Unsupervised Data-Driven Feature Vector Normalization With Acoustic Model Adaptation for Robust Speech Recognition

Luis Buera, Antonio Miguel, Óscar Saz, Alfonso Ortega, and Eduardo Lleida, *Member, IEEE*

***Abstract*—In this paper, an unsupervised data-driven robust speech recognition approach is proposed based on a joint feature vector normalization and acoustic model adaptation. Feature vector normalization reduces the acoustic mismatch between training and testing conditions by mapping the feature vectors towards the training space. Model adaptation modifies the parameters of the acoustic models to match the test space. However, since neither is optimal, both approaches use an intermediate space between training and testing spaces to map either the feature vectors or acoustic models. The joint optimization of both approaches provides a common intermediate space with a better match between normalized feature vectors and adapted acoustic models. In this paper, feature vector normalization is based on a minimum mean square error (MMSE) criterion. A Class Dependent Multi-Environment Model LInear Normalization (CD-MEMLIN) based on two classes (silence/speech) with a Cross Probability Model (CD-MEMLIN-CPM) is used. CD-MEMLIN-CPM assumes that each class of clean and noisy spaces can be modeled with a Gaussian mixture model (GMM), training a linear transformation for each pair of Gaussians in an unsupervised data-driven training process. This feature vector normalization maps the recognition space feature vector to a normalized space. The acoustic model adaptation maps the training space to the normalized space by defining a set of linear transformations over an expanded HMM-state space, compensating for those degradations that the feature vector normalization is not able to model, like rotations. Experiments have been carried out with the Spanish SpeechDat Car database and Aurora 2 databases using both the standard Mel-frequency cepstral coefficient (MFCC) and advanced ETSI front-ends. Consistent improvements were reached for both corpora and front-ends. Using the standard MFCC front-end, a 92.08% average improvement on WER for Spanish SpeechDat Car and a 69.75% average improvement for clean condition evaluation of Aurora 2 was obtained, improving those results reached with ETSI advanced front-end (83.28% and 67.41%, respectively). Using the ETSI advanced front-end with the proposed solution, a 75.47% average improvement was obtained for the clean condition evaluation of Aurora 2 database.**

***Index Terms*—Acoustic model adaptation, data-driven feature vector normalization, linear transformation matrices, robust speech recognition, unsupervised.**

Manuscript received October 27, 2008; revised May 18, 2009. First published June 30, 2009; current version published November 20, 2009. This work was supported by the MEC of the Spanish government under Project TIN 2005-08660-C04-01. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yariv Ephraim.

L. Buera was with the Department of Electronic Engineering and Communications, University of Zaragoza, 50009 Zaragoza, Spain. He is now with the Speech Technology Group, Toshiba Research Europe, Cambridge CB4 0GZ, U.K. (e-mail: luis.buera@crl.toshiba.co.uk).

A. Miguel, Ó. Saz, A. Ortega and E. Lleida are with the Department of Electronic Engineering and Communications, University of Zaragoza, 50009 Zaragoza, Spain.

Digital Object Identifier 10.1109/TASL.2009.2026441

I. Introduction

AUTOMATIC speech recognition (ASR) systems can achieve satisfactory performance under controlled conditions. However, when training and testing acoustic conditions differ, the accuracy of the systems rapidly degrades. To compensate for the different effects which cause the mismatch between training and recognition spaces, robustness techniques have been developed along three main lines of research [1]:

- robust feature vector extraction methods, to generate acoustic vectors less affected by the noise;
- acoustic model adaptation methods, which map acoustic models from training space to recognition space;
- feature vector adaptation/normalization methods, which map testing space feature vectors to the training space.

In this paper, we focus on "unsupervised" methods, which are those that do not require a transcription of the training data. Unsupervised adaptation can provide more user friendly solutions for ASR applications because they do not require active enrollment by the speakers. Also, they can be very useful in those situations where the transcriptions are not available, or they are very expensive to obtain.

Feature vector adaptation/normalization methods fall into one of these three main classes [2]: high-pass filtering, model-based techniques, and empirical compensation. High-pass filtering research involves methods such as Cepstral Mean Normalization (CMN) [3], [4] and RelAtive SpecTral Amplitude (RASTA) processing [5], which are included in almost all ASR systems because they are simple and effective. Model-based methods assume that the mismatch between training and recognition spaces can be represented by a structural model of environmental degradation. So, the corresponding parameters of the structural model are estimated and the appropriate inverse operation is applied to compensate the recognition signal. Examples of model-based methods are Vector Taylor Series for feature normalization (VTS) [6], Codeword Dependent Cepstral Normalization (CDCN) [7], and Spectral Subtraction (SS) [8]. Finally, empirical compensation methods use direct cepstral comparisons and are entirely data-driven. Typically, a training phase based on stereo data (clean and noisy signals simultaneously recorded) is required to estimate some transformations, although "blind" approaches, which use only noisy training data, have been developed [9], [10]. Some algorithms which are based on this approach include multivariate Gaussian-based cepstral normalization (RATZ) [9], Stereo-based Piecewise

Linear Compensation for Environments (SPLICE) [11], Probabilistic Optimum Filtering (POF) [12] and Multi-Environment Model-based LInear Normalization (MEMLIN) [10].

In general, acoustic model adaptation methods produce better results than other robustness research lines because they can model the uncertainty caused by the noise statistics in a more accurate way [13]. However these methods require more data and computing time. Furthermore, the performance of acoustic model adaptation methods degrades dramatically when the transcription of the adaptation data is not available (unsupervised techniques) [14]. For this case, a previous step to provide an estimation of the transcription of the adaptation data is needed (usually a recognition process). Unfortunately, this approach can not provide satisfactory performance when the adaptation data are highly noise corrupted or the recognition task is complex (e.g., large vocabulary, spontaneous speech, etc.) because the word accuracy might be not good enough. On the other hand, several feature vector normalization/adaptation techniques, which in general do not need as much training data and computational time as acoustic model adaptation methods, have proved to be very effective under adverse conditions [10], [11]. Thus, in order to combine the advantages of the acoustic model adaptation and the feature vector adaptation/normalization methods in an unsupervised framework, we consider an unsupervised hybrid solution in this work. By hybrid solution, we mean a combination of a feature vector normalization/adaptation technique with an acoustic model adaptation technique.

Previous work [10] has shown that MEMLIN and its supervised Phoneme Dependent extension (PD-MEMLIN) are effective in compensating for the effects of adverse dynamic car conditions, thus improving the performance of techniques based on similar criteria, e.g., SPLICE, RATZ. In order to use unsupervised data in the training process, and following the same philosophy as PD-MEMLIN, a class dependent generalization is proposed in this paper. Thus, Class-Dependent MEMLIN (CD-MEMLIN) is defined and a simple class definition, silence and speech, is used throughout the paper. Note that this class definition works in an unsupervised framework by using a voice activity detector (VAD) over the clean training speech to classify the frames. CD-MEMLIN assumes that each acoustic class of clean and noisy spaces can be modeled with Gaussian Mixture Models (GMMs), and a linear transformation is trained for each pair of Gaussians.

A critical point in the performance of MEMLIN [15] is the cross-probability model, which is the *a posteriori* probability of the clean model Gaussian given the noisy one. So, in this work we propose a Cross-Probability Model (CPM) based on a set of GMMs, which will be applied over CD-MEMLIN, defining the technique CD-MEMLIN-CPM. In this approach, noisy feature vectors associated to each pair of Gaussians (clean and noisy) are modeled with a GMM, obtaining a time-dependent and dynamic solution for the CPM.

The proposed unsupervised hybrid solution uses CD-MEMLIN-CPM for feature normalization followed by an acoustic model adaptation method based on linear transformations over an expanded HMM-state space. Hence, clean and normalized spaces are modeled with GMMs and a set of linear transformations is obtained, estimating one linear transformation per pair of Gaussians (clean and normalized) using linear regression. In recognition, each normalized feature vector is recognized with the augMented stAte space acousTic dEcoder (MATE) [16] using expanded acoustic models, which are generated from the reference models and the set of linear transformations. The approach of the proposed hybrid solution is to map the reference acoustic models to the normalized space, compensating those degradations that the feature vector normalization/adaptation techniques are not able to model, like rotations [17].

To compare the performance of the proposed methods, two databases have been used: the Spanish SpeechDat Car database [18], [19], which represents a real dynamic environment, and the widely used Aurora 2 database [20], which does not represent a real environment because the noise has been artificially added, but is a reference database to compare robustness techniques.

This paper is organized as follows. In Section II, a brief overview of CD-MEMLIN is included. In Section III, the cross-probability model importance is studied in a qualitative way and CD-MEMLIN-CPM is proposed. In Section IV, the online unsupervised hybrid solution, which combines CD-MEMLIN-CPM with the novel acoustic model adaptation method based on linear transformations is presented. The results of the proposed techniques with Spanish SpeechDat Car and Aurora 2 databases are detailed in Section V. Finally, the summary, conclusions and future work directions are discussed in Section VI.

II. CD-MEMLIN Overview

Class-Dependent Multi-Environment Model-based LInear Normalization (CD-MEMLIN) is an empirical feature vector normalization/adaptation technique based on a general MMSE framework where each class is modeled with a GMM for the clean and noisy spaces. Hence, three approximations are considered.

A. CD-MEMLIN Approximations

- Clean feature vectors $\mathbf{x}$ are modeled with a GMM of N_x components for each class c

$$p(\mathbf{x}|c) = \sum_{s_{x,c}}^{N_x} p(s_{x,c}|c)p(\mathbf{x}|s_{x,c},c) \tag{1}$$

$$p(\mathbf{x}|s_{x,c},c) = \mathcal{N}\left(\mathbf{x};\boldsymbol{\mu}_{s_{x,c}},\boldsymbol{\Sigma}_{s_{x,c}}\right) \tag{2}$$

  where $\boldsymbol{\mu}_{s_{x,c}}$ $\boldsymbol{\Sigma}_{s_{x,c}}$ and $p(s_{x,c}|c)$ are the mean vector, the diagonal covariance matrix, and the *a priori* probability associated with the clean model Gaussian $s_{x,c}$ for the c class. Note that all classes are modeled with the same number of Gaussians (N_x) for simplicity.
- The noisy space is split into several basic environments, e, which represent different acoustic conditions. Furthermore, the corresponding feature vectors $\mathbf{y}$ are modeled as

a GMM of N_y components for each basic environment and class

$$p(\mathbf{y}|e,c) = \sum_{s^e_{y,c}}^{N_y} p\left(s^e_{y,c}|e,c\right) p\left(\mathbf{y}|s^e_{y,c},e,c\right) \quad (3)$$

$$p\left(\mathbf{y}|s^e_{y,c},e,c\right) = \mathcal{N}\left(\mathbf{y};\boldsymbol{\mu}_{s^e_{y,c}},\boldsymbol{\Sigma}_{s^e_{y,c}}\right) \quad (4)$$

where $s^e_{y,c}$ denotes the corresponding Gaussian of the noisy model for the $e1$ basic environment and c class; $\boldsymbol{\mu}_{s^e_{y,c}}$, $\boldsymbol{\Sigma}_{s^e_{y,c}}$, and $p(s^e_{y,c}|e,c)$ are the mean vector, the diagonal covariance matrix, and the *a priori* probability associated to $s^e_{y,c}$. Again, observe that all classes are modeled with the same number of Gaussians per basic environment (N_y) for simplicity.

