boldsymbol{x}}-\tilde{\boldsymbol{y}}_k)^{\mathrm{T}}\tilde{\Sigma}_k^{-1}(\tilde{\boldsymbol{x}}-\tilde{\boldsymbol{y}}_k)\right) \\ &= \frac{1}{(2\pi)^{N/2}\sqrt{|V\Sigma_k V^{\mathrm{T}}|}} \\ &\quad \times \exp\left(-\frac{1}{2}[(\boldsymbol{x}-\boldsymbol{y}_k)^{\mathrm{T}}V^{\mathrm{T}}]\right. \\ &\quad \left.\times\,[(V^{\mathrm{T}})^{-1}\Sigma_k^{-1}V^{-1}][V(\boldsymbol{x}-\boldsymbol{y}_k)]\right) \\ &= \frac{1}{|V|(2\pi)^{N/2}\sqrt{|\Sigma_k|}} \\ &\quad \times \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{y}_k)^{\mathrm{T}}\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{y}_k)\right) \\ &= \frac{1}{|V|}p(\boldsymbol{x}\,|\,G_k). \end{aligned} \tag{12}$$

As can be observed, the effect of applying a transformation V to the feature space is simply the introduction of a scale factor $1/|V|$ which does not depend on the Gaussian (only on the transformation V). So, as the scale factor is the same for all the Gaussian pdfs, the probabilities $P(G_k\,|\,\boldsymbol{x})$ of each Gaussian generating the input vector (and therefore the performance of the speech recognizer) remain invariant to the application of transformations of the feature space.

3.2.2. Partial probability weighting (PPW)

Due to the invariance of the acoustic evaluation to the application of feature space transformations, the SGDWT formalism is not applicable to improve those recognizers based on mixtures of Gaussians.

The mechanism involved in the optimization of an Euclidean-distance-measure based recognizer is the enhancement, by the application of a SGDWT transformation, of those components most relevant for the acoustic discrimination. Similarly, for those mixture-of-Gaussians based recognizers using diagonal covariance matrices (widely extended in the literature and practical implementations (Young et al., 1997; Moreno and Eberman, 1997; Rubio et al., 1997)) the probability of a vector being generated by each Gaussian of the mixture $p(\boldsymbol{x}\,|\,G_k)$ can be expressed as a product of partial probabilities, each one corresponding to each component, and, in this case, some components of the feature vector can be enhanced by the application of exponential weights w_n to the partial probabilities,

$$\tilde{p}(\boldsymbol{x}\,|\,G_k) = \prod_{n=1}^{N}\left[\frac{1}{\sqrt{2\pi\sigma_k^2(n)}} \times \exp\left(-\frac{(x(n)-y_k(n))^2}{2\sigma_k^2(n)}\right)\right]^{w_n}, \tag{13}$$

where $x(n)$ are the components of the input vector, $y_k(n)$ are the components of the mean vector of the Gaussian pdf G_k and $\sigma_k^2(n)$ are the elements in the main diagonal of the diagonal covariance matrix Σ_k,

$$\Sigma_k(p,q) = \begin{cases} 0 & \text{if} \quad p \neq q, \\ \sigma_k^2(p) & \text{if} \quad p = q. \end{cases} \tag{14}$$

This way of weighting the contribution of the different components to the acoustic evaluation, that we have named *Partial Probability Weighting* (PPW), have a precedent in the *codebook exponents* considered in (Young et al., 1997).

The discriminative training of the PPW exponents that we propose is based on discriminant functions derived from Eq. (13). In a mixture-of-Gaussians HMM based recognizer, each state s of a certain HMM is modeled as a mixture of Gaussians,

$$\begin{aligned} p(\boldsymbol{x}\,|\,s) &= \sum_k p(\boldsymbol{x}\,|\,G_k)P(G_k\,|\,s) \\ &= \sum_k p(\boldsymbol{x}\,|\,G_k)b_k(s). \end{aligned} \tag{15}$$

In order to reduce the complexity of the cost function, during the discriminative training we are using a HMM-based classifier that models every phoneme class λ_i as a single state HMM and then the modified probability of an input vector $\boldsymbol{x}$ given the class λ_i can be written as

$$\tilde{p}(\boldsymbol{x}\,|\,\lambda_i) = \sum_k \tilde{p}(\boldsymbol{x}\,|\,G_k)P(G_k\,|\,\lambda_i), \tag{16}$$

and if the covariance matrices Σ_k of the Gaussian pdfs are diagonal, the functions $\tilde{p}(\boldsymbol{x}|G_k)$ can be written as in the equation (13) and a discriminant function can be derived,

$$\begin{aligned} g_i(X_m,\lambda_i) &= \log\tilde{p}(X_m|\lambda_i) \\ &= \frac{1}{T}\sum_{t=1}^{T}\log\tilde{p}(\boldsymbol{x}_t|\lambda_i) \\ &= \frac{1}{T}\sum_{t=1}^{T}\log\left[\sum_k \tilde{p}(\boldsymbol{x}_t|G_k)P(G_k|\lambda_i)\right] \\ &= \frac{1}{T}\sum_{t=1}^{T}\log\left[\sum_k P(G_k|\lambda_i)\right. \\ &\quad \left.\times\prod_{n=1}^{N}\left[\frac{1}{\sqrt{2\pi\sigma_k^2(n)}}\exp\left(-\frac{(x(n)-y_k(n))^2}{2\sigma_k^2(n)}\right)\right]^{w_n}\right]. \end{aligned} \tag{17}$$

The formulation of the PPW procedure can be easily extended to a classifier (for the discriminative training) representing each phoneme class as an HMM with more than one state. However, it would increase the complexity of the cost function to be minimized and also the coarticulation effects should be considered and could make it not recommendable. Moreover, in order to avoid problems derived from the coarticulation and continuity effects, some solutions could be explored, like applying temporal windows to the input sequence of vectors (similarly as proposed in Section 3.1.2).

