# CROSS-PROBABILITY MODEL BASED ON GMM FOR FEATURE VECTOR NORMALIZATION IN CAR ENVIRONMENTS

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

Communication Technologies Group (GTC).
Aragon Institute of Engineering Research (I3A) University of Zaragoza, Spain.
{lbuera,amiguel,lleida,ortega,oskarsaz}@unizar.es

**ABSTRACT**

In previous works, in order to develop a robust man-machine interface based on speech for cars, Multi-Environment Model based LInear Normalization, MEMLIN, was presented and it was proved to be effective to compensate environment mismatch. MEMLIN is an empirical feature vector normalization technique which models clean and noisy spaces with Gaussian Mixture Models, GMMs; and the probability of the clean model Gaussian, given the noisy model one and the noisy feature vector (cross-probability model) is a critical point. In previous works the cross-probability model was approximated as time-independent in a training process. However, in this paper, an estimation based on GMM is considered for MEMLIN. Some experiments with SpeechDat Car and Aurora 2 databases were carried out in order to study the performance of the proposed estimation of the cross-probability model, obtaining important improvements: 75.53% and 62.49% of mean improvement in Word Error Rate, WER, for MEMLIN with SpeechDat Car and a reduced set of Aurora2 database, respectively (82.86% and 67.52% if time-independent cross-probability model is applied). Although the behaviour of the technique is satisfactory, using clean acoustic models in decoding produces a mismatch because the normalization is not perfect. So, retraining acoustic models in the normalized space is proposed, reaching 97.27% of mean improvement with SpeechDat Car database.

***Index Terms—*** Feature vector normalization, MEMLIN, GMM, retraining.

## 1. INTRODUCTION

Since cars are more and more considered as business offices, drivers need a safe way to communicate and interact with either other humans or machines. For safety reason, traditional visual and tactile man-machine interfaces, such as displays, buttons and knobs are not satisfactory but speech, as the most convenient and natural way of communication, is an appropriate and complementary solution which can reduce distractions. Hence, Automatic Speech Recognition (ASR) provides safety and comfort, and it is possible to follow the philosophy "Eyes on the road and hands on the steering wheel", which should drive every in-vehicle system design. The problem of robust ASR in car environments has attracted much attention in the recent years and a new market demands for systems which allow the driver to control non critical devices or tasks like phone dialing, RDs-tuner, air conditioner, satellite navigation systems, remote information services access, Web browsing... For this purpose, robustness in challenging car environment still needs to be improved.

When training and testing acoustic conditions differ, the accuracy of ASR systems rapidly degrades. To compensate for this mismatch, robustness techniques have been developed along the following two main lines of research: acoustic model adaptation methods, and feature vector adaptation/normalization methods. Also, hybrid solutions, which are effective under certain conditions, can be generated by combining both kind of techniques, [1]. In general, acoustic model adaptation methods produce the best results [2] because they can model the uncertainty caused by the noise statistics. However, these methods require more data and computing time than do feature vector adaptation/normalization methods, which do not produce as good results but provide more on line solutions. So, finally, the choice of a robustness technique depends on the characteristics of the application in each situation.

Feature vector adaptation/normalization methods fall into one of three main classes [3]: high-pass filtering, which contains very simple methods such Cepstral Mean Normalization, CMN, model-based techniques, which assumes a structural model of environmental degradation, and empirical compensation, which uses direct cepstral comparisons. In any case, and independently of the class, some algorithms assume a prior probability density function (pdf) for the estimation variable. In those cases, a Bayesian estimator can be used to estimate the clean feature vector. The most commonly used criterion is to minimize the Mean Square Error (MSE), and the optimal estimator for this criterion, Minimum Mean Square Error (MMSE), is the mean of the posterior pdf. Methods, such as Stereo-based Piecewise Linear Compensation for Environments (SPLICE) [4], or Multi-Environment Model-based LInear Normalization (MEMLIN) [5] use the MMSE estimator to compute the estimated clean feature vector.

Previous works [5] show that MEMLIN is effective to compensate the effects of dynamic and adverse car conditions. MEMLIN is an empirical feature vector normalization technique based on stereo data and the MMSE estimator. MEMLIN splits the noisy space into several basic environments and each of them and clean feature space are modelled using GMMs. Therefore, a bias vector transformation is associated with each pair of Gaussians from the clean and the noisy basic environment spaces. A critical point in MEMLIN is the estimation of the cross-probability model (the probability of the clean model Gaussian, given the noisy model one, and the noisy feature vector). In [5], a time-independent solution is considered to compute this probability, but this work focuses on a different solution [6], which consists on modelling the noisy feature vectors associated to each pair of Gaussians from the clean and the noisy basic environment spaces with a GMM. Furthermore, adapting acoustic

This work has been supported by the national project TIN 2005-08660-C04-01.

models to the normalized space is proposed to reduce the mismatch between compensated feature vectors and clean acoustic models.