- Finally, the relationship between $\mathbf{x}$ and $\mathbf{y}$ is considered linear within each pair of Gaussians per class, $s_{x,c}$ and $s^e_{y,c}$: $p(\mathbf{x}|\mathbf{y},s_{x,c},s^e_{y,c},e,c) = \mathcal{N}(\mathbf{x};\mathbf{y}-\mathbf{r}_{s_{x,c},s^e_{y,c}},\boldsymbol{\Sigma}_{s_{x,c},s^e_{y,c}})$, where $\mathbf{r}_{s_{x,c},s^e_{y,c}}$ is the bias vector transformation between noisy and clean feature vectors associated to each pair of Gaussians ($s_{x,c}$ and $s^e_{y,c}$) and $\boldsymbol{\Sigma}_{s_{x,c},s^e_{y,c}}$ is the corresponding covariance matrix. Although CD-MEMLIN proposes a linear degradation model of the signal space based on a bias vector, different approximations could be considered, such as first order polynomial or even nonlinear estimates [10].

B. CD-MEMLIN Enhancement

In order to estimate the clean feature vector $\hat{\mathbf{x}}_t$ for each time index t, the MMSE estimator is applied combining the three approximations

$$\begin{aligned}\hat{\mathbf{x}}_t &= E[\mathbf{x}|\mathbf{y}_t] \\ &= \sum_{e}^{N_e}\sum_{c}^{N_c}\sum_{s^e_{y,c}}^{N_y}\sum_{s_{x,c}}^{N_x} p\left(e,c,s^e_{y,c},s_{x,c}|\mathbf{y}_t\right) \\ &\quad \times E\left[\mathbf{x}|\mathbf{y}_t,e,c,s^e_{y,c},s_{x,c}\right] \\ &= \mathbf{y}_t - \sum_{e}^{N_e}\sum_{c}^{N_c}\sum_{s^e_{y,c}}^{N_y}\sum_{s_{x,c}}^{N_x} \mathbf{r}_{s_{x,c},s^e_{y,c}} p(e|\mathbf{y}_t)p(c|\mathbf{y}_t,e) \\ &\quad \times p\left(s^e_{y,c}|\mathbf{y}_t,e,c\right) p\left(s_{x,c}|\mathbf{y}_t,e,c,s^e_{y,c}\right) \end{aligned} \quad (5)$$

where the operator $E[]$ is the expected value, N_c is the number of classes, and N_e is the number of basic environments. Here, $p(e|\mathbf{y}_t)$ is the *a posteriori* probability of the basic environment; $p(c|\mathbf{y}_t,e)$ is the *a posteriori* probability of the class c, given the noisy feature vector $\mathbf{y}_t$ and the basic environment; and $p(s^e_{y,c}|\mathbf{y}_t,e,c)$ is the *a posteriori* probability of the noisy model Gaussian $s^e_{y,c}$ given the noisy feature vector $\mathbf{y}_t$, the basic environment e, and the class c. These three terms are computed for each test feature vector combining (3) and (4) in the recognition phase [10]. Finally, the Cross-Probability Model (CPM) $p(s_{x,c}|\mathbf{y}_t,e,c,s^e_{y,c})$, is the probability of the clean model Gaussian $s_{x,c}$, given the noisy feature vector $\mathbf{y}_t$, the basic environment e, the class c, and the noisy model Gaussian $s^e_{y,c}$. The CPM term along with the bias vector transformation, $\mathbf{r}_{s_{x,c},s^e_{y,c}}$, is estimated in an unsupervised training phase using stereo data.

C. CD-MEMLIN Training

Given a stereo data training corpus for each basic environment and class, $(\mathbf{X}_{e,c},\mathbf{Y}_{e,c}) = \{(\mathbf{x}^{e,c}_1,\mathbf{y}^{e,c}_1);\ldots;(\mathbf{x}^{e,c}_{t_{e,c}},\mathbf{y}^{e,c}_{t_{e,c}});\ldots;(\mathbf{x}^{e,c}_{T_{e,c}},\mathbf{y}^{e,c}_{T_{e,c}})\}$, with $t_{e,c}=1,\ldots,T_{e,c}$, the bias vector transformation $\mathbf{r}_{s_{x,c},s^e_{y,c}}$, is estimated by minimizing frame-by-frame the mean weighted square error, with respect to $\mathbf{r}_{s_{x,c},s^e_{y,c}}$ [10].

On the other hand, the cross-probability model $p(s_{x,c}|\mathbf{y}_t,e,c,s^e_{y,c})$ is simplified by avoiding the time dependence given by the noisy feature vector $\mathbf{y}_t$, i.e., $p(s_{x,c}|\mathbf{y}_t,e,c,s^e_{y,c}) \simeq p(s_{x,c}|e,c,s^e_{y,c})$. Thus, the term $p(s_{x,c}|e,c,s^e_{y,c})$ can be obtained with (1), (2), (3), and (4) as

$$\begin{aligned} &p\left(s_{x,c}|e,c,s^e_{y,c}\right) \\ &= \frac{p\left(s^e_{y,c}|e,c\right) p\left(\mathbf{y}|s^e_{y,c},e,c\right) p(s_{x,c}|c) p(\mathbf{x}|s_{x,c},c)}{\sum_{s_{x,c}}^{N_x} p\left(s^e_{y,c}|e,c\right) p\left(\mathbf{y}|s^e_{y,c},e,c\right) p(s_{x,c}|c) p(\mathbf{x}|s_{x,c},c)}. \end{aligned} \quad (6)$$

In summary, CD-MEMLIN estimates a linear model compensation based on a bias vector transformation for each pair of noisy and clean model Gaussians per basic environment and class. Thus, the mapping space associated with each CD-MEMLIN transformation is more enclosed and has less uncertainty than the ones corresponding to RATZ or SPLICE [10]. Note that RATZ assumes a bias vector transformation for each clean model Gaussian, and SPLICE defines a bias vector transformation per noisy model Gaussian. Also observe that MEMLIN is a simplified version of CD-MEMLIN when just one class is considered. On the other hand, if one class per phoneme is defined, CD-MEMLIN would present the same solution as PD-MEMLIN [10].

The clean estimated feature vector for CD-MEMLIN $\hat{\mathbf{x}}_t$ (5) can be seen as a shifted version of the noisy vector $\mathbf{y}_t$: $\hat{\mathbf{x}}_t = \mathbf{y}_t + \mathbf{g}_t$, where $\mathbf{g}_t$ is the complete time dependent bias vector. Thus, a direct correspondence between CD-MEMLIN and the acoustic model adaptation techniques can be appreciated. Usually, $\hat{\mathbf{x}}_t$ is decoded using clean acoustic models. However, it provides the same solution as decoding the noisy feature vector $\mathbf{y}_t$ using adapted acoustic models where the adapted acoustic models are built frame-by-frame modifying just the mean vectors of the clean acoustic models. Thus, for each time index t the adapted mean vectors $\boldsymbol{\mu}^{adap}_t$, would be computed as $\boldsymbol{\mu}^{adap}_t = \boldsymbol{\mu} - \mathbf{g}_t$, where $\boldsymbol{\mu}$ is the corresponding mean vectors of the clean acoustic models. Note that it is assumed that the acoustic models are composed of HMMs with GMMs as the observation generation probability density functions (pdfs) for all the states. This correspondence between CD-MEMLIN and acoustic model adaptation techniques can be applied to all the feature vector normalization/adaptation methods which consist of a linear transformation composed only by a bias vector (e. g., CMN, RATZ, SPLICE).

The supervised training process is a major limitation for use in user friendly applications because active enrollment by the speakers is required. Thus, to maintain the unsupervised framework, a simple two-class definition is used for CD-MEMLIN in

this work, silence and speech, so that a voice activity detector (VAD) can be used to label each training feature vector.

Finally, if CD-MEMLIN is inspected, two critical points can be detected in order to improve the performance of the technique: the degradation model of the signal space, which has been approximated as linear, and the cross-probability model, which has been considered time-independent. The first point was studied for MEMLIN in [10], where different linear and non linear solutions were presented, while this work is focused on the second issue.

III. Cross-Probability Model Based on GMMs

A. Cross-Probability Model Performance

To study the performance of the cross-probability model in a qualitative way, the comparative pdfs and log-scattergrams between the first Mel frequency cepstral coefficients (MFCCs) in nonsilence frames for different signals are depicted in Fig. 1. Fig. 1(a.1) and (a.2) represents the relationship between clean and noisy feature coefficients (the noisy space was selected from the Spanish SpeechDat car database: low speed, rough road). Fig. 1(b.1) and (b.2) shows the relationship between clean and CD-MEMLIN-normalized feature coefficients. Finally, Fig. 1(c.1) and (c.2) represents the relationship between clean and CD-MEMLIN-normalized coefficients when the cross-probability model is computed over $s_{x,c}$ based on the corresponding clean feature vectors as $(p(s_{x,c}|\mathbf{x}_t))$. Note that this is the oracle solution although it could not be applied in a real situation. In all cases, CD-MEMLIN is applied modeling the noisy and clean spaces with 64 Gaussians per class $(N_x = N_y = 64)$.

Fig. 1(a.1) and (a.2) shows clearly the effects of the real car environment. The pdf of clean first MFCCs is affected [Fig. 1(a.1)], shifting the mean and reducing the variance (typical effects of convolutional distortion and additive noise, respectively). Also, the random nature of the environmental degradation increases the uncertainty between the feature coefficients: a given clean feature coefficient can generate different noisy features, and vice versa [Fig. 1(a.2)]. The effects of the noisy environment are compensated when CD-MEMLIN is applied (Fig. 1(b.1) and (b.2)). The pdf of normalized first MFCCs is approximated to the pdf of clean signal first MFCCs [Fig. 1(b.1)], and the uncertainty is reduced, although there is still a considerable uncertainty between clean and normalized coefficients [Fig. 1(b.2)]. The peak that appears in Fig. 1(b.1) is because of the transformation of noisy feature vectors towards the clean silence. This problem could be solved if an efficient VAD were used not only in the training process but also during the normalization process [10]. Note that in spite of the qualitative improvement obtained with CD-MEMLIN, there is still a large mismatch between clean and normalized feature coefficients. Finally, if CD-MEMLIN is applied with the oracle cross-probability model, the pdf of the normalized first MFCCs is almost the same as that of the clean MFCCs [Fig. 1(c.1)], and the uncertainty is drastically reduced [Fig. 1(c.2)]. These results verify the qualitative importance