4. Recognition experiments

We have performed recognition experiments in order to evaluate the effect of the application of DFW formalism to CSR systems based on both, an Euclidean distance measure and mixture of Gaussian pdfs. The analysis for both categories of recognizers is based on the recognition experiments using DHMM and SCHMM recognizers, respectively.

4.1. Experimental conditions

For these experiments we have used the Spanish databases EUROM1 (Llisterri et al., 1993) for training and MINIGEO (Casacuberta et al., 1991; Díaz-Verdejo et al., 1998) for recognition. A version of these databases decimated to 8 kHz has been utilized. The front-end module includes pre-emphasis and segmentation into frames. Each frame is then represented by a vector that contains an energy coefficient, a cepstral vector containing 14 Mel Frequency Cepstral Coefficients (MFCC) (Davis and Mermelstein, 1980; Young et al., 1997), and the delta parameters (or first-order regression coefficients) and delta–delta parameters (or second-order regression coefficients) associated to these coefficients (Furui, 1986), which amounts to 45 components. The MFCC coefficients are obtained using a filter bank with 24 Mel scaled triangular filters. 24 context-independent PLUs are considered in order to represent the Spanish phonemes. Each PLU is modeled as a three states HMM with left-to-right topology and the silence is modeled as a one state HMM. Three DHMM-based recognition systems have been implemented using VQ codebooks with 128, 256 and 512 centroids. The SCHMM-based recognizers were implemented using mixtures of 128, 256 and 512 Gaussian pdfs. Both categories of recognizers have been trained with the maximum likelihood criterion (Rabiner and Juang, 1993).

Two different speaker independent recognition tasks have been prepared. The first one, labelled as MGEO, consists in the recognition of continuous speech with 203 words in the vocabulary. The perplexity estimated for this task is 5.9 (using a bigrammar). The second task, labelled MGEO-PHON, consists in the recognition of the phoneme-like units. For this task, a phoneme bigrammar was estimated from the training database. In this case, the number of elements in the vocabulary is 25 (24 PLUs plus silence) and a perplexity of 9.6 was estimated using the phoneme bigrammar.

4.2. DFW for DHMM-based recognition systems

According to previous discussions, in order to estimate the DWT transformation to be applied to DHMM-based recognizers, the cost function is based on a single Gaussian classifier, which models every class as a single spherical Gaussian pdf. In order to prepare the DFE classifier, the training

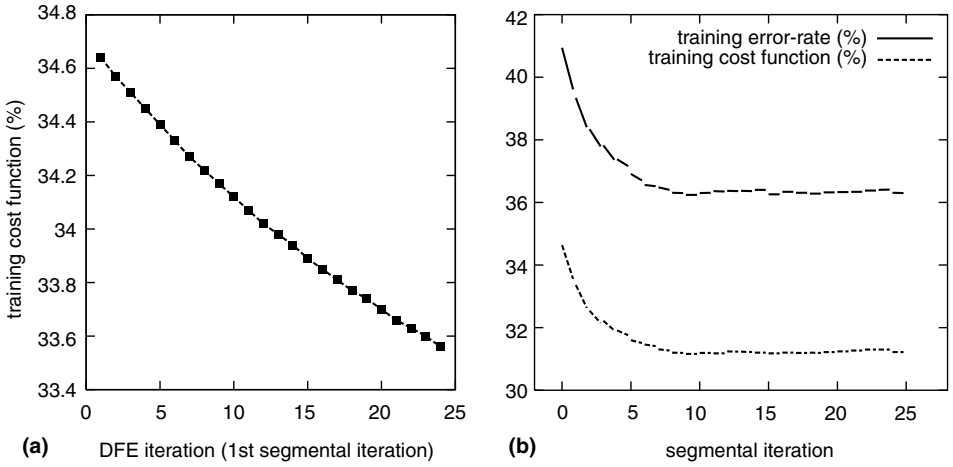

Fig. 3. Evolution of the training error-rate and the cost function during the estimation of the SGDWT transformation for the 256 centroids DHMM recognizer: (a) evolution of the cost function for the first segmental iteration; (b) evolution of the training error rate and the cost function with the segmental iterations.

database is automatically segmented into phonemes and labelled via the Viterbi algorithm using the recognizer and the phonetic transcription of each training sentence. Using the segments corresponding to each phoneme class, the single Gaussian pdf representing this class in the DFE classifier can be estimated. From the DFE classifier and the segments, the cost function can be obtained and the transformation of the feature space can be iteratively re-estimated.

4.2.1. Estimation of the DWT transformation

The estimation of the transformation is performed by a segmental algorithm like that represented in Fig. 2. At every segmental iteration, the recognizer is updated to the new transformation and a new segmentation of the training database is obtained. At every DFE iteration the transformation is iteratively re-estimated according to the DFE formalism. Usually, in order to homogenize the relative weight of the different components in the MFCC-based representations, a Statistically Weighted (SW) transformation is applied (Tohkura, 1987; Young et al., 1997). The SW transformation normalizes each component by multiplying it by the inverse of its standard deviation. This transformation has been utilized as reference transformation for the experiments and as initialization for the discriminative training procedure.