This paper is organized as follows: In Section 2, an overview of MEMLIN is detailed. In Section 3, some experiments are presented to show the importance of the cross-probability model estimation. The GMM based solution considered to compute the cross-probability model is explained in Section 4. The acoustic model retrained is explained in Section 5. The results with Spanish SpeechDat Car [7] and Aurora2 [8] databases are included in Section 6. Finally, the conclusions are presented in Section 7.

## 2. MEMLIN OVERVIEW

MEMLIN is an empirical feature vector normalization technique which uses stereo data in order to estimate the different compensation linear transformations in a previous training process. The clean feature space is modelled as a mixture of Gaussians. The noisy space is split into several basic acoustic environments and each one is modelled as a mixture of Gaussians. The linear transformations are estimated for all basic environments between a clean Gaussian and a noisy Gaussian. A scheme of MEMLIN can be shown in Fig. 1.

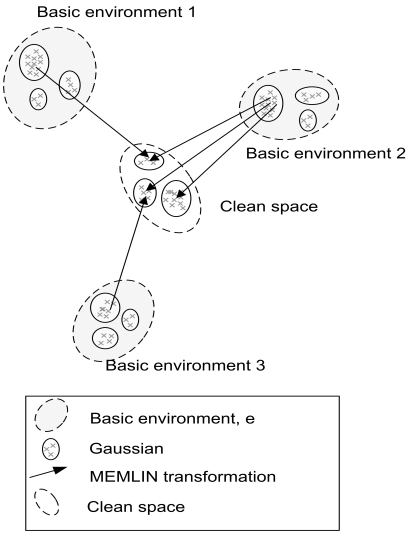


**Fig. 1**. Scheme of MEMLIN.

### 2.1. MEMLIN approximations

• Clean feature vectors, $\mathbf{x}_t$, are modelled using a GMM of $C$ components

$$p(\mathbf{x}_t) = \sum_{s_x=1}^{C} p(\mathbf{x}_t|s_x)p(s_x), \tag{1}$$

$$p(\mathbf{x}_t|s_x) = \mathcal{N}(\mathbf{x}_t; \mu_{s_x}, \mathbf{\Sigma}_{s_x}), \tag{2}$$

where $t$ is the time index and $\mu_{s_x}$, $\mathbf{\Sigma}_{s_x}$, and $p(s_x)$ are the mean vector, the diagonal covariance matrix, and the a priori probability associated with the clean model Gaussian $s_x$.

• Noisy space is split into several basic environments, $e$, and the noisy feature vectors, $\mathbf{y}_t$, are modeled as a GMM of $C'$ components for each basic environment (assuming that all the the basic environments are modelled with the same number of components)

$$p_e(\mathbf{y}_t) = \sum_{s_y^e=1}^{C'} p(\mathbf{y}_t|s_y^e)p(s_y^e), \tag{3}$$

$$p(\mathbf{y}_t|s_y^e) = \mathcal{N}(\mathbf{y}_t; \mu_{s_y^e}, \mathbf{\Sigma}_{s_y^e}), \tag{4}$$

where $s_y^e$ denotes the corresponding Gaussian of the noisy model for the $e$ basic environment; $\mu_{s_y^e}$, $\mathbf{\Sigma}_{s_y^e}$, and $p(s_y^e)$ are the mean vector, the diagonal covariance matrix, and the a priori probability associated with $s_y^e$.

• Clean feature vectors can be approximated as a linear function, $\Psi$, of the noisy feature vector which depends on the basic environments, and the clean and noisy model Gaussians: $\mathbf{x} \approx \Psi(\mathbf{y}_t, s_x, s_y^e) = \mathbf{y}_t - \mathbf{r}_{s_x,s_y^e}$, where $\mathbf{r}_{s_x,s_y^e}$ is the bias vector transformation between noisy and clean feature vectors for each pair of Gaussians, $s_x$ and $s_y^e$.