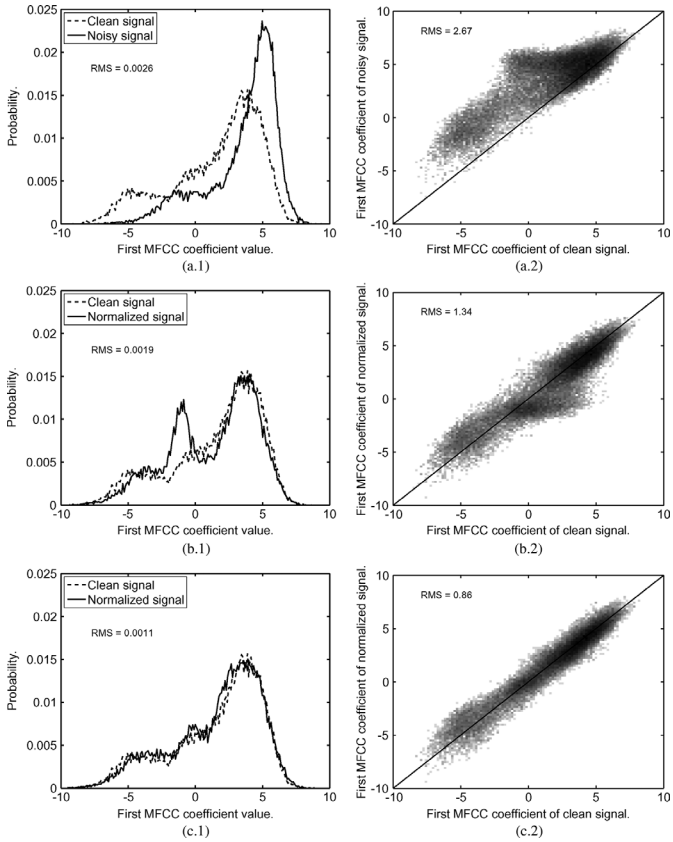

Fig. 1. PDFs and Log-scattergrams between the first MFCC coefficient in nonsilence frames computed over different signals.

of the estimation of the cross-probability model in the performance of CD-MEMLIN algorithm. The same assertion has been concluded for MEMLIN in previous work [15].

B. GMM for Cross-Probability Model

To improve the static cross-probability model (6) for CD-MEMLIN, we propose to model the noisy feature vectors associated to each pair of Gaussians of the acoustic class (speech/silence) c ($s_{x,c}$ and $s^e_{y,c}$) with a GMM of N'_y components

$$p\left(\mathbf{y}_t|s_{x,c}, s^e_{y,c}, e, c\right) = \sum_{s'_y}^{N'_y} p\left(\mathbf{y}_t|s'_y, s_{x,c}, s^e_{y,c}, e, c\right) p\left(s'_y|s_{x,c}, s^e_{y,c}, e, c\right) \quad (7)$$

$$p\left(\mathbf{y}_t|s'_y, s_{x,c}, s^e_{y,c}, e, c\right) = \mathcal{N}\left(\mathbf{y}_t; \boldsymbol{\mu}_{s'_y, s_{x,c}, s^e_{y,c}}, \boldsymbol{\Sigma}_{s'_y, s_{x,c}, s^e_{y,c}}\right) \quad (8)$$

where $\boldsymbol{\mu}_{s'_y, s_{x,c}, s^e_{y,c}}$, $\boldsymbol{\Sigma}_{s'_y, s_{x,c}, s^e_{y,c}}$, and $p(s'_y|s_{x,c}, s^e_{y,c}, e, c)$ are the mean, the diagonal covariance matrix, and the *a priori* probability corresponding to the s'_y Gaussian of the cross-probability model associated to $s_{x,c}$ and $s^e_{y,c}$. To train these three parameters, a previous training process with stereo data for each basic environment and class is applied using the Expectation–Maximization (EM) algorithm [21]. Hence, each noisy training feature vector associated to a basic environment e and class c $\mathbf{y}^{e,c}_{t_{e,c}}$ is labeled with the most probable noisy model Gaussian [(3)

and (4) are used] and the most probable clean model Gaussian, which is computed with the corresponding clean training feature vector $\mathbf{x}_{t_{e,c}}^{e,c}$ using (1) and (2).

Once the cross-probability GMM parameters are estimated, $p(s_{x,c}|e,c,\mathbf{y}_t,s_{y,c}^e)$ can be estimated combining (6), (7), and (8) as

$$p\left(s_{x,c}|e,c,\mathbf{y}_t,s_{y,c}^e\right) = \frac{p\left(\mathbf{y}_t|e,c,s_{x,c},s_{y,c}^e\right)p\left(s_{x,c}|e,c,s_{y,c}^e\right)}{\sum_{s_{x,c}}^{N_x} p\left(\mathbf{y}_t|e,c,s_{x,c},s_{y,c}^e\right)p\left(s_{x,c}|e,c,s_{y,c}^e\right)}. \quad (9)$$

Note that the time-independence assumption considered in the first approach of CD-MEMLIN has been avoided, while the training process is still unsupervised. On the other hand, $p(s_{x,c}|e,c,s_{y,c}^e)$ represents in this case the *a priori* probability of the trained GMMs for the cross-probability model (since the pair of Gaussians $(s_{x,c}, s_{y,c}^e)$ are not equiprobable) and has to be estimated in the training process.

IV. Unsupervised Hybrid Compensation Technique

Previous works [10], [22], [23] show that feature vector normalization/adaptation techniques such as RATZ, SPLICE, and MEMLIN are effective to compensate the effects of noise, obtaining satisfactory improvements. However, these kinds of techniques have intrinsical limitations, e.g., not taking into account several kinds of degradations like rotations, because the feature vector coefficients are considered independent. In order to compensate for this weakness, a linear transformation can be used in the feature vector domain. However, this solution does not result in very good performance [17]. Thus, we propose an unsupervised hybrid compensation technique which combines CD-MEMLIN-CPM with an acoustic model adaptation method based on a set of linear transformations over an expanded HMM-state space.

The scheme of the proposed hybrid compensation technique is depicted in Fig. 2. It is composed of two phases: training and decoding. In the unsupervised training phase, the available clean and noisy stereo training data are used to estimate the parameters of the corresponding feature vector normalization/adaptation method ("Training normalization," which in this case is CD-MEMLIN-CPM). Furthermore, the noisy training feature vectors are compensated using the corresponding method ("Normalization") and a set of linear transformations is estimated with the normalized and clean stereo training data by linear regression ("Matrix estimation"). In the decoding phase, each compensated testing feature vector ("Normalization") is recognized using augMented stAte space acousTic dEcoder (MATE) [16] ("MATE decoder") with expanded acoustic models, which are obtained with the reference acoustic models and the set of linear transformations. During the search process, a linear transformation per frame is implicity selected by ML criterion in a modified Viterbi algorithm which can be seen as a 3-D Viterbi. Note that, although CD-MEMLIN-CPM is the selected feature vector normalization/adaptation technique for the proposed hybrid method in this work, so that a VAD should be included in the training phase to complete the generic scheme showed in Fig. 2, any other algorithm could be used in the same way.

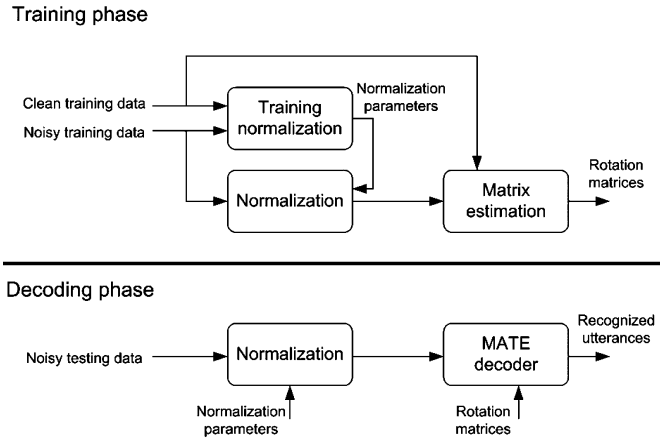

Fig. 2. Scheme of the proposed unsupervised hybrid compensation technique.

A. Matrix Estimation

In order to estimate the set of linear transformations, three approximations are considered.

- Clean feature vectors $\mathbf{x}$ are modeled using a GMM of N_x' components

$$p(\mathbf{x}) = \sum_{s_x}^{N_x'} p(s_x)p(\mathbf{x}|s_x) \quad (10)$$

$$p(\mathbf{x}|s_x) = \mathcal{N}\left(\mathbf{x};\boldsymbol{\mu}_{s_x},\boldsymbol{\Sigma}_{s_x}\right) \quad (11)$$

  where $\boldsymbol{\mu}_{s_x}$, $\boldsymbol{\Sigma}_{s_x}$, and $p(s_x)$ are the mean vector, the diagonal covariance matrix, and the *a priori* probability associated with the clean model Gaussian s_x.
- Normalized feature vectors $\hat{\mathbf{x}}$ are modeled using a GMM of $N_{\hat{x}}$ components

$$p(\hat{\mathbf{x}}) = \sum_{s_{\hat{x}}}^{N_{\hat{x}}} p(\hat{\mathbf{x}}|s_{\hat{x}})p(s_{\hat{x}}) \quad (12)$$

$$p(\hat{\mathbf{x}}|s_{\hat{x}}) = \mathcal{N}\left(\hat{\mathbf{x}};\boldsymbol{\mu}_{s_{\hat{x}}},\boldsymbol{\Sigma}_{s_{\hat{x}}}\right) \quad (13)$$

  where $\boldsymbol{\mu}_{s_{\hat{x}}}$, $\boldsymbol{\Sigma}_{s_{\hat{x}}}$, and $p(s_{\hat{x}})$ are the mean vector, the diagonal covariance matrix, and the *a priori* probability associated with the normalized model Gaussian $s_{\hat{x}}$.
- Given a pair of clean and normalized model Gaussians (s_x and $s_{\hat{x}}$), normalized feature vectors can be approximated as a linear function of the clean feature vectors: $\hat{\mathbf{x}}_t \approx \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t$, where $\mathbf{A}_{s_x,s_{\hat{x}}}$ is the linear transformation between the feature vectors $\hat{\mathbf{x}}_t$ and $\mathbf{x}_t$ associated to the pair of Gaussians s_x and $s_{\hat{x}}$. Observe that a more complex model could be considered, although we select this simple one (only a linear transformation) because it is assumed that other kinds of degradations between noisy and clean frames are compensated by the feature vector normalization/adaptation technique (CD-MEMLIN-CPM in this work)

Thus, a set of linear transformations can be defined as

$$\mathcal{A} = \{\mathbf{A}_{s_x,s_{\hat{x}}}\}_{s_x=1,s_{\hat{x}}=1}^{N_x,N_{\hat{q}x}} = \{\mathbf{A}_n\}_{n=1}^{N} \quad (14)$$

where the index n, which represents each pair of Gaussians s_x and $s_{\hat{x}}$, has been included to simplify the notation. Also, N denotes the number of the pair of the Gaussians: $N = N_{s_x} \times N_{s_{\hat{x}}}$.