Fig. 3 shows the evolution of the cost function and the training error rate during the discriminative training of the DWT transformation. These plots correspond to the estimation of the transformation for the 256 centroids DHMM recognizer. Similar plots are obtained for the 128 and the 512 centroids recognizers. For a small enough value of the learning factor η, a monotonic minimization of the cost function is observed during the DFE iterations. Since the segmentation is not a continuous process, the monotonic minimization of the cost function is not guaranteed for the segmental iterations, and sometimes a small increment of the cost function is observed when a new segmentation is performed. However, in Fig. 3(b), a fast minimization of the cost function is obtained for the first segmental iterations, and beyond the iteration number 10 only small modifications on both, the training error rate and the cost function, are observed.

The evolution of the transformations with the segmental iterations is shown in Fig. 4. The figure represents the standard deviation of the transformed components, which represents the relative contribution of each component to the Euclidean distance measure after the transformation is applied. The reference transformation used as initialization (the SW transformation) consists in a normalization of all the components in the feature

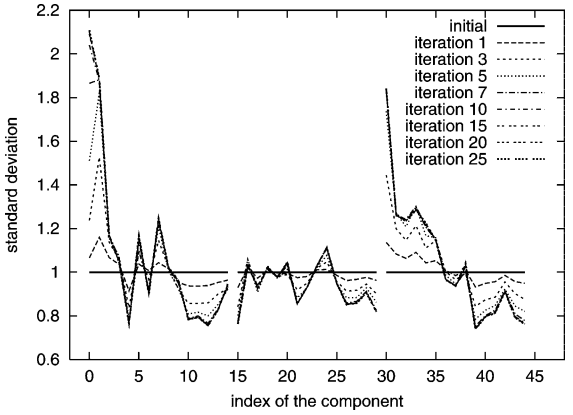

Fig. 4. Evolution of the transformation with the segmental iterations for the 256 centroids DHMM recognizer. We have plotted the standard deviation of the transformed components.

vector that makes all the standard deviations equal to 1. The first coefficient is the energy parameter. The next 14 are the MFCC coefficients. The next 15 are the corresponding delta parameters and the last 15 are the delta–delta parameters. In this figure, the fast evolution of the transformation in the first segmental iterations can be observed. The evolution of the transformation beyond the segmental iteration number 10 is insignificant.

4.2.2. Recognition experiments with DHMM recognizers

Recognition experiments have been performed applying the transformation obtained at every segmental iteration. Fig. 5 shows the recognition performance as a function of the segmental iteration, using the 256 centroids DHMM recognizer. The recognition results corresponds to the MGEO task. In this figure, a significant reduction of the recognition word error rate is observed. From an error of 8.2% (using the initial SW transformation) the error rate is reduced to 6.0%. A fast reduction is observed during the first segmental iterations and from the iteration number 7 onwards, only small variations are observed in the error rate.

The effect of applying the SGDWT transformation is not only the reduction of the error rate. The average number of active nodes in the recognition tree is also reduced for a given pruning threshold, because of the increment of the accuracy in the recognizer. This reduces the requested memory and the recognition time. In Fig. 5, the evolution of the average number of active nodes and the recognition time with the segmental iteration are also shown. The recognition time is related to the duration of the sentence. Again, the reduction of the average number of active nodes and the recognition time is obtained for the first segmental iterations and only a small reduction is observed from the iteration number 10 onwards.

In order to reduce the influence of the continuity of the speech and the coarticulation effects over the discriminative estimation of the transformation, according to the discussion in Section 3.1.2, we have applied a temporal window to the sequence of vectors associated to each phoneme during the DFE training of the transformation. For a sequence with T frames, we have applied a temporal window $W(T,t)$ with $1 \leqslant t \leqslant T$, with a half sine wave shape, described by

$$W(T,t) = \omega \sin\left(\pi \frac{t}{T+1}\right), \tag{18}$$

where ω is a normalization constant,

$$\omega = \frac{T}{\sum_{t=1}^{T} \sin\left(\pi \frac{t}{T+1}\right)}. \tag{19}$$

This temporal window enhances the contribution of the central part of the phonemes and reduces the contribution of the frames near the limits and, therefore, focuses the discriminative training of the transformation on the stationary part of the phonemes.

The transformations obtained when the temporal window is applied has been labeled SGDWT-W. These transformations were obtained by a segmental procedure similar to that utilized for the SGDWT transformations. In this case, the evolution of the cost function and the training error rate, and the evolution of the transformation are similar to those depicted in Figs. 3 and 4. Fig. 6 compares the transformations obtained with the DFE formalism with and without applying the temporal window during the discriminative training of the transformation. In this figure, the standard deviations of the transformed components are represented for the SW (reference), the