### 2.2. MEMLIN enhancement

With those approximations, MEMLIN transforms the MMSE estimation expression, $\hat{\mathbf{x}}_t = E[\mathbf{x}|\mathbf{y}_t]$, into

$$\hat{\mathbf{x}}_t = \mathbf{y}_t - \sum_{e}\sum_{s_y^e}\sum_{s_x} \mathbf{r}_{s_x,s_y^e} p(e|\mathbf{y}_t)p(s_x|\mathbf{y}_t, e, s_y^e), \tag{5}$$

where $p(e|\mathbf{y}_t)$ is the a posteriori probability of the basic environment; $p(s_y^e|\mathbf{y}_t, e)$ is the a posteriori probability of the noisy model Gaussian, $s_y^e$, given the feature vector and the basic environment. To estimate those terms ($p(e|\mathbf{y}_t)$ and $p(s_y^e|\mathbf{y}_t, e)$), expressions (3) and (4) are applied as described in [5]. Finally, the cross-probability model, $p(s_x|\mathbf{y}_t, e, s_y^e)$, is the probability of the clean model Gaussian, $s_x$, given the noisy feature vector, the basic environment and the noisy model Gaussian. The cross-probability model can be estimated avoiding the time dependence given by the noisy feature vector in a training phase using stereo data for each basic environment $(\mathbf{X}_e^{Tr}, \mathbf{Y}_e^{Tr}) = (\mathbf{x}_1^{Tr,e}, \mathbf{y}_1^{Tr,e}); ...; (\mathbf{x}_{t_e}^{Tr,e}, \mathbf{y}_{t_e}^{Tr,e}); ...; (\mathbf{x}_{T_e}^{Tr,e}, \mathbf{y}_{T_e}^{Tr,e})$, with $t_e \in [1, T_e]$ [5] as

$$p(s_x|\mathbf{y}_t, e, s_y^e) \simeq p(s_x|s_y^e) = \frac{\sum_{t_e} p(\mathbf{x}_{t_e}^{Tr,e}|s_x)p(\mathbf{y}_{t_e}^{Tr,e}|s_y^e)p(s_x)p(s_y^e)}{\sum_{t_e}\sum_{s_x} p(\mathbf{x}_{t_e}^{Tr,e}|s_x)p(\mathbf{y}_{t_e}^{Tr,e}|s_y^e)p(s_x)p(s_y^e)}. \tag{6}$$

On the other hand, the bias vector transformation, $\mathbf{r}_{s_x,s_y^e}$, is also computed using the stereo data in the previous training phase [5].

## 3. CROSS-PROBABILITY MODEL PERFORMANCE

To study the performance of the cross-probability model in a qualitative way, the histograms and log-scattegrams between the first Mel Frequency Cepstral Coefficients (MFCCs) in non-silence frames for different signals are depicted in Fig. 2.

Figure 2.a, which represents the clean and noisy feature coefficients in real car conditions, shows the effects of car noise. The pdf of clean first MFCCs is clearly affected (Fig.2.a.1), and the uncertainty is increased (Fig.2.a.2).

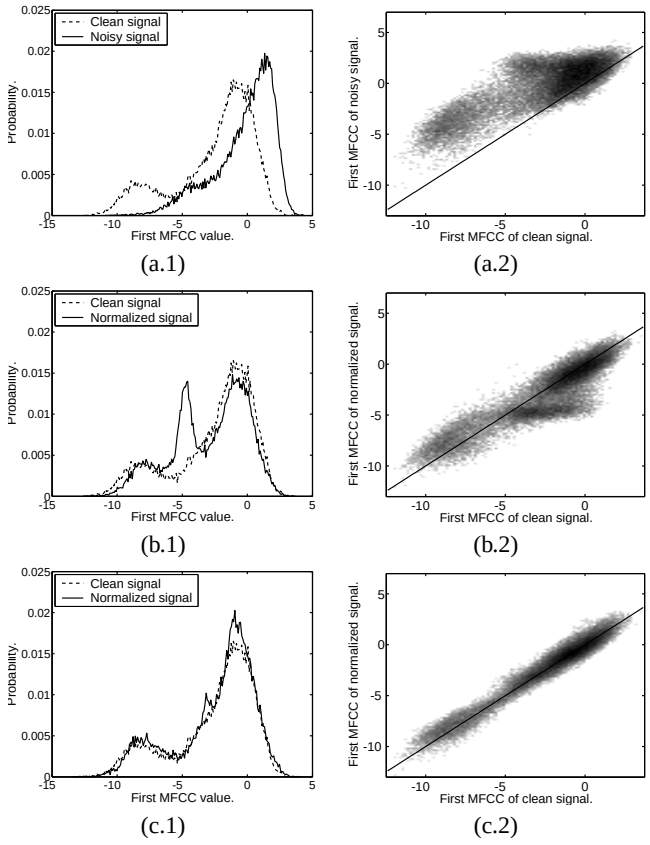


**Fig. 2**. Log-scattegrams and histograms between the first MFCC in non-silence frames for different signals. The line in the log-scattergrams represents the function $x = y$.