In order to estimate the linear transformation $\mathbf{A}_n$, clean and normalized stereo data are used in the previous training phase: $(\mathbf{X}, \hat{\mathbf{X}}) = \{(\mathbf{x}_1, \hat{\mathbf{x}}_1); \ldots; (\mathbf{x}_t, \hat{\mathbf{x}}_t); \ldots; (\mathbf{x}_T, \hat{\mathbf{x}}_T)\}$, with $t = 1, \ldots, T$. Observe that $\hat{\mathbf{X}}$ is obtained by compensating all the noisy training data $\mathbf{Y}$, with the selected feature vector compensation/adaptation technique ("Normalization" in training phase in Fig. 2). Thus, $\mathbf{A}_n$ is estimated by minimizing the defined mean weighted square error $\boldsymbol{\xi}_n$, (15) with respect to $\mathbf{A}_n$ (16) (all the details have been included in Appendix I)

$$\begin{aligned}\boldsymbol{\xi}_n &= \frac{1}{T}\sum_{t}^{T} p(s_x|\mathbf{x}_t)p(s_{\hat{x}}|\hat{\mathbf{x}}_t) \\ &\quad \cdot Tr\left[(\hat{\mathbf{x}}_t - \mathbf{A}_n\mathbf{x}_t)(\hat{\mathbf{x}}_{qt} - \mathbf{A}_n\mathbf{x}_t)^T\right] \qquad (15)\\ \mathbf{A}_n &= \mathbf{A}_{s_x,s_{\hat{x}}} = \underset{\mathbf{A}_n}{\operatorname{argmin}}\{\boldsymbol{\xi}_n\} \\ &= \left[\sum_{t}^{T} p(s_x|\mathbf{x}_t)p(s_{\hat{x}}|\hat{\mathbf{x}}_t)\left(\hat{\mathbf{x}}_t \cdot (\mathbf{x}_t)^T\right)\right] \\ &\quad \cdot \left[\sum_{t}^{T} p(s_x|\mathbf{x}_t)p(s_{\hat{x}}|\hat{\mathbf{x}}_t)\left(\mathbf{x}_t \cdot (\mathbf{x}_t)^T\right)\right]^{-1} \qquad (16)\end{aligned}$$

where the operator $Tr[]$ denotes the trace and $()^T$ is the transpose; $p(s_x|\mathbf{x}_t)$ is the *a posteriori* probability of the clean model Gaussian s_x, given the clean training feature vector $\mathbf{x}_t$, and $p(s_{\hat{x}}|\hat{\mathbf{x}}_t)$ is the *a posteriori* probability of the normalized model Gaussian $s_{\hat{x}}$, given the normalized training feature vector $\hat{\mathbf{x}}_t$. Both probabilities can be estimated by combining (10) and (11), for the first case (17), and (12) and (13) for the second one (18):

$$p(s_x|\mathbf{x}_t) = \frac{p(\mathbf{x}_t|s_x)p(s_x)}{\sum_{s_x}^{N'_x} p(\mathbf{x}_t|s_x)p(s_x)} \qquad (17)$$

$$p(s_{\hat{x}}|\hat{\mathbf{x}}_t) = \frac{p(\hat{\mathbf{x}}_t|s_{\hat{x}})p(s_{\hat{x}})}{\sum_{s_{\hat{x}}}^{N_{\hat{x}}} p(\hat{\mathbf{x}}_t|s_{\hat{x}})p(s_{\hat{x}})}. \qquad (18)$$

B. MATE Decoder

In order to select the linear transformation $\mathbf{A}_t$, associated with each normalized testing feature vector $\hat{\mathbf{x}}_t$, from the set of estimated linear transformations $\mathbf{A}_n$, a Maximum-Likelihood (ML) criterion is applied in the decoding process using the MATE decoder ("MATE decoder" in Fig. 2). Hence, the reference acoustic models are modified in a similar way as described in [16], where the set of linear transformations in this case are the linear transformations $\mathbf{A}_n$ previously estimated. Thus, each state (q) of the reference space HMM acoustic models $(q = 1, \ldots, Q)$ is expanded into N states (q, n) assuming the linear approximation $\hat{\mathbf{x}}_t \approx \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t = \mathbf{A}_n\mathbf{x}_t$. The goal of the state expansion is to reduce the mismatch between the reference space acoustic models and the normalized feature vectors for each linear transformation. Note that each expanded state is specialized in one of the linear transformations previously estimated.

Assuming that a component s_q in the pdf mixture of the original state q follows a normal distribution: $\mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{s_q}, \boldsymbol{\Sigma}_{s_q})$, the corresponding expanded state component for the nth linear transformation $s_{q,n}$, follows the normal distribution $\mathcal{N}(\hat{\mathbf{x}}_t; \mathbf{A}_n\boldsymbol{\mu}_{s_q}, \mathbf{A}_n\boldsymbol{\Sigma}_{s_q}\mathbf{A}_n^T)$. So, the pdf for the expanded state (q, n) $p(\hat{\mathbf{x}}_t|q, n)$ is a GMM composed by the previously defined expanded components, where the *a priori* component probabilities are considered unaltered: $p(s_{q,n}) = p(s_q)$. Hence,

$$p(\hat{\mathbf{x}}_t|q, n) = \sum_{s_q}^{N_{s_q}} p(s_q)\mathcal{N}\left(\hat{\mathbf{x}}_t; \mathbf{A}_n\boldsymbol{\mu}_{s_q}, \mathbf{A}_n\boldsymbol{\Sigma}_{s_q}\mathbf{A}_n^T\right) \qquad (19)$$

where N_{s_q} is the number of the Gaussians of the pdf of the original state q.

To complete the parameter set of the expanded acoustic models, the expanded state transition probabilities $\boldsymbol{\Gamma}$ are computed as

$$\boldsymbol{\Gamma} = a_{q',n',q,n} \simeq \frac{a_{q',q}}{N} \qquad (20)$$

where $a_{q',n',q,n}$ is the transition probability from the expanded state (q', n') to (q, n), and $a_{q',q}$ is the transition probability from the original state (q') to (q), which is part of the reference acoustic model parameters. Observe that $a_{q',n',q,n}$ could be estimated with the EM algorithm [16]. However, in this work the transition probabilities are considered equiprobable for simplicity.

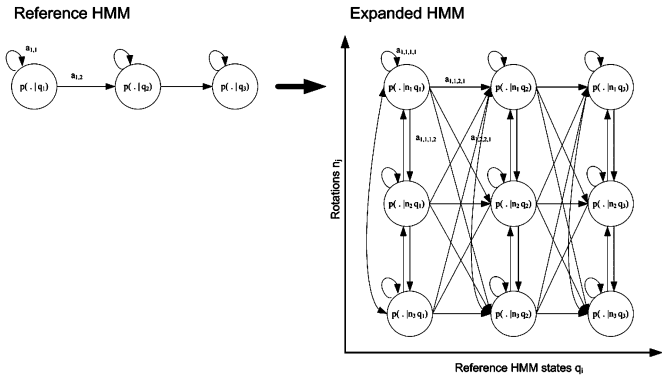

Fig. 3. Example of the proposed expanded acoustic models. A three state left-to-right reference HMM (left side) is transformed using three linear transformations (right side).

A graphical example of the proposed expanded acoustic models is presented in Fig. 3. A standard three state left-to-right reference HMM (left side), which is defined by the distributions $p(.|q_i)$ $i = 1, \ldots, 3$, and the transition probabilities $a_{i,j}$, $i = 1, \ldots, 3$ $j = i, \ldots, 3$, is expanded using three linear transformations (right side). Observe how each reference HMM state is transformed into three new states, modifying the distributions $(p(.|q_i, n_j)$ $i = 1, \ldots, 3$ $j = 1, \ldots, 3)$, and the transition probabilities $(a_{i,j,i',j'}$ $i = 1, \ldots, 3$ $j = 1, \ldots, 3$ $i' = i, \ldots, 3$ $j' = j, \ldots, 3)$. Note that the expanded acoustic models, from a generative point of view, can be seen as a more flexible speech production process under adverse environment conditions. In fact, the expanded acoustic model could generate sequences of rotated feature vectors more suitable to the normalized space.

Once the clean acoustic models have been expanded, the classic Viterbi search algorithm for decoding unlabeled sequences has to be modified to use the new expanded models. Thus, given a normalized testing utterance ("Normalization"

in the decoding phase in Fig. 2), the sequence of expanded states that maximizes the likelihood determines implicity the linear transformation $\mathbf{A}_t$ for each normalized feature vector. The search algorithm under this framework can be performed by computing recursively the score state variable $\phi_{q,n}(t)$ for the state (q, n) and the time index t

$$\phi_{q,n}(t) = \max_{q',n'} \{\phi_{q',n'}(t-1) \cdot a_{q',n',q,n} \cdot p(\hat{\mathbf{x}}_t | q, n)\}. \quad (21)$$

It can be observed that the proposed searching solution is similar to the approach presented in [24], where a 3-D Viterbi algorithm was developed to compensate the effects of non stationary noise. However, note that the presented hybrid solution can be seen as decoding each CD-MEMLIN-CPM-normalized feature vector $\hat{\mathbf{x}}_t = \mathbf{y}_t + \mathbf{g}_t$ with the corresponding expanded acoustic models, where the mean vectors and covariance matrices are adapted frame-by-frame as $\mathbf{A}_t\boldsymbol{\mu}$ and $\mathbf{A}_t\boldsymbol{\Sigma}\mathbf{A}_t^T$, respectively. This solution provides the same results as decoding the noisy feature vector $\mathbf{y}_t$ with adapted acoustic models, where the adapted mean vectors and covariance matrices are modified frame-by-frame as $\mathbf{A}_t\boldsymbol{\mu} - \mathbf{g}_t$ and $\mathbf{A}_t\boldsymbol{\Sigma}\mathbf{A}_t^T$, respectively. From this point of view, the presented hybrid technique is conceptually similar to maximum-likelihood linear regression (MLLR), where linear transformations are included jointly in acoustic models. However, both approaches are quite different. Thus, the linear transformations for the proposed hybrid technique are estimated using a different criterion than MLLR. Furthermore, the unsupervised MLLR version needs a previous step to provide a hypothesis of the transcription of the adaptation data (usually a decoding process); so that the performance of the unsupervised MLLR solution can degrade dramatically in high noise acoustic conditions or when the adaptation task is complex (e.g., large vocabulary, spontaneous speech...) because the hypothesis of the transcription could not be precise enough. These problems do not affect the proposed hybrid technique, which does not require the transcription of the adaptation data. Finally, observe that the proposed hybrid solution does not use a compact solution with extended matrices as in MLLR because of the computational cost. So that, given a certain complexity level, the proposed solution provides a better performance than the compact approach because it is not possible, from a practical point of view, to include all the transformations of CD-MEMLIN in the MATE framework.

V. Experimental Results

To study the performance of the proposed unsupervised online compensation techniques, a set of experiments were carried out using two databases: the first one is the Spanish SpeechDat Car database [18], [19], which is composed of real, dynamic, and complex environments. The second is the Aurora 2 database [20], which does not represent real environments because the noise has been artificially added, but it has been widely used to compare robustness techniques.

In both cases, the recognition task is isolated and continuous digit recognition. As the feature set, the standard ETSI front-end [25] features plus energy are computed every 10 ms using a 25-ms Hamming window. Also, the corresponding delta- and delta-delta-coefficients are included to complete the 39-dimensional feature vectors. Online cepstral mean normalization is applied to testing and training data. The feature vector normalization/adaptation techniques are applied over the 12 MFCCs and energy, whereas the derivatives are computed over the normalized static coefficients. The acoustic models are composed of a 16-state HMM for each digit with a three-state begin-end silence HMM and a one-state inter-word silence HMM. In all cases, each pdf state is composed of a mixture of three Gaussian components.