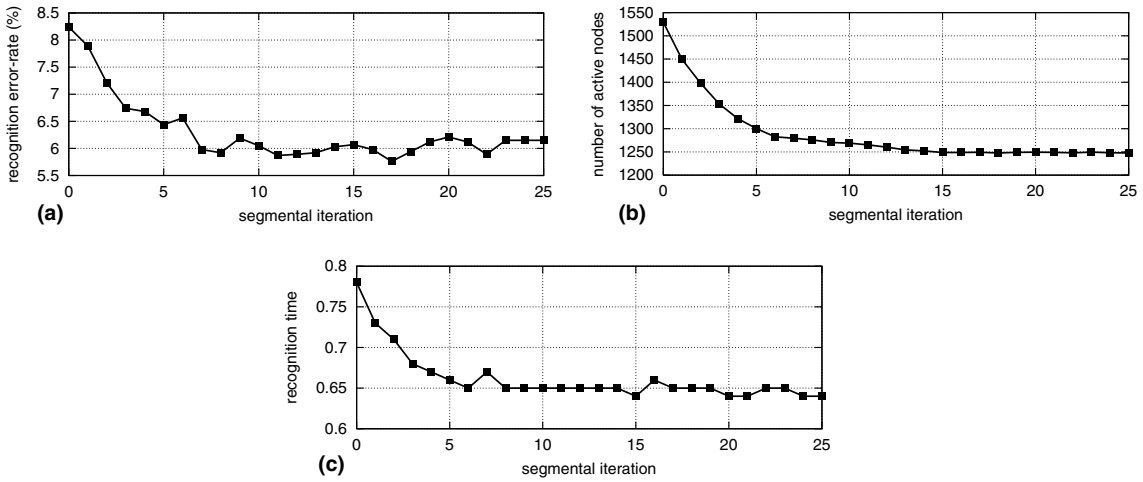

Fig. 5. Evolution of the recognition performance with the segmental iterations. The results correspond to the 256 centroids DHMM recognition system and the MGEO task: (a) word error rate; (b) average number of active nodes; (c) recognition time.

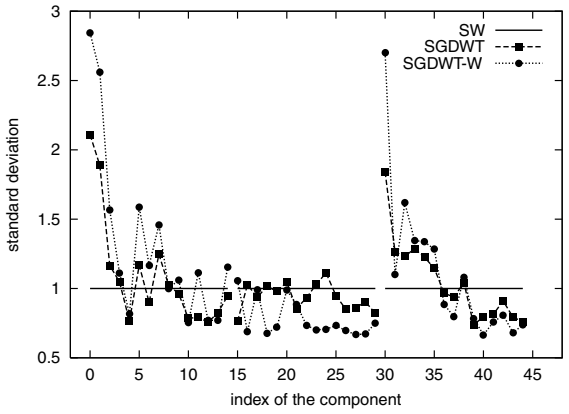

Fig. 6. Standard deviation of the transformed components for the transformations SW, SGDWT and SGDWT-W (256 centroids DHMM recognition system, segmental iteration number 25).

SGDWT and the SGDWT-W transformations. These transformations corresponds to the 256 centroids DHMM recognizer and the segmental iteration number 25.

The recognition performance for the 128, 256 and 512 centroids DHMM recognizers for both tasks (MGEO and MGEO-PHON) is represented in Figs. 7 and 8. In these figures, the effect of applying the different transformations of the representation space (SW, SGDWT and SGDWT-W) can be evaluated. The results presented in these plots are the average for the last 10 segmental iterations. The application of the discriminatively trained transformations reduces significantly the error rate for both tasks (Fig. 7). The application of temporal windows for the estimation of the SGDWT-W transformation improves the performance of the recognizer with respect to the SGDWT transformation. For the MGEO task, the application of the SGDWT transformation reduces the error rate by 22% and the SGDWT-W transformation by 26%. The improvement of the recognition performance is also important for the phoneme recognition task. Fig. 8 represents the average number of active nodes in the recognition tree and the average recognition time (related to the duration of the sentence) for the MGEO task when the different transformations are applied. In these plots, an important reduction of the computational requirements can also be observed. The application of the discriminatively trained transformations reduces the recognition time by 15% and the average number of active nodes by 20%. Since no pruning threshold is applied for the MGEO-PHON task the number of active nodes and the recognition time are not affected by the application of transformations in this case.

The experimental results show how the recognition performance is affected by the application of transformations of the feature space in those rec-

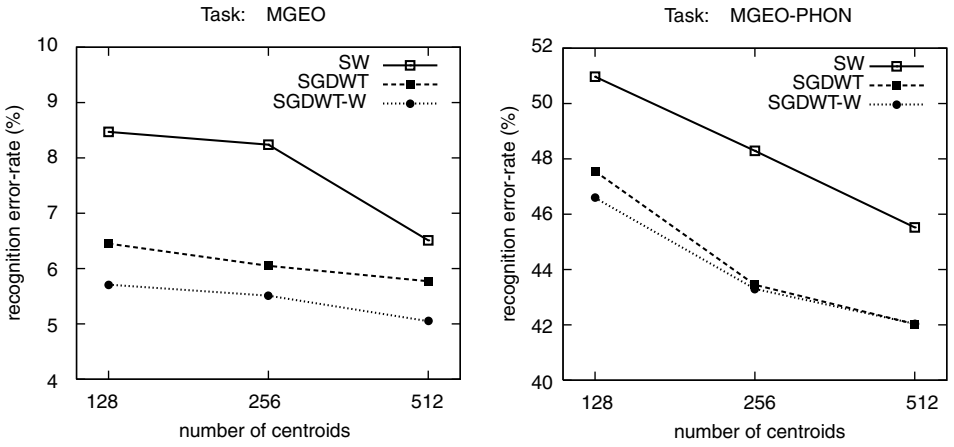

Fig. 7. Recognition results obtained with the DHMM recognizers by applying the SW, SGDWT and SGDWT-W transformations.