In Fig. 2.b and 2.c, clean and normalized coefficients with MEMLIN (128 Gaussians are considered to model the clean and basic environment spaces) are represented. The pdf of normalized first MFCCs has been approximated to the clean signal one (Fig. 2.b.1), and the uncertainty has been reduced (Fig. 2.b.2). The peak that appears in Fig. 2.b.1 is due to the transformation of noisy feature vectors towards the clean silence.

Finally, Fig. 2.c represents clean and normalized coefficients with MEMLIN when the cross-probability model is computed with the corresponding clean feature vector as (7). 128 Gaussians are used to model the different spaces. In this case the pdf of the normalized signal is almost the same that the clean one (Fig. 2.c.1) and the uncertainty is drastically reduced (Fig. 2.c.2). Furthermore, the WER results in this case are almost the same that we would obtain with clean signal. These results verify the importance of a good estimation of the cross-probability model in MEMLIN algorithm.

$$p(s_x|\mathbf{y}_t, e, s_y^e) \simeq \frac{p(s_x)p(\mathbf{x}_t|s_x)}{\sum_{s_x} p(s_x)p(\mathbf{x}_t|s_x)}. \quad (7)$$

## 4. CROSS-PROBABILITY MODEL BASED ON GMM

To improve the time-independent cross-probability model (6), we propose to model the noisy feature vectors associated to a pair of Gaussians ($s_x$ and $s_y^e$) with a GMM of $C''$ components (assuming that the noisy feature vectors are modelled with the same number of Gaussianas for all pairs $s_x$ and $s_y^e$). Since the estimation of the corresponding GMMs for each basic environment can be considered independent, they are not indexed to simplify the notation. Hence we present a model of the noisy feature vectors associated to the pair of Gaussians $s_x$ and $s_y$

$$p(\mathbf{y}_t|s_x, s_y) = \sum_{s'_y=1}^{C''} p(\mathbf{y}_t|s_x, s_y, s'_y)p(s'_y|s_x, s_y), \quad (8)$$

$$p(\mathbf{y}_t|s_x, s_y, s'_y) = \mathcal{N}(\mathbf{y}_t; \mu_{s_x,s_y,s'_y}, \mathbf{\Sigma}_{s_x,s_y,s'_y}), \quad (9)$$

where $\mu_{s_x,s_y,s'_y}$, $\mathbf{\Sigma}_{s_x,s_y,s'_y}$, and $p(s'_y|s_x, s_y)$ are the mean vector, the diagonal covariance matrix, and the a priori probability associated with $s'_y$ Gaussian of the cross-probability GMM associated with $s_x$ and $s_y$. To train these three parameters, the EM algorithm [9] is applied.

Let a set of clean and noisy stereo data available to learn the corresponding cross-probability GMM parameters $(\mathbf{X}, \mathbf{Y}) = \{(\mathbf{x}_1, \mathbf{y}_1), ...(\mathbf{x}_n, \mathbf{y}_n)..., (\mathbf{x}_N, \mathbf{y}_N)\}$. Each $\mathbf{y}_n$ can be seen as an incomplete component-labelled frame, which is completed by two indicator vectors. The first one is $\mathbf{w}_n \in \{0, 1\}^{C'}$, with 1 in the position corresponding to the $s_y$ Gaussian which generates $\mathbf{y}_n$ and zeros elsewhere ($\mathbf{W} = \{\mathbf{w}_1, ..., \mathbf{w}_N\}$). The second indicator vector is $\mathbf{z_n} \in \{0, 1\}^{C''}$, with 1 in the position corresponding to the $s'_y$ Gaussian of the cross-probability GMM which generates $\mathbf{y}_n$ and zeros elsewhere ($\mathbf{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_N\}$). Each $\mathbf{x}_n$ can be seen also as an incomplete component-labelled frame, which is completed by one indicator vector: $\mathbf{v}_n \in \{0, 1\}^C$, with 1 in the position corresponding to the $s_x$ Gaussian which generates $\mathbf{x}_n$ and zeros elsewhere ($\mathbf{V} = \{\mathbf{v}_1, ..., \mathbf{v}_N\}$). The indicator vectors are called missing data, too. So, the complete data pdf is

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{z}) \quad &\simeq \quad p(\mathbf{v}, \mathbf{w})p(\mathbf{x}|\mathbf{v}, \mathbf{w}) \times \\ &\quad p(\mathbf{v}, \mathbf{w}, \mathbf{z})p(\mathbf{y}|\mathbf{v}, \mathbf{w}, \mathbf{z}), \end{aligned} \quad (10)$$

where it is assumed that $\mathbf{x}$ is independent of $\mathbf{y}$ and $\mathbf{z}$. Since the missing data are Multinomial, the complete data pdf can be expressed as (11), where $v_{s_x}$, $w_{s_y}$ and $z_{s'_y}$ are the components of $\mathbf{v}$, $\mathbf{x}$ and $\mathbf{z}$ associated to the Gaussians $s_x$, $s_y$ and $s'_y$, respectively.