A. Results With SpeechDat Car Database

Seven basic environments were defined for the Spanish SpeechDat Car database.

- E1: car stopped, motor running.
- E2: town traffic, closed windows, and climatizer off (silent conditions).
- E3: town traffic and noisy conditions (windows open, and/or climatizer on).
- E4: low speed, rough road, and silent conditions.
- E5: low speed, rough road, and noisy conditions.
- E6: high speed, good road, and silent conditions.
- E7: high speed, good road, and noisy conditions.

In this study, two simultaneously recorded channels of the database (stereo data) have been used: the CLose talK channel (CLK), which recorded the clean signal with a Shune SM-10A microphone, and Hands-Free channel (HF), which recorded the noisy signal using a Peiker ME15/V520-1 microphone located on the ceiling of the car in front of the driver. The signal-to-noise ratio (SNR) range for the CLK signals goes from 20 to 30 dB, and the HF SNR goes from 5 to 20 dB. The unsupervised training process has been carried out with the CLK and HF signals of the training set.

A training corpus for each basic environment is needed to learn the corresponding parameters of the proposed techniques: bias vector transformations, cross-probability models, and linear transformations. For this purpose, 16 108 utterances for all basic environments and different tasks: isolated and continuous digits, spelling, dates, commands, and names are used. To train the acoustic models for the connected digits task, a set of the training corpus composed by the digit task utterances are used (1896 utterances for all basic environments). The testing corpus is composed of 1086 utterances for all basic environments and different speakers from the training corpus. The composition of the training and testing corpora is explained in detail in Table I, including the number of utterances and words for each basic environment.

The word error rate (WER) baseline results for each basic environment are presented in Table II, where AWER is the average WER in %, which is computed proportionally to the number of words of each basic environment (see Table I). The "Train" column refers to the signals used to obtain the corresponding acoustic HMMs: if they are trained with all clean training utterances, the column is marked as CLK, and if the column is marked as HF, the acoustic models are trained with all noisy training utterances (multi-condition training). Furthermore HF† indicates that specific acoustic HMMs for each basic environment are applied in decoding (environment match condition).

TABLE I
NUMBER OF UTTERANCES AND WORDS FOR TRAINING AND TESTING CORPORA USED IN ALL THE EXPERIMENTS WITH SPANISH SPEECHDAT CAR DATABASE

	E1	E2	E3	E4	E5	E6	E7	Total
# Utterances train	3,393	3,122	2,356	2,106	2,550	2,038	543	16,108
# Utterances train (digits)	400	368	272	248	304	240	64	1,896
# Words train	10,542	9,652	7,160	6,517	7,908	6,265	1,673	49,717
# Words train (digits)	2,105	1,930	1,431	1,301	1,596	1,249	336	9,948
# Utterances test	199	223	136	152	200	120	56	1,086
# Words test	1,049	1,166	715	798	1,049	630	294	5,701

TABLE II
WER BASELINE RESULTS WITH SPANISH SPEECHDAT CAR DATABASE, IN %, FROM THE DIFFERENT BASIC ENVIRONMENTS $(\mathrm{E1}, \ldots, \mathrm{E7})$

Train	Test	E1	E2	E3	E4	E5	E6	E7	AWER (%)
CLK	CLK	0.95	2.32	0.70	0.25	0.57	0.32	0.00	0.91
CLK	HF	3.05	13.29	15.52	27.32	31.36	35.56	53.06	21.48
HF	HF	3.81	6.86	3.50	3.76	4.96	4.44	3.06	4.63
HF †	HF	1.14	4.37	1.68	2.13	2.10	2.06	23.13	3.42
CLK-AFE	CLK-AFE	1.14	2.32	0.70	0.13	0.48	0.00	0.00	0.88
CLK-AFE	HF-AFE	1.43	4.80	3.50	3.51	6.48	2.38	13.95	4.35

On the other hand, "Test" column indicates which signals are used for decoding: clean (CLK), or noisy (HF). Also results with the ETSI Advanced front-end (AFE) [26] are included in Table II for comparison. They are marked as CLK-AFE and HF-AFE for clean and noisy signals, respectively.

Table II shows the effect of real car conditions, which produces a significant increase in WER for all the basic environments (Train CLK, Test HF) compared to that of the clean conditions, (Train CLK, Test CLK). When acoustic models are retrained (ML criterion) using all basic environment signals (Train HF) the AWER decreases considerably to 4.63%. Finally, the most competitive results (3.42% AWER) are obtained when specific acoustic models are retrained for each basic environment with ML criterion (Train HF †) despite the poor WER reached with the E7 basic environment due to the reduced amount of training data for that condition (64 utterances, see Table I). However, this option is not possible in a real situation because the basic environment can not be known for each testing utterance. Furthermore, observe that AFE provides a very similar performance with matched clean conditions (Train CLK-AFE, Test CLK-AFE), while a significant improvement is reached when noisy testing data is decoded with clean acoustic models (Train CLK-AFE, Test HF-AFE) due to additional Wienner filtering and "SNR-dependent" processing.

In order to study the performance of the proposed techniques, the Average Improvement in WER (AIMP), in %, is defined. Thus, given an AWER, the corresponding AIMP is computed as

$$\mathrm{AIMP} = \frac{100(\mathrm{AWER} - \mathrm{AWER}_{\mathrm{CLK-HF}})}{\mathrm{AWER}_{\mathrm{CLK-CLK}} - \mathrm{AWER}_{\mathrm{CLK-HF}}} \tag{22}$$

where $\mathrm{AWER}_{\mathrm{CLK-CLK}}$ is the average WER obtained under clean conditions (Train CLK, Test CLK: 0.91 in this case), and $\mathrm{AWER}_{\mathrm{CLK-HF}}$ is the baseline (Train CLK, Test HF: 21.48 in this case). So, A 100% AIMP would be achieved when AWER equals the one obtained under clean conditions.

Fig. 4 shows the AIMP for CD-MEMLIN when a different number of Gaussians $(N_c \times N_y)$ per basic environment is considered (4, 8, 16, 32, 64, and 128). Note that the number of Gaussians per basic environment can give us a qualitative idea of the computational cost of the feature vector normalization/adaptation process in the decoding phase. Furthermore, MEMLIN and SPLICE with Environmental Model selection [11], which is the multi-environment version of SPLICE, are included to compare ($N_c = 1$ in both cases). In case of CD-MEMLIN, the class labels for each training feature vector (silence or speech) were obtained with a simple voice activity detector (VAD) based on an energy threshold, which was applied over the clean training feature vectors. It can be observed that CD-MEMLIN produces a consistent improvement with all numbers of Gaussians per basic environment with respect to SPLICE with environmental model selection and MEMLIN, reaching 79.02% AIMP (5.23% AWER). Also CD-MEMLIN reduces the mapping space at the level of the two classes (silence/speech), adapting in a better way the bias vector transformations to the acoustic models and providing smaller projection spaces for the bias vector transformations than other techniques based on the same principles (e.g., RATZ, SPLICE, MEMLIN, etc.). In order to obtain more specific transformations, the number of transformations per Gaussian per basic environment with CD-MEMLIN (N_x) is higher than SPLICE, which is 1, although the computing cost in the normalization process is almost the same because the most costly part is the computation of the scores of the noisy model Gaussians.

Previous work [10] indicates that using more classes (one per phoneme, PD-MEMLIN) can provide better results, but in that case a supervised training process is required. This is shown in Fig. 4, where the AIMP for PD-MEMLIN is also included. In that case the labels for training data are obtained with Viterbi

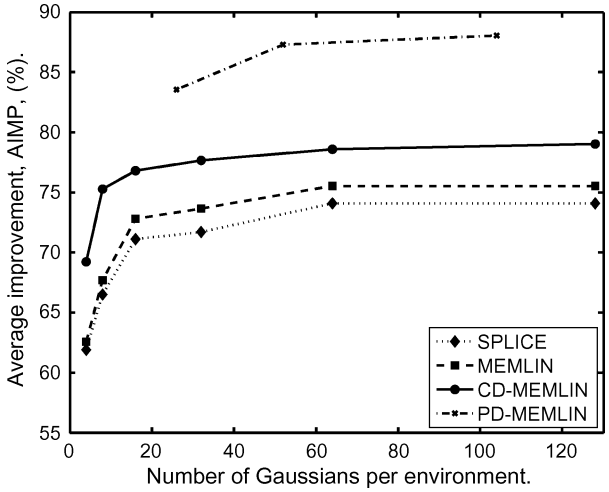

Fig. 4. Average improvement in WER, AIMP, in % with Spanish SpeechDat Car database for different feature vector normalization/adaptation techniques: SPLICE with environmental model selection, MEMLIN, CD-MEMLIN, and PD-MEMLIN with different number of Gaussians per basic environment.

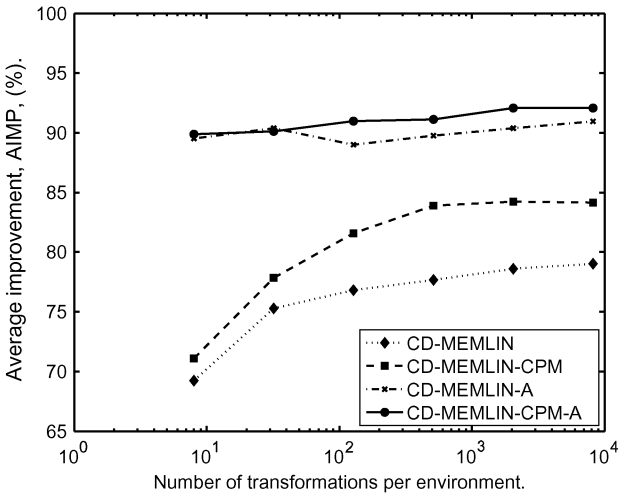

Fig. 5. Average improvement in WER, AIMP, in % with Spanish SpeechDat Car database for different techniques: CD-MEMLIN, CD-MEMLIN with cross-probability model based on GMM, CD-MEMLIN-CPM, and the corresponding hybrid techniques based on CD-MEMLIN-CPM-A and CD-MEMLIN-A.

forced alignment and 1, 2, or 4 Gaussians per phoneme and basic environment are used, so that the total number of Gaussians per basic environment is $N_c \times N_y$, where $N_c = 26$ (25 Spanish phonemes plus the silence). Furthermore, Fig. 4 shows that the performance of SPLICE is improved by MEMLIN in all conditions, as it has been observed in a previous work [10].