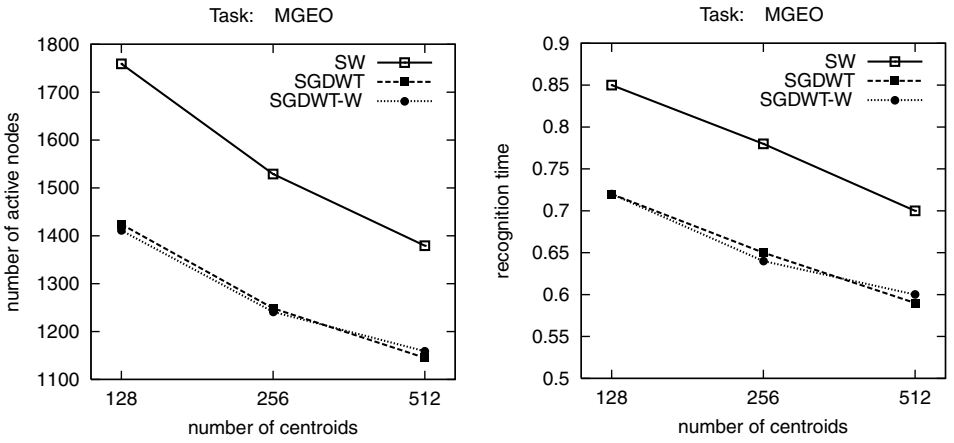

Fig. 8. Average number of active nodes and recognition time for the task MGEO using the DHMM recognizers when the SW, SGDWT and SGDWT-W transformations are applied.

ognizers for which the acoustic evaluation is based on the Euclidean distance measure, like the DHMM ones. For this kind of recognizers, the performance can be substantially improved by the application of transformations trained with a discriminative criterion, like those based on the DFW formalism.

4.3. DFW for SCHMM based recognition systems

Those recognizers for which the acoustic evaluation is based on a mixture of multivariated Gaussian pdfs, like the Continuous-HMM or the SCHMM recognizers, are not affected by the application of a transformation of the feature space. For this reason, for this kind of recognizers, the element estimated by the DFW method is a set of exponential weights that are applied to the partial probabilities associated to each component of the feature vector. This method for tuning the relative contribution of the different components to the acoustic evaluation is named PPW.

4.3.1. Estimation of the PPW weights

The PPW procedure allows the enhancement of some components in the feature vector for the

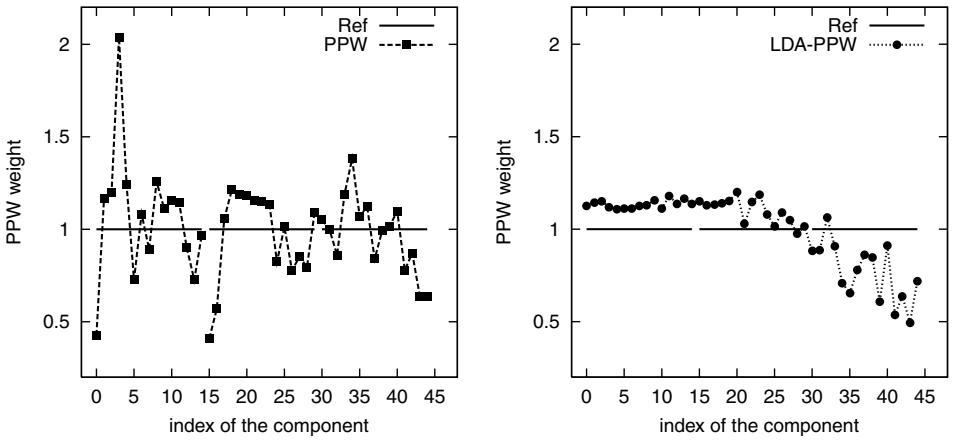

Fig. 9. Reference, PPW and LDA-PPW weights obtained for the 256 Gaussians SCHMM recognizer.

acoustic evaluation. However, the directions to be enhanced depend on the basis vectors utilized to represent the feature vectors, and a non-diagonal transformation of the feature space (including a rotation) could modify the effect of the PPW procedure. For this reason, an adequate transformation (including a rotation of the feature space) applied before the estimation and the application of the PPW weights could increase the improvement derived from the discriminative weighting of components. The transformation can be estimated with the Linear Discriminant Analysis (LDA) formalism (Fukunaga, 1990; Milner, 1997; Kumar and Andreou, 1998; Jin and Waibel, 2000), which selects the basis vectors with a discriminative criterion and allows the arrangement of the new components according to their discriminative capability.

In the recognition experiments for SCHMM recognizers, we have obtained the PPW weights without and with applying a LDA transformation to the feature space before the estimation of the exponential weights. The classifier utilized for the discriminative training of the PPW weights represented each class as a single state HMM according to Eq. (18). The single state model was defined from the center state of the corresponding model in the SCHMM recognition system (Milner, 1997) in order to reduce the complexity of the cost function, and a temporal window $W(T,t)$ has been applied for the definition of the discriminant functions in order to reduce the coarticulation effect. All the weights were set to 1 as initialization.

Fig. 9 shows the set of PPW weights obtained by applying the DFW formalism. The reference weights, used as initialization, are also represented. When the LDA transformation is applied, the new components are rearranged according to their discriminative capability (taking into account the eigenvalues associated to the discrimination matrix) (Jin and Waibel, 2000). The transformed components more relevant for the discrimination are those with lower indexes in the new feature vector. For this reason, when LDA is applied, the PPW weights associated to higher index components tends to be smaller than those for lower indexes, as can be observed in Fig. 9.