Once the complete data pdf is obtained, the EM algorithm is applied iteratively in two steps: the Expectation (E) step, which estimates the expected values of the missing data, and the Maximization (M) step, which obtains the parameters of the cross-probability GMM using the estimated missing data.

### 4.1. The E step

To evaluate the E step, the function $Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})$ is defined as $Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)}) = E[log(p(\mathbf{X}, \mathbf{Y}, \mathbf{V}, \mathbf{W}, \mathbf{Z}|\mathbf{\Theta}))|\mathbf{X}, \mathbf{Y}, \mathbf{\Theta}^{(k)}]$, where the operator $E[\bullet]$ is the expected value, $k$ is the iteration index and $\mathbf{\Theta}$ includes all the unknown parameters of the cross-probability GMM we pretend to estimate. So, $Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})$ is expressed as (12), where

$$(v_{s_x} w_{s_y})^{(k)} \simeq E[v_{s_x}|\mathbf{x}_n]E[w_{s_y}|\mathbf{y}_n], \quad (13)$$

$$(v_{s_x} w_{s_y} z_{s'_y})^{(k)} \simeq (v_{s_x} w_{s_y})^{(k)} E[z_{s'_y}|\mathbf{y}_n, v_{s_x}, w_{s_y}, \mathbf{\Theta}^{(k)}], \quad (14)$$

$$p(\mathbf{x},\mathbf{y},\mathbf{v},\mathbf{w},\mathbf{z}) \simeq \prod_{s_x}\prod_{s_y}[p(v_{s_x}=1,w_{s_y}=1)p(\mathbf{x}|v_{s_x}=1,w_{s_y}=1)]^{v_{s_x}w_{s_y}} \times \prod_{s_x}\prod_{s_y}\prod_{s'_y}[p(v_{s_x}=1,w_{s_y}=1,z_{s'_y}=1)p(\mathbf{y}|v_{s_x}=1,w_{s_y}=1,z_{s'_y}=1)]^{v_{s_x}w_{s_y}z_{s'_y}}. \quad (11)$$

$$Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)}) = \sum_n\sum_{s_x}\sum_{s_y}(v_{s_x}w_{s_y})^{(k)}[log(p(s_x)p(s_y)) + log(p(\mathbf{x}_n|v_{s_x}=1,w_{s_y}=1))] + \sum_n\sum_{s_x}\sum_{s_y}\sum_{s'_y}(v_{s_x}w_{s_y}z_{s'_y})^{(k)}[log(p(s_x)p(s_y)p(s'_y|s_x,s_y)) + log(p(y_n|v_{s_x}=1,w_{s_y}=1,z_{s'_y}=1))]. \quad (12)$$

$$E[z_{s'_y}|\mathbf{y}_n,v_{s_x},w_{s_y},\mathbf{\Theta}^{(k)}] = \frac{p(s'_y|s_x,s_y)^{(k)}N(\mathbf{y}_n|\mu^{(k)}_{s_x,s_y,s'_y},\mathbf{\Sigma}^{(k)}_{s_x,s_y,s'_y})}{\sum_{s'_y}p(s'_y|s_x,s_y)^{(k)}N(\mathbf{y}_n|\mu^{(k)}_{s_x,s_y,s'_y},\mathbf{\Sigma}^{(k)}_{s_x,s_y,s'_y})}. \quad (15)$$

where it is assumed that $v_{s_x}$ and $w_{s_y}$ are independent, $E[v_{s_x}|\mathbf{x}_n,\mathbf{y}_n,\mathbf{\Theta}^{(k)}] \simeq E[v_{s_x}|\mathbf{x}_n]$ and $E[w_{s_y}|\mathbf{x}_n,\mathbf{y}_n,\mathbf{\Theta}^{(k)}] \simeq E[w_{s_y}|\mathbf{y}_n]$. $E[z_{s'_y}|\mathbf{y}_n,v_{s_x},w_{s_y},\mathbf{\Theta}^{(k)}]$ is estimated with (8) and (9) as (15), and $E[v_{s_x}|\mathbf{x}_n]$ and $E[w_{s_y}|\mathbf{y}_n]$ are computed in a similar way with (1) and (2), and with (3) and (4), respectively, assuming that there is only one basic environment. Although, in this work, to simplify, $E[v_{s_x}|\mathbf{x}_n]$ and $E[w_{s_y}|\mathbf{y}_n]$ values are 1, if the corresponding Gaussians are the most probable ones, and 0 in any other case (hard Gaussian estimation approach).