The AIMPs for CD-MEMLIN-CPM, the hybrid technique based on CD-MEMLIN-CPM and the acoustic model adaptation based on linear transformations (which is called CD-MEMLIN-CPM-A for simplicity) are depicted in Fig. 5. Also, the results of CD-MEMLIN and the hybrid technique based on CD-MEMLIN (CD-MEMLIN-A) are included for comparison. The AIMP is presented for a range of Transformations per Environment($\mathrm{TpE} = N_c \times N_y \times N_x$). The GMMs for the cross-probability model are composed of two Gaussians while 16 linear transformations are used ($N'_x = N_{\hat{x}} = 4$) to extend the acoustic models. There is a significant improvement

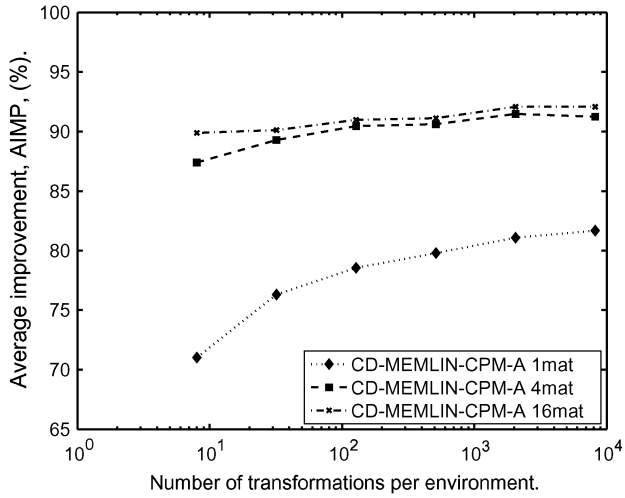

Fig. 6. Average improvement in WER, AIMP, in % with Spanish SpeechDat Car database for CD-MEMLIN-CPM-A with 1, 4, and 16 linear transformations when different number of transformations per basic environment is used.

TABLE III
BEST AVERAGE WER (AWER), AVERAGE IMPROVEMENT IN WER (AIMP) IN %, AND NUMBER OF TRANSFORMATIONS OBTAINED FOR SPANISH SPEECHDAT CAR DATABASE

Train	Test	TpE	AWER (%)	AIMP (%)
HF MLLR	HF	–	5.28	78.77
CLK	HF CD-MEMLIN	8,192	5.23	79.02
CLK	HF CD-MEMLIN-CPM	2,048	4.16	84.22
CLK	HF CD-MEMLIN-CPM-A	2,048	2.54	92.08

(84.22% AIMP, 4.16% AWER) that CD-MEMLIN-CPM obtains with respect to CD-MEMLIN, especially when the basic environments are modeled with high number of Gaussians. Also, it can be verified that the proposed hybrid solution CD-MEMLIN-CPM-A, provides the best results for almost all the transformations per environment, although very similar performance is reached with CD-MEMLIN-A. This indicates that the MATE decoder with linear transformations is a good solution to combine with feature vector normalization/adaptation techniques. In fact, the performance with 32 components per class and the basic environment (92.08% AIMP, 2.54% AWER) is significantly better than any presented result of the proposed techniques in this Section. Even if matched training conditions (Train HF, Test HF: 81.93% AIMP, 4.63% AWER) or environment match conditions (Train HF † Test HF: 87.81% AIMP, 3.42% AWER) are considered, the performances are slightly inferior with respect to the one obtained with CD-MEMLIN-CPM-A. This is due to the fact that the noisy space is more heterogenous than the normalized one. Furthermore, note that a reduced number of Gaussians per class and basic environment is enough to obtain satisfactory results (89.87% AIMP, 3.00% AWER with only two components per class and basic environment). Also, observe that the improvements obtained with CD-MEMLIN-CPM-A, CD-MEMLIN-A and even with CD-MEMLIN-CPM are superior to the one reached with AFE (4.35% AWER). Fig. 6 shows the evolution of the average improvement when CD-MEMLIN-CPM is used with 1, 4, and 16 linear transformations. Increasing from 4 to

TABLE IV
WORD ACCURACY BASELINE RESULTS WITH AURORA 2 DATABASE WHEN STANDARD ETSI FRONT-END AND HTK ARE USED

Clean training, multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restauran	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	98,89	99,03	99,05	99,26	99,06	98,89	99,03	99,05	99,26	99,06	99,17	99,09	99,13	99,07
20 dB	96,75	90,54	97,08	96,20	95,14	90,14	95,86	89,95	94,79	92,69	93,37	95,13	94,25	93,98
15 dB	91,53	72,19	88,55	90,03	85,58	74,52	88,15	73,84	81,24	79,44	86,03	89,09	87,56	83,52
10 dB	75,53	47,61	63,53	72,29	64,74	51,89	66,05	49,27	55,20	55,60	71,94	75,03	73,49	62,83
5 dB	47,34	22,91	30,75	39,08	35,02	26,80	36,28	24,60	24,96	28,16	50,63	50,57	50,60	35,39
0 dB	22,44	5,53	10,71	14,25	13,23	7,12	17,35	10,50	9,50	11,12	24,53	23,64	24,09	14,56
-5dB	10,65	0,12	6,83	6,85	6,11	0,95	8,62	5,28	6,14	5,25	12,90	11,19	12,05	6,95
Average	66,72	47,76	58,12	62,37	**58,74**	50,09	60,74	49,63	53,14	**53,40**	65,30	66,69	**66,00**	**58,06**

Multicondition training, multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restauran	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	98,59	98,67	98,57	98,83	98,67	98,59	98,67	98,57	98,83	98,67	98,62	98,67	98,65	**98,66**
20 dB	97,82	97,94	98,24	97,47	97,87	97,73	97,61	97,61	97,66	97,65	97,67	97,58	97,63	**97,73**
15 dB	96,65	97,43	97,70	96,88	97,17	96,16	96,67	96,60	96,27	96,43	96,50	96,31	96,41	**96,72**
10 dB	94,38	95,47	96,18	94,11	95,04	92,94	94,86	93,71	93,92	93,86	93,83	93,92	93,88	**94,33**
5 dB	89,01	88,21	87,53	87,60	88,09	85,05	86,58	87,53	85,16	86,08	83,11	84,16	83,64	**86,39**
0 dB	67,85	63,18	54,10	63,71	62,21	60,88	63,06	66,27	58,07	62,07	46,21	56,35	51,28	**59,97**

16 linear transformations slightly improves the performance while the computational cost is increased roughly by a factor of 4. However, the use of only 1 linear transformation does not give any additional improvement over the CD-MEMLIN-CPM results.

The most representative results (AWER and AIMP) obtained with the different techniques are summarized in Table III. Furthermore, the performance of unsupervised MLLR, where the transcriptions obtained from the decoding of the noisy training data are assumed as the true ones, is presented in the table to complete the comparison (Train HF MLLR, Test HF). In this case, 12 effective transformations are computed: one transformation per digit and two more for the short inter-word silence and the long silence. Note that the performance in this case (78.77%AIMP, AWER 5.28%) is inferior to that obtained with matched training condition and the unsupervised proposed techniques.

No constraints have been assumed in estimating the set of linear transformations $\mathbf{A}_n$ so that the covariance matrices of the expanded states could not be diagonal $(\mathbf{A}_n\mathbf{\Sigma}_{\mathbf{s_q}}\mathbf{A}_n^T)$. However, in order to present a fair comparison, the covariance matrices were diagonalized before they were used in the MATE decoder. Observe that the diagonalization of the covariance matrices would not be needed if the feature vectors were transformed with $\mathbf{A}_n^{-1}$ and the log-determinant of the Jacobian was added in the likelihood computation $(log(\mathbf{A}_n))$. However, we discarded this solution because it is less efficient than the proposed approach, although a small improvement could be obtained.

The computational cost associated to CD-MEMLIN-CPM or CD-MEMLIN-CPM-A is more expensive than CD-MEMLIN, although the number of transformations per basic environment is the same. However, some solutions can be proposed in order to reduce the computational cost. Thus, the static cross-probability model (6) could be used to determine the *a priori* most probable pair of Gaussians, so that not all the GMMs trained to model the noisy feature vectors associated to each pair of Gaussians would need to be computed, just the most probable ones. On the other hand, the computational cost of the MATE decoder, which is required in the proposed hybrid solution, could be reduced if classic pruning techniques were considered [27].

B. Results With Aurora 2 Database

For the Aurora 2 work, identical utterances from the clean training set and the multicondition training set have been used in the unsupervised training process for the proposed techniques. Thus, the noise types from set B and C are kept as unseen conditions, while the system is tuned on the noise types from set A. Furthermore, some SNRs remain unseen even for set A (0 dB and -5 dB), because they are not included in the multicondition training set. 422 utterances per kind of noise and SNR are used in the training phase (20 basic environments: 4 kinds of noise $\times$ 5 different SNRs), while the testing set contains 70 070 utterances. The testing and training tasks are continuous and isolated digits. Since the purpose of the presented techniques is to reduce the mismatch between training and recognition spaces, we present only the results for clean training and multicondition testing. All the improvements we present in this section are computed with respect to the results reached with standard ETSI front-end and HTK [28] (typical reference system: 58.06% average word accuracy [20]). Complete baseline results, including clean and multicondition training, have been included in Table IV.

The average improvements (AIMP) for CD-MEMLIN, CD-MEMLIN-CPM, and CD-MEMLIN-CPM-A with Aurora 2 database are depicted in Fig. 7 for different numbers of Gaussians per environment. As it has been observed previously that CD-MEMLIN-CPM-A provides a slight improvement with respect to CD-MEMLIN-A, CD-MEMLIN-A results have not been included in this subsection. In this case, AIMP is computed as the average of the improvements of the different recognition conditions because the high variability of the SNRs. Thus, the high and medium SNR environments, which are the more interesting conditions in real applications, are

TABLE V
WORD ACCURACY AND IMPROVEMENT OBTAINED FOR AURORA 2 DATABASE WITH THE PROPOSED HYBRID TECHNIQUE BASED ON CD-MEMLIN-CPM AND ACOUSTIC MODEL ADAPTATION BASED ON LINEAR TRANSFORMATIONS (CD-MEMLIN-CPM-A)

Clean training, multicondition testing															
	A					B					C				Percentage
	Subway	Babble	Car	Exhibition	Average	Restauran	Street	Airport	Station	Average	Subway M	Street M	Average	Average	Improvement
Clean	99,20	99,03	99,22	99,29	99,19	99,20	99,03	99,22	99,29	99,19	98,89	99,09	98,99	99,15	6,90%
20 dB	98,46	98,22	98,45	98,24	98,34	98,31	97,92	98,04	98,52	98,20	97,85	97,70	97,78	98,17	63,97%
15 dB	97,94	97,37	97,67	97,41	97,60	96,86	96,91	96,50	97,36	96,91	96,57	96,00	96,28	97,06	79,29%
10 dB	95,83	93,76	95,26	95,50	95,09	92,70	91,86	92,21	94,10	92,72	91,61	91,56	91,58	93,44	81,04%
5 dB	90,99	81,77	86,40	89,21	87,09	81,95	79,81	81,46	83,12	81,59	75,77	79,17	77,47	82,97	72,73%
0 dB	72,93	54,01	62,89	72,06	65,47	57,35	56,17	60,17	58,75	58,11	44,43	52,27	48,35	59,10	51,71%
-5dB	41,35	27,49	29,81	40,03	34,67	29,50	29,32	31,36	29,88	30,02	20,81	26,82	23,82	30,64	25,30%
Average	91,23	85,03	88,14	90,49	**88,72**	85,44	84,53	85,67	86,37	**85,50**	81,24	83,34	**82,29**	**86,15**	
Improvement	71,89%	77,49%	70,48%	72,26%	73,03%	76,96%	63,00%	76,53%	75,25%	72,94%	58,08%	55,54%	56,81%		**69,75%**
ETSI Adv.	56,32%	77,13%	70,10%	62,72%	66,57%	74,24%	63,17%	79,30%	76,35%	73,27%	59,91%	54,84%	57,37%		**67.41%**