4.3.2. Recognition experiments with SCHMM recognizers

The recognition results using the SCHMM recognizers for both tasks (MGEO and MGEO-PHON) are presented in Figs. 10 and 11. The plots in Fig. 10 represent the error rate. As can be observed, the application of the PPW weights improves the accuracy of the recognizers with respect to the reference, specially when LDA and DFW are combined. The combination of LDA and DFW has reduced the error rate by 13% with respect to the baseline results for the MGEO task. In contrast to the DHMM recognizers, the SCHMM

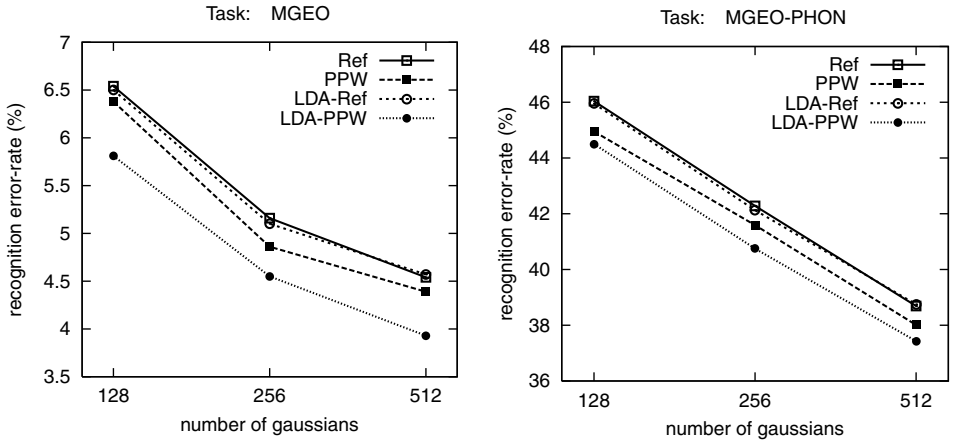

Fig. 10. Recognition results obtained with the SCHMM recognizers for the MGEO and the MGEO-PHON tasks. Comparison of the performance obtained by applying the reference and PPW weights without and with LDA.

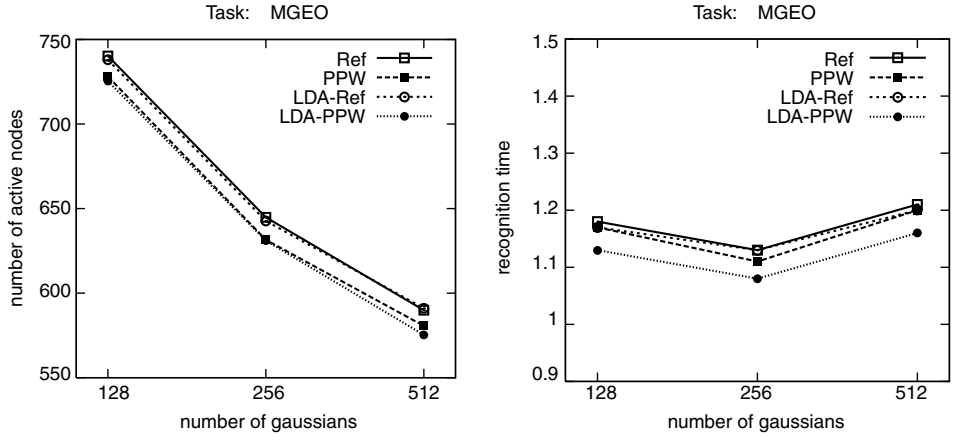

Fig. 11. Average number of active nodes and recognition time for the task MGEO using the SCHMM recognizers when the reference and PPW weights are applied without and with LDA.

recognition systems are less sensible to the application of the weights. This can be justified taking into account that in the DHMM systems the feature vector is substituted by the symbol associated to the nearest centroid, and a modification of the Euclidean distance measure could drastically modify the acoustic observation, since a different centroid could be the nearest one when the new distance measure is utilized. In the case of the SCHMM recognizers, the acoustic evaluation is based on the probabilities of the different Gaussian pdfs given the input vector, and a modification of the PPW weights produces a smooth modification of the acoustic observation, since the effect is a modification of the relative probability of each Gaussian.

The application of LDA itself does not provide significant improvements in the recognition performance. However, when LDA and DFW are combined, significant improvements are obtained in the recognition performance for both tasks. The improvements of LDA–PPW with respect to PPW are obtained because the LDA transformation provides a set of basis vectors more adequate

than the original one for the application of the DFW.

The increment of computational load derived from the application of the PPW weights is irrelevant compared to the recognition time and is compensated by the reduction in the recognition time derived from the improvement in the discriminative capability. Fig. 11 represents the average number of active nodes and the average recognition time (related to the duration of the sentence) for the task MGEO when the different PPW weights are applied. Only a small reduction in the number of nodes is appreciated for both sets of PPW weights and a small reduction in the recognition time is observed in the case of LDA-PPW weights.

The above recognition results show that the DFW formalism can be successfully applied to improve the performance of continuous speech recognizers in which the acoustic evaluation is based on a mixture of Gaussian pdfs. In this case, an improvement of the discriminative capability of the recognition system is achieved, which leads to an improvement in the recognition performance and a reduction in the computational requirements (number of active nodes and recognition time) more significant when DFW is combined with LDA.