### 4.2. The M step

To obtain the maximum likelihood estimates for the unknown parameters of the cross-probability GMM, $Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})$ is maximized with respect to them. So, the corresponding expressions for the $(k+1)$th iteration are

$$p(s'_y|s_x,s_y)^{(k+1)} = \frac{\sum_n(v_{s_x}w_{s_y}z_{s'_y})^{(k)}}{\sum_n\sum_{s'_y}(v_{s_x}w_{s_y}z_{s'_y})^{(k)}}. \quad (16)$$

$$\mu^{(k+1)}_{s_x,s_y,s'_y} = \frac{\sum_n(v_{s_x}w_{s_y}z_{s'_y})^{(k)}\mathbf{y}_n}{\sum_n(v_{s_x}w_{s_y}z_{s'_y})^{(k)}}. \quad (17)$$

$$\mathbf{\Sigma}^{(k+1)}_{s_x,s_y,s'_y} = \frac{1}{\sum_n(v_{s_x}w_{s_y}z_{s'_y})^{(k)}} \times \sum_n(v_{s_x}w_{s_y}z_{s'_y})^{(k)}(\mathbf{y}_n-\mu^{(k)}_{s_x,s_y,s'_y})(\mathbf{y}_n-\mu^{(k)}_{s_x,s_y,s'_y})^t. \quad (18)$$

As it has been indicated, for MEMLIN, the cross-probability GMM parameters have to be estimated independently for each basic environment using the labeled training corpus $(\mathbf{X}_{Tr,e},\mathbf{Y}_{Tr,e})$. So, the expressions (8) and (9) are transformed into

$$p(\mathbf{y}_t|s_x,s^e_y,e) = \sum_{s'_y=1}^{C''}p(\mathbf{y}_t|s_x,s^e_y,s'_y,e)p(s'_y|s_x,s^e_y,e), \quad (19)$$

$$p(\mathbf{y}_t|s_x,s^e_y,s'_y,e) = \mathcal{N}(\mathbf{y}_t;\mu_{s_x,s^e_y,s'_y},\mathbf{\Sigma}_{s_x,s^e_y,s'_y}), \quad (20)$$

where $\mu_{s_x,s^e_y,s'_y}$, $\mathbf{\Sigma}_{s_x,s^e_y,s'_y}$, and $p(s'_y|s_x,s^e_y,e)$ are the mean vector, the diagonal covariance matrix, and the a priori probability associated with $s'_y$ Gaussian of the cross-probability GMM associated with $s_x$ and $s^e_y$. So, $p(s_x|\mathbf{y}_t,e,s^e_y)$ can be obtained as

$$p(s_x|\mathbf{y}_t,e,s^{e,ph}_y) = \frac{p(\mathbf{y}_t|s_x,s^e_y,e)}{\sum_{s_x}p(\mathbf{y}_t|s_x,s^e_y,e)}. \quad (21)$$

## 5. NORMALIZED SPACE ACOUSTIC MODELS

Feature vector normalization techniques try to map the noisy feature vectors to the clean space. However this mapping is not perfect and a new normalized space is created, which is different from the clean one. Thus, a further improvement can be obtained adapting the clean acoustic models towards the normalized space. For this purpose, the noisy training data are normalized in the same way as testing data and the original clean acoustic models are adapted with those data towards the new normalized space. If there are enough data, Maximum Likelihood (ML) algorithm can be used, but a model adaptation method should be applied otherwise (Maximum A Posteriori, MAP [10], MLLR [11]...). In this work, once the MEMLIN normalized space acoustic models are obtained, the normalized testing data can be recognized directly with them.

## 6. RESULTS

### 6.1. Results with SpeechDat Car database

To observe the performance of the cross-probability GMM proposed in a real, dynamic, and complex environment, a set of experiments were carried out using the Spanish SpeechDat Car database [7]. Seven basic environments were defined: car stopped, motor running (E1), town traffic, windows close and climatizer off (silent conditions) (E2), town traffic and noisy conditions: windows open and/or climatizer on (E3), low speed, rough road, and silent conditions (E4), low speed, rough road, and noisy conditions (E5), high speed, good road, and silent conditions (E6), and high speed, good road, and noisy conditions (E7).

The clean signals are recorded with a CLose talK (CLK) microphone (Shure SM-10A), and the noisy ones are recorded by a Hands-Free (HF) microphone placed on the ceiling in front of the driver (Peiker ME15/V520-1). The SNR range for CLK signals goes from 20 to 30 dB, and for HF ones goes from 5 to 20 dB.