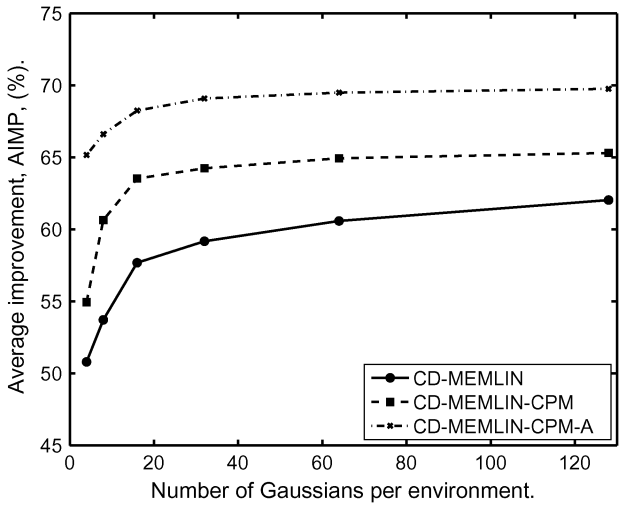

Fig. 7. Average improvement in word accuracy in % with Aurora 2 database for different techniques: CD-MEMLIN, CD-MEMLIN with cross-probability model based on GMM (CD-MEMLIN-CPM), and the proposed hybrid technique based on CD-MEMLIN-CPM and acoustic model adaptation method based on linear transformations (CD-MEMLIN-CPM-A). Different numbers of transformations per basic environment are used.

more important in the computation of the AIMP. The GMMs for the cross-probability model are composed of two Gaussians ($N'_y = 2$) and again 16 linear transformations are used to extend the HMM-based acoustic models ($N'_x = N_{\hat{x}} = 4$). It can be observed how CD-MEMLIN provides a consistent improvement over the baseline for all numbers of transformations per basic environment, reaching 62.03% AIMP (82.76% average word accuracy) when 64 Gaussians per class and basic environment are used. It has been observed with the Spanish SpeechDat Car database, CD-MEMLIN-CPM overcomes the performance obtained with CD-MEMLIN for all numbers of Gaussians (65.31% AIMP, 84.11% average word accuracy), showing again the good performance of the proposed GMM-based model for the cross-probability. Finally, Fig. 7 also includes the AIMP for the proposed hybrid solution when CD-MEMLIN-CPM is used as feature vector normalization/adaptation technique (CD-MEMLIN-CPM-A). Again, a reasonable and consistent improvement for all the number of Gaussins per environment is obtained with respect to CD-MEMLIN and CD-MEMLIN-CPM (69.75% AIMP, 86.15% average word accuracy). In fact, they are slightly better than the ones obtained with ETSI Advanced Front-End (AFE) (67.41% MIMP, 85.97% average word accuracy). Furthermore, fewer Gaussians are needed to reach very competitive results (65.16% AIMP, 84.17% average word accuracy with only two Gaussians per class and basic environment). Note also that the proposed hybrid solution over CD-MEMLIN-CPM even overcomes the performance obtained with multicondition training for high SNR (clean, 20 dB, and 15 dB), although no transcription is used in the training process as multicondition training does.

In order to compare the best results reached with the proposed techniques concerning ETSI AFE, the complete best results (CD-MEMLIN-CPM-A with 64 Gaussians per class and basic environment, two Gaussians to model the cross-probability model and 16 linear transformations) are included in Table V. An important improvement can be observed in set A, 73.03% AIMP (88.72% average word accuracy), because the training process of the proposed hybrid technique is applied over the same kinds of noise. Also, competitive results have been obtained for set B (72.94% AIMP, 85.50% average word accuracy). Although set B includes unseen types of additive noise in the training phase, the average improvement is quite similar to that of set A. The performance is not as competitive for set C (56.81% AIMP, 82.29% average word accuracy), whose utterances include unseen convolutional distortion and additive noise. Thus, we can conclude that the transformations that have been learned in the training process may not compensate for the degradation produced by the environments of set C. Similar conclusions could be obtained from the complete results of CD-MEMLIN or CD-MEMLIN-CPM, hence they have not been included. Compared to the ETSI AFE, the behavior of the proposed hybrid technique under seen conditions (set A) is much better (73.03% AIMP versus 66.57% AIMP in Set A), while for set B and set C the average improvements are slightly inferior. From these results, we can conclude that a reasonable future approach could be to improve the performance of the presented technique under unseen conditions.

The proposed hybrid solution (CD-MEMLIN-CPM-A) provides very satisfactory performance in medium and high SNR conditions, which are the most important ones in real applications, obtaining more than 93% average word accuracy for 10 dB, 15 dB, 20 dB, and clean conditions.

TABLE VI
WORD ACCURACY AND IMPROVEMENT OBTAINED FOR AURORA 2 DATABASE WITH THE PROPOSED HYBRID TECHNIQUE BASED ON CD-MEMLIN-CPM AND ACOUSTIC MODEL ADAPTATION BASED ON LINEAR TRANSFORMATIONS (CD-MEMLIN-CPM-A). ETSI ADVANCED FRONT-END (AFE) IS USED

Clean training, multicondition testing															
	A					B					C				Percentage
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average	Improvement
Clean	99,14	99,00	99,17	99,35	99,17	99,14	99,00	99,17	99,36	99,17	99,05	98,94	99,00	99,13	5,78%
20 dB	98,37	98,46	98,72	98,27	98,46	98,56	97,83	98,81	98,92	98,53	98,10	98,16	98,13	98,42	67,81%
15 dB	97,34	97,59	98,15	97,20	97,57	97,55	97,13	97,71	97,66	97,51	97,18	96,77	96,98	97,43	81,09%
10 dB	94,68	95,17	96,30	96,26	95,35	95,07	94,07	95,23	96,22	95,15	93,72	93,21	93,46	94,89	84,66%
5 dB	88,83	88,13	91,30	88,93	89,32	86,99	87,75	88,87	89,68	88,32	85,47	82,66	84,08	87,87	80,29%

To complete the experiments, the hybrid technique CD-MEMLIN-CPM-A was applied with the ETSI AFE. The results, which are included in Table VI, show that the proposed solution is compatible with the ETSI AFE, providing consistent and significant improvements with respect to CD-MEMLIN-CPM-A with the standard ETSI (Table V), specifically for low SNR environments, while the performance for medium-high SNR environments is also more competitive (more than 94% average word accuracy is obtained for 10 dB, 15 dB, 20 dB, and clean conditions). Furthermore, results obtained for unseen conditions (set B and set C), overcome the ones obtained with the standard ETSI because of the robust front-end.

VI. SUMMARY, CONCLUSION, AND FUTURE WORK

All the proposed techniques in this work need an unsupervised training phase (no transcription is used), so these methods can provide user friendly solutions for ASR applications as they do not require active enrollment by the speakers.

Although satisfactory performance was obtained with PD-MEMLIN in previous works, in this case a two-class simplified version (CD-MEMLIN) is applied as a starting point in order to use an unsupervised stereo data based training process. Thus, it is assumed that each acoustic class (silence/speech) of clean and noisy spaces can be modeled with GMMs and a linear transformation which is trained for each pair of Gaussians. This method has been shown to be more effective than techniques based on similar framework, e.g., SPLICE or MEMLIN, because the mapping space associated with each CD-MEMLIN transformation is more constrained and has less uncertainty.

Also, a study of the cross-probability model of CD-MEMLIN (the *a posteriori* probability of the clean model Gaussian given the noisy model Gaussian) has been provided. Qualitative results have demonstrated that this model is a critical point in CD-MEMLIN, and important improvements can be obtained if it is properly estimated. In this paper, we propose a solution which consists of modeling the noisy feature vectors associated with each pair of Gaussians with a GMM. This approach applied over CD-MEMLIN defines the CD-MEMLIN with Cross-Probability Model based on GMMs (CD-MEMLIN-CPM), which provides a consistent improvement over CD-MEMLIN.

Finally, in order to compensate for some kinds of degradations such as rotations, an online unsupervised hybrid compensation technique has been proposed in this work. The hybrid solution is composed of the combination of a feature vector normalization/adaptation technique (CD-MEMLIN-CPM in this case) and an acoustic model adaptation technique based on a set of linear transformations. Clean and normalized spaces are modeled following both GMMs, so that a linear transformation is defined for each pair of Gaussians (clean model and normalized model). The linear transformations are estimated with clean and CD-MEMLIN-CPM-normalized training data by linear regression. Thus, in testing, each CD-MEMLIN-CPM normalized frame is decoded using augMented stAte space acousTic dEcoder (MATE). In order to use MATE decoder the reference acoustic models are expanded using the linear transformations. The results show that the hybrid solution clearly overcomes the performance of the proposed feature vector normalization/adaptation techniques because the clean reference acoustic models are mapped into the normalized space, even beating the performance of ETSI AFE with Spanish SpeechDat Car and Aurora 2 databases. Finally, it can be observed that the proposed hybrid solution and ETSI AFE are compatible, obtaining very competitive results for clean condition evaluation with Aurora 2 database when they are combined (near 90% average word accuracy). Furthermore, it has been observed that just a few Gaussians per basic environment are needed to obtain satisfactory results.

However, some of the most satisfactory techniques we have presented in this work have three main limitations: the computational cost, the need for stereo data in the training process and the limited improvement under unseen recognition environments.

The high computational cost of the hybrid solution is because of the MATE decoder, which can be seen as a 3-D Viterbi because it is composed of three axes (time, states, and linear transformations) instead of the standard two axes (time and states). In order to minimize the computational cost, classic pruning techniques can be considered to remove the unlikely paths. Also, a reduced number of linear transformations can be used without seriously decreasing the performance as it has been shown. On the other hand, the cross-probability model based on GMMs also increases the computational cost with respect to the solutions based on *a priori* cross-probability model. In this case, only the GMMs associated with the most probable *a priori* pair of Gaussians should be computed, so the computational cost would be dramatically reduced.

Sometimes, stereo training data are not available. In these situations, standard empirical feature vector normalization/adaptation techniques cannot be applied and a nonstereo training process has to be developed. Some non stereo ("blind") solutions have been presented in previous works [9], [10] to overcome this weakness. However, the final performance is not satisfactory. Thus, to develop a competitive "blind" training process for the techniques we have introduced in this work is one of the research lines in which we are working on.

Although satisfactory performance is obtained for the proposed hybrid solution based on CD-MEMLIN-CPM in medium and high SNR scenarios with Aurora 2 database (specifically with set A), the results are not as satisfactory for unseen conditions (set B and set C).