5. Conclusions

The importance of the feature extraction block for all the pattern classification problems and, in particular, for automatic speech recognition is well known. An important effort has been dedicated to the improvement of the feature extraction block for speech recognizers (Junqua and Wakita, 1989; Tohkura, 1987; Juang et al., 1987; Peinado et al., 1990; Furui, 1986). The DFE method provides a formalism for the discriminative optimization of the feature extractor (Biem and Katagiri, 1994, 1997; Paliwal et al., 1995). One of the applications derived from the DFE method is the DFW (Biem and Katagiri, 1993, 1997; de la Torre et al., 1996a,b).

This work has been devoted to the application of the DFW formalism to improve the performance of CSR systems based on hidden Markov modeling. In this work we try to optimize the recognizers by tuning the contribution of the different components to the acoustic evaluation. This tuning is performed with a discriminative criterion.

Two different categories of recognizers are considered in this work: those for which the acoustic evaluation is based on an Euclidean distance measure (like the DHMM recognizers) and those based on a mixture of Gaussian pdfs (like the continuous and the semi-continuous HMM recognizers). For the systems in the first category, the discriminative training is applied to estimate a transformation of the feature space. For those recognizers in the second category, the acoustic evaluation is not affected by the application of a transformation of the feature space, as discussed in Section 3.2.1. In this case, we have applied exponential weights to the partial probabilities associated to each component, and so, this method is named Partial Probability Weighting. In both cases, the objective of the DFW method is a fine tuning of the contribution of the different components to the acoustic evaluation.

Those systems based on an Euclidean distance measure are very sensible to the application of feature space transformations and important improvements can be achieved by the application of transformations properly estimated. The recognizers based on mixtures of Gaussian pdfs are less sensible to the application of the PPW weights. The reason of it is the different nature of the acoustic evaluation in both cases.

The application of the DFW formalism to the CSR systems has improved the performance for both categories of recognizers. In the case of the DHMM systems, the application of the DWT transformations has reduced significantly the recognition error rate. Reductions in the average number of active nodes and the recognition time have also been observed. The application of temporal windows for the estimation of the transformation leads to additional improvements in the performance of the DHMM recognizers.

In the case of SCHMM recognizers the improvements are significant, although not so important as in the discrete case. We have combined the DFW with LDA (Fukunaga, 1990). The ap-

plication of a LDA transformation before the estimation of the PPW exponential weights selects the basis vectors (and then the proper directions that will be enhanced by the PPW weights) with a discriminative criterion. This way, the combination of LDA and DFW provides more significant improvements in the discriminative capability of the recognition systems which leads to a better recognition performance for mixture of Gaussians HMM-based continuous speech recognizers.

The application of both, transformations or PPW weights, does not implies an increment in the recognition time. Moreover, the application of the transformations or the PPW weights properly estimated with the DFW formalism provides a reduction in the recognition time due to the improvement of the discriminative capability of the feature extractor.

References

Bacchiani, M., Aikawa, K., 1994. Optimization of time-frequency masking filters using minimum classification error criterion. In: Proc. Internat. Conf. on Acoust. Speech Signal Process., ICASSP-94, Vol. 2, pp. 275–278.

Bahl, L.R., Brown, P.F., de Souza, P.V., Mercer, R.L., 1986. Maximum mutual information estimation of hidden Markov models parameters for speech recognition. In: Proc. ICASSP-86, Tokyo, pp. 49–52.

Biem, A., Katagiri, S., 1993. Feature extraction based on minimum classification error/generalized probabilistic descent method. In: Proc. ICASSP-93, Vol. 2, pp. 275–278.

Biem, A., Katagiri, S., 1994. Filter bank design based on discriminative feature extraction. In: Proc. ICASSP-94, Vol. 1, pp. 485–488.

Biem, A., Katagiri, S., 1997. Cepstrum-based filter bank design using discriminative feature extraction training at various levels. In: Proc. Internat. Conf. on Acous. Speech Signal Process., ICASSP-97, Vol. 2, pp. 1503–1506.

Biem, A., Katagiri, S., Juang, B.H., 1997. Pattern recognition using discriminative feature extraction. IEEE Trans. Signal Process. 45 (2), 500–504.

Biem, A., McDermott, E., Katagiri, S., 1995. Discriminative filter bank model for speech recognition. In: Proc. EuroSpeech-95, Vol. 1, pp. 545–548.

Casacuberta, F., Garcia, R., Llisterri, J., Nadeu, C., Pardo, J.M., Rubio, A., 1991. Development of spanish corpora for speech research (albayzin). In: Proc. Workshop Internat. Cooperation and Standarization of Speech Databases and Speech I/O Assessment Methods, September, pp. 26–28.

Chou, W., Juang, B.H., Lee, C.H., 1992. Segmental gpd training of hmm based speech recognizer. In: Proc. ICASSP-92, pp. I,473–476.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuosly spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28 (4), 357–366.