For speech recognition, the feature vectors are composed of the 12 MFCCs, the energy, first and second derivatives, giving a final feature vector of 39 coefficients computed every 10 ms using a 25 ms Hamming window. On the other hand, in this work, the feature vector normalization methods are applied only to the 12 MFCCs and energy, whereas the derivatives are computed over the normalized static coefficients

The recognition task is isolated and continuous digits recognition. The acoustic models are composed by 16-state 3 Gaussian continuous density HMM to model the 10 Spanish digits and 2 silence models for long (three-state 6 Gaussian continuous density HMM) and interword (one-state 6 Gaussian continuous density HMM) silences are used.

| Train | Test | E1 | E2 | E3 | E4 | E5 | E6 | E7 | MWER (%) |
|---|---|---|---|---|---|---|---|---|---|
| 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 | 21.49 |
| 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 |

**Table 1**. WER baseline results, in %, from the different basic environments (E1,..., E7).

The Word Error Rate (WER) baseline results for each basic environment are presented in Table 1, where MWER is the Mean WER computed proportionally to the number of words in each basic environment. Cepstral mean normalization is applied to testing and training data. "Train" column refers to the signals used to obtain the corresponding acoustic HMMs: CLK if they are trained with all clean training utterances, and HF and if they are trained with all noisy ones. HF† indicates that specific acoustic HMMs for each basic environment are applied in the recognition task (environment match condition). "Test" column indicates which signals are used for recognition: clean, CLK, or noisy, HF.

Table 1 shows the effect of real car conditions, which increases the WER in all of the basic environments, ("Train" CLK, "Test" HF), concerning the rates for clean conditions, ("Train" CLK, "Test" CLK). When acoustic models are retrained using all basic environment signals, ("Train" HF) MWER decreases. Finally, and in spite of the high WER reached for the basic environment E7 due to the reduced number of training utterances, 3.42% of MWER is obtained for environment match condition.

Figure 3 shows the mean improvement in WER (MIMP) in % for MEMLIN and MEMLIN with Cross-Probability model based on GMM (MEMLIN CPGMM). Also the results with SPLICE with Environmental Model Selection (SPLICE EMS) [4] are included. MIMP is computed as

$$MIMP = \frac{100(MWER - MWER_{CLK-HF})}{MWER_{CLK-CLK} - MWER_{CLK-HF}}, \quad (22)$$

where $MWER_{CLK-CLK}$ is the mean WER obtained with clean conditions (0.91 in this case), and $MWER_{CLK-HF}$ is the baseline (21.49). So, A 100% MIMP would be achieved when MWER equals the one obtained under clean conditions. The cross-probability GMMs are composed by 2 Gaussians for each pair of clean and noisy Gaussians. It can be observed the important improvement of MEMLIN CPGMM concerning MEMLIN: from 62.57% to 75.79% with 4 Gaussians per basic environment and from 74.08% to 82.86% with 64 Gaussians.

Although the number of Gaussians to model the basic environments could be the same for MEMLIN and MEMLIN CPGMM, the computing time is not the same. To reduce it, only the cross-probability GMMs of the most probable pairs of Gaussians could be computed in normalization. Some experiments were carried out considering this alternative, showing that similar results can be obtained computing only a reduced number of pair of Gaussians [6].

Table 2 shows the corresponing matching condition results (MWER and MIMP) when normalized acoustic models are used (clean and noisy condition results, Train CLK, Test CLK and Train HF, Test HF, can be observed in Table 1 to compare). In Train HF MEMLIN and Train HF MEMLIN CPGMM, the noisy training data normalized with MEMLIN or MEMLIN CPGMM are used to retrain the corresponding new acoustic models with the ML algorithm.

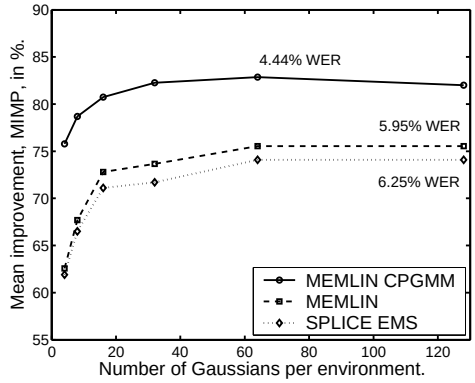


**Fig. 3**. Mean improvement in WER, MIMP, in % for MEMLIN, MEMLIN with Cross-Probability model based on GMM, MEMLIN CPGMM, and SPLICE with Environmental Model Selection, SPLICE EMS.

The number of Gaussians per basic environment is included next to the normalization techniques and for MEMLIN CPGMM, the noisy feature vectors for each pair of Gaussianas $s_x$ and $s_y^e$ are modelled with 2 components (there is not significant differences in recognition if the basic environments are modelled with different number of Gaussians). Clearly there are significant improvements when normalized space acoustic models are used. It can be observed that the improvement with respect to using clean acoustic models is significant (4.44% and 5.95% of MWER for MEMLIN CPGMM and MEMLIN, respectively), and the comparison is even satisfactory if we compare the results with the ones reached with environment match condition ("Train" HF, "Test" HF and "Train" HF †, "Test" HF). This is because the normalized space is not as heterogenous as the noisy one and the training process can be more effective.