APPENDIX I
ESTIMATION OF ROTATION MATRICES

Let a set of clean and normalized training stereo data to learn the corresponding linear transformations $(\mathbf{X},\hat{\mathbf{X}}) = \{(\mathbf{x}_1,\hat{\mathbf{x}}_1);\ldots;(\mathbf{x}_t,\hat{\mathbf{x}}_t);\ldots;(\mathbf{x}_T,\hat{\mathbf{x}}_T)\}$, with $t = 1,\ldots,T$. Thus, the mean weighted square error $\boldsymbol{\xi}_n$ is defined for each pair of Gaussians s_x and $s_{\hat{x}}$ as

$$\boldsymbol{\xi}_n = \frac{1}{T}\sum_{t}^{T} p(s_x|\mathbf{x}_t)p(s_{\hat{x}}|\hat{\mathbf{x}}_t)\cdot Tr\left[(\hat{\mathbf{x}}_t-\mathbf{A}_n\mathbf{x}_t)(\hat{\mathbf{x}}_t-\mathbf{A}_n\mathbf{x}_t)^T\right]. \tag{B.1}$$

In order to estimate the linear transformation $\mathbf{A}_{s_x,s_{\hat{x}}}$, the defined mean weighted square error (B.1) is minimized with respect to $\mathbf{A}_{s_x,s_{\hat{x}}}$, applying some basic matrix properties

$$\begin{aligned}\frac{\partial\boldsymbol{\xi}_{s_x,s_{\hat{x}}}}{\partial\mathbf{A}_{s_x,s_{\hat{x}}}} &= \frac{1}{T}\sum_{t}^{T} p(s_x|\mathbf{x}_t)p(s_{\hat{x}}|\hat{\mathbf{x}}_t)\frac{\partial}{\partial\mathbf{A}_{s_x,s_{\hat{x}}}}\\ &\quad\times\Big[Tr\Big[\hat{\mathbf{x}}_t(\hat{\mathbf{x}}_t)^T-\hat{\mathbf{x}}_t(\mathbf{x}_t)^T\left(\mathbf{A}_{s_x,s_{\hat{x}}}\right)^T\\ &\qquad-\mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t(\hat{\mathbf{x}}_t)^T\\ &\qquad+\mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t(\mathbf{x}_t)^T\left(\mathbf{A}_{s_x,s_{\hat{x}}}\right)^T\Big]\Big]\\ &= \mathbf{0}.\end{aligned} \tag{B.2}$$

Thus,

$$\begin{aligned}\mathbf{0} = \frac{1}{T}\sum_{t}^{T} p(s_x|\mathbf{x}_t)p(s_{\hat{x}}|\hat{\mathbf{x}}_t)\left(-\hat{\mathbf{x}}_t(\mathbf{x}_t)^T-\hat{\mathbf{x}}_t(\mathbf{x}_t)^T\right.\\ \left.+\mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t(\mathbf{x}_t)^T+\mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t(\mathbf{x}_t)^T\right).\end{aligned} \tag{B.3}$$

Finally, it is obtained the corresponding expression for $\mathbf{A}_{s_x,s_{\hat{x}}}$

$$\begin{aligned}\mathbf{A}_{s_x,s_{\hat{x}}} &= \sum_{t}^{T} p(s_x|\mathbf{x}_t)p(s_{\hat{x}}|\hat{\mathbf{x}}_t)\hat{\mathbf{x}}_t(\mathbf{x}_t)^T\\ &\quad\times\left(\sum_{t}^{T} p(s_x|\mathbf{x}_t)p(s_{\hat{x}}|\hat{\mathbf{x}}_t)\mathbf{x}_t(\mathbf{x}_t)^T\right)^{-1}.\end{aligned} \tag{B.4}$$

REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 3, no. 16, pp. 261–291, 1995.

[2] R. M. Stern, B. Raj, and P. J. Moreno, "Compensation for environmental degradation in automatic speech recognition," in *Proc. ESCA Tutorial Research Workshop Robust Speech Recognition for Unknown Communication Channels*, Pont-au-Mousson, France, Apr. 1997, pp. 33–42.

[3] N. Hanai and R. M. Stern, "Robust speech recognition in the automobile," in *Proc. ICSLP*, Yokohama, Japan, Sep. 1994, pp. 1339–1342.

[4] U. Yapanel, X. Zhang, and J. Hansen, "High performance digit recognition in real car environments," in *Proc. ICSLP*, Denver, CO, Sep. 2002, pp. 793–796.

[5] H. Hermansky and N. Morgan, "RASTA processing for speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[6] D. Y. Kim, C. K. Un, and N. S. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," *IEEE Trans. Signal Process.*, vol. 5, no. 3, pp. 57–59, Mar. 1998.

[7] A. Acero, "Acoustical and environmental robustness in automatic speech recognition," Ph.D. dissertation, Elect. Comput. Eng. Dept., Carnegie Mellon Univ., Pittsburgh, PA, Sep. 1990.

[8] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[9] P. Moreno, "Speech recognition in noisy environments," Ph.D. dissertation, Elect. Comput. Eng. Dept., Carnegie-Mellon Univ., , Apr. 1996.

[10] L. Buera, E. Lleida, A. Miguel, A. Ortega, and O. Saz, "Cepstral vector normalization based on stereo data for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 15, no. 3, pp. 1098–1113, Mar. 2007.

[11] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the AURORA2 database," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 217–220.

[12] L. Neumeyer and M. Weintraub, "Probabilistic optimum filtering for robust speech recognition," in *Proc. ICASSP*, Adelaide, Australia, Apr. 1994, vol. 1, pp. 417–420.

[13] L. Neumeyer and M. Weintraub, "Robust speech recognition in noise using adaptation and mapping techniques," in *Proc. ICASSP*, Detroit, MI, May 1995, vol. 1, pp. 141–144.

[14] M. Padmanabhan, G. Saon, and G. Zweig, "Lattice-based unsupervised MLLR for speaker adaptation," in *Proc. ASR*, 2000, vol. 2, pp. 128–132.

[15] L. Buera, E. Lleida, J. Nolazco, A. Miguel, and A. Ortega, "Time-dependent cross-probability model for multi-environment model based linear normalization," in *Proc. ICSLP*, Sep. 2006, pp. 1555–1558.

[16] A. Miguel, E. Lleida, A. Juan, L. Buera, A. Ortega, and O. Saz, "Local transformation models for speech recognition," in *Proc. ICSLP*, Pittsburgh, PA, 2006, pp. 1598–1601.

[17] S. Molau, "Normalization in the acoustic feature space for improved speech recognition," Ph.D. dissertation, Univ. of Aachen, Aachen, Germany, Feb. 2003.

[18] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, "Speechdat-car. A large speech database for automotive environments," in *Proc. LREC*, Athens, Greece, 2000, vol. 2, pp. 895–900.

[19] H. van den Heuvel, J. Boudy, R. Comeyne, S. Euler, A. Moreno, and G. Richard, "The speechdat-car multilingual speech databases for in-car applications: Some first validation results," in *Proc. Eurospeech*, Budapest, Hungary, Sep. 1999, vol. 5, pp. 2279–2282.

[20] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR2000*, Paris, France, Sep. 2000, pp. 29–32.

[21] A. P. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 9, no. 1, pp. 1–37, 1977.

[22] P. Moreno, "Speech Recognition in Noisy Environments," Ph.D. dissertation, Elect. Comput. Eng. Dept., Carnegie Mellon Univ., Pittsburgh, PA, Apr. 1996.

[23] J. Droppo, L. Deng, and A. Acero, "Evaluation of the splice algorithm on the Aurora2 database," in *Proc. Eurospeech*, Sep. 2001, vol. 1, pp. 217–220.

[24] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1990, pp. 845–848.

[25] ETSI, "Speech processing transmission and quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms," Apr. 2000, ETSI ES 201 108 version 1.1.2, Tech. Rep..

[26] ETSI, "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," Oct. 2002, ETSI ES 202 050 version 1.1.1, Tech. Rep..

[27] H. Ney and S. Ortmanns, "Dynamic programming search for continuous speech recognition," *IEEE Signal Process. Mag.*, vol. 16, no. 5, pp. 64–83, Sep. 1999.

[28] S. Young, G. Evermann, M. Gales, T. Hain, D. Tershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woolland, The "HTK Book (for HTK Version 3.3)," Cambridge Univ. Eng. Dept., Apr. 2005.

Luis Buera was born in Lleida, Spain, in 1978. He received the M.Sc. degree in telecommunication engineering and the Ph.D. degree in signal processing from the University of Zaragoza (UZ), Zaragoza, Spain, in 2002 and 2007, respectively.

From 2002 to 2007, he was with the Communications Technologies Group, Department of Electronic Engineering and Communications, UZ, under a research grant. Currently, he is with the Speech Technology Group, Toshiba Research Europe, Cambridge, U.K.

Antonio Miguel was born in Zaragoza, Spain, in 1977. He received the M.Sc. degree in telecommunication engineering and the Ph.D. degre from the University of Zaragoza (UZ), Zaragoza, Spain, in 2001 and 2008.

From 2000 to 2006, he was with the Communications Technologies Group, Department of Electronic Engineering and Communications, UZ, under a research grant. Since 2006, he has been an Assistant Professor in the same department. Currently, his research interest lies in the field of acoustic modeling for ASR.

Óscar Saz was born in Zaragoza, Spain, in 1980. He received the M.Sc. degree in telecommunication engineering from the University of Zaragoza (UZ), Zaragoza, Spain, in 2004. He is currently working towards the Ph.D. degree from UZ.

From 2004 to 2006, he was with the Communications Technologies Group, Department of Electronic Engineering and Communications, UZ, under a research grant. Currently, his research interests are in the field of speaker adaptation and personalization of ASR systems, specially oriented to users with pathological speech.

Alfonso Ortega was born in Teruel, Spain, in 1976. He received the M.Sc. degree in telecommunication engineering and the Ph.D. degree from the University of Zaragoza (UZ), Zaragoza, Spain, in 2000 and 2005, respectively.

In 1999, he joined, under a research grant, the Communications Technologies Group, UZ, where he has been an Assistant Professor since 2001. He is also involved as a Researcher with the Aragon Institute of Engineering Research (I3A). Currently, his research interest lies in the signal processing field applied to speech technologies.

Eduardo Lleida (M'89) was born in Spain in 1961. He received the M.Sc. degree in telecommunication engineering and the Ph.D. degree in signal processing from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 1985 and 1990, respectively.

From 1986 to 1988, he was involved in his doctoral work at the Department of Signal Theory and Communications, UPC. From 1989 to 1990, he was an Assistant Professor and from 1991 to 1993, he was an Associate Professor in the Department of Signal Theory and Communications, UPC. From February 1995 to January 1996, he was a consultant in speech recognition with AT&T Bell Laboratories, Murray Hill, NJ. Currently, he is an Associate Professor of signal theory and communications in the Department of Electronic Engineering and Communications, University of Zaragoza, Zaragoza, Spain, Spain, where he is heading a research team in speech recognition and signal processing. He is managing several speech-related project in Spain. He has coauthored more than 100 technical papers in the field of speech and speaker recognition, speech enhancement and recognition in adverse acoustic environments, acoustic modeling, confidence measures, and spoken dialogue systems.