Díaz-Verdejo, J.E., Peinado, A.M., Rubio, A.J., Segarra, E., Prieto, N., Casacuberta, F., 1998. Albayzin: a task-oriented spanish speech corpus. In: Proc. First Internat. Conf. on Language Resources and Evaluation LREC-98, Vol. 1, pp. 497–501.

de la Torre, A., Peinado, A.M., Rubio, A.J., Segura, J.C., Sánchez, V.E., 1996a. Minimum classification error transformation for improving speech recognition systems. In: Proc. EUSIPCO-96, Vol. 3, pp. 1575–1578.

de la Torre, A., Peinado, A.M., Rubio, A.J., Sánchez, V.E., Díaz, J.E., 1996b. An application of minimum classification error to feature space transformations for speech recognition. Speech Communication 20 (3–4), 273–290.

de la Torre, A., Peinado, A.M., Rubio, A.J., Garcia, P., 1997. Discriminative feature extraction for speech recognition in noise. In: Proc. EuroSpeech-97, Vol. 1, pp. 291–294.

Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. Wiley, New York.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press, New York.

Furui, S., 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans. ASSP 34 (February), 52–59.

Huang, X.D., Jack, M.A., 1989. Unified techniques for vector quantisation and hidden markov modeling using semi-continuous models. In: Proc. ICASSP-89, Glasgow (Scotland), May, pp. 639–642.

Huang, X., Lee, K.F., Hon, H.W., 1990. On semi-continuous hidden markov modeling. In: Proc. ICASSP-90, pp. 689–692.

Jin, Q., Waibel, A., 2000. Application of lda to speaker recognition. In: Proc. ICSLP-00, Vol. 2, pp. 250–253.

Juang, B.H., Katagiri, S., 1992. Discriminative learning for minimum error classification. IEEE Trans. Signal Process. 40 (12), 3043–3054.

Juang, B.H., Rabiner, L.R., Wilpon, J.G., 1987. On the use of bandpass liftering in speech recognition. IEEE Trans. ASSP 35 (7), 947–954.

Juang, B.H., Chou, W., Lee, C.H., 1997. Minimum classification error rate methods for speech recognition. IEEE Trans. Speech Audio Process. 5 (3), 257–265.

Junqua, J.C., Wakita, H., 1989. A comparative study of cepstral lifters and distance measures for all pole models of speech in noise. In: Proc. ICASSP-89, pp. 476–479.

Kumar, N., Andreou, A.G., 1998. Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. Speech Communication 26, 283–297.

Lee, C.H., Rabiner, L.R., Pieraccini, R., Wilpon, J.G., 1990. Acoustic modeling for large vocabulary speech recognition. Comput. Speech Language 4, 127–165.

Lee, K.F., 1990. Context-dependent phonetic hidden markov models for speaker-independent continuous speech recognition. IEEE Trans. ASSP 38, 599–623.

Llisterri, J., Aguilar, L., Blecua, B., Machuca, M., Mota, C., Ríos, A., Moreno, A., 1993. Spanish eurom. 1. ESPRIT PROJECT 6819 (SAM-A).

McDermott, E., Katagiri, S., 1994. Protoype-based minimum classification error/generalized probabilistic descent training for various speech units. Comput. Speech Language 8 (4), 351–368.

Milner, B., 1997. Cepstral-time matrices and lda for improved connected digit and sub-word recognition accuracy. In: Proc. EuroSpeech-97, Vol. 1, pp. 405–408.

Moreno, P.J., Eberman, B., 1997. A new algorithm for robust speech recognition: the delta vector taylor series approach. In: Proc. EuroSpeech-97, Vol. 5, pp. 2599–2602.

Paliwal, K.K., Bacciani, M., Sagisaka, Y., 1995. Minimum classification error training algorithm for feature extractor and pattern classifier in speech recognition. In: Proc. EuroSpeech-95, Vol. 1, pp. 541–544.

Peinado, A.M., Ramesh, P., Roe, D.B., 1990. On the use of energy information for speech recognition using HMM. In: Proc. EUSIPCO-90, Barcelona, September, Vol. 2, pp. 1243–1246.

Peinado, A.M., Rubio, A.J., Segura, J.C., Sánchez, V., Díaz, J.E., 1995. Mce estimation of vq parameters for mvqhmm speech recognition. In: Proc. EuroSpeech-95, Vol. 1, pp. 533–536.

Peinado, A.M., Segura, J.C., Rubio, A.J., García, P., Pérez, J.L., 1996. Discriminative codebook design using multiple vector quantization in hmm-based speech recognizers. IEEE Trans. Speech Audio Process. 4 (2), 88–95.

Rabiner, L.R., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ.

Rubio, A.J., García, P., de la Torre, A., Segura, J.C., Díaz-Verdejo, J., Benítez, M.C., Sánchez, V., Peinado, A.M., López-Soler, J.M., Pérez Córdoba, J.L., 1997. Stacc: an automatic service for information access using csr through telephone line. In: Proc. EuroSpeech-97, Vol. 4, pp. 1779–1782.

Segura, J.C., Rubio, A.J., Peinado, A.M., García, P., Román, R., 1994. Multiple VQ hidden Markov modeling for speech recognition. Speech Communication 14 (2), 163–170.

Tohkura, Y., 1987. A weighted cepstral distance measure for speech recognition. IEEE Trans. ASSP 35 (10), 1414–1422.

Watanabe, H., Yamaguchi, T., Katagiri, S., 1995. Discriminative metric design for pattern recognition. In: Proc. ICASSP-95, pp. 3439–3442.

Watanabe, H., Yamaguchi, T., Katagiri, S., 1997. Discriminative metric design for robust pattern recognition. IEEE Trans. Signal Process. 45 (11), 2655–2662.

Young, S., 1996. A review of large-vocabulary continuous speech recognition. IEEE Signal Process. Magazine 13 (5), 45–57.

Young, S., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 1997. The HTK Book, Cambridge University.