### 6.2. Results with Aurora2 database

Aurora2 database [8] is built from TIDigits database utterances that have been digitally corrupted by passing them through a linear filter and/or by adding different types of noises at SNRs ranging from 20dB to -5dB. This does not define a real environment because not all kind of degradations are included i.e. Lombard effect [12]; but, in spite of this weakness, Aurora2 is one of the most used database and it is almost a standard database to compare different techniques.

In this work, the MEMLIN and MEMLIN CPGMM parameters were trained using identical utterances from the clean training set and the multi-condition training set. This tunes the normalization parameters on the noise types from set A, keeping sets B and C as unseen conditions. Although the results for the three sets were obtained, in this work we only present the results with car noise contaminated signals, which is considered as testing corpus and it

| Train | Test | MWER (%) | MIMP (%) |
|---|---|---|---|
| HF MEMLIN 64 | HF MEMLIN 64 | 1.67 | 96.33 |
| HF MEMLIN CPGMM 128 | HF MEMLIN CPGMM 128 | 1.47 | 97.27 |

**Table 2**. Best MWER and MIMP obtained with MEMLIN and MEMLIN CPGMM and matched acoustic models.

| Train | Test | -5dB | 0dB | 5dB | 10dB | 15dB | 20dB | clean | MWER (%) | MIMP (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| CLK | HF | 6.83 | 10.71 | 30.75 | 63.53 | 88.55 | 97.08 | 99.05 | 58.12 | – |
| CLK | HF MEMLIN 64 | 24.58 | 50.76 | 78.68 | 92.53 | 97.26 | 98.33 | 99.25 | 83.51 | 62.49 |
| CLK | HF MEMLIN CPGMM 64 | 26.67 | 55.53 | 82.98 | 94.40 | 97.53 | 98.51 | 99.25 | 85.79 | 67.52 |

**Table 3**. Best results obtained with MEMLIN and MEMLIN CPGMM with car noise contaminated signals of Aurora2 database.

is marked as HF to maintain the nomenclature. The parameters for speech recognition (acoustic models and feature vectors) are obtained as the same way as it is indicated in Subsection 6.2.

The recognition results obtained with Aurora2 database are presented in Table 3. It can be observed that MEMLIN and MEMLIN CPGMM maintain the satisfactory performance, obtaining a mean improvement of 62.49% and 67.52%, respectively (the improvement is computed in this case as ETSI recommendation).

## 7. CONCLUSIONS

In this paper we have focussed on an approach of MEMLIN where the cross-probability model is estimated by modelling the noisy feature vectors associated to each pair of Gaussians from the clean and the noisy basic environment spaces with a GMM. MEMLIN obtains an improvement in WER of 75.53% with 128 Gaussians per environment with SpeechDat Car database in Spanish, whereas MEMLIN with cross-probability model based on GMM reaches 82.86% for 64 Gaussians to model each basic environment. If we consider Aurora2 database, and the recognition test is composed only by the car noise corrupted signals, the improvements are, modelling each basic environments with 64 Gaussians, 62.49% and 67.52%, respectively. On the other hand, in order to reduce the mismatch between normalized feature vectors and clean acoustic models, we propose to obtain acoustic models which represent the normalized space. Applying this procedure to SpeechDat Car database, important improvements are obtained: 96.33% and 97.27% if the normalization technique is MEMLIN or MEMLIN CPGMM with 64 and 128 Gaussianas per basic environment, respectively.

## 8. REFERENCES


[1] A. Sankar and C. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 190–202, May 1996.

[2] Leonardo Neumeyer and Mitchel Weintraub, "Robust Speech Recognition in Noise Using Adaptation and Mapping Techniques," in *Proceedings of ICASSP* . Detroit, USA, May 1995, vol. 1, pp. 141–144.

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

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

[5] L. Buera, E. Lleida, A. Miguel, and A. Ortega, "Multi-environment models based linear normalization for robust speech recognition in car conditions," in *Proceedings of ICASSP*. Montreal, Canada, May 2004, vol. 1, pp. 1013–1016.

[6] L. Buera, E. Lleida, A. Miguel, and A. Ortega, "Time-dependent cross-probability model for multi-environment model based linear normalization," in *Proceedings of ICSLP*. Pittsburgh, USA, 2006.

[7] Asuncion Moreno, Borge Lindberg, Christoph Draxler, Gael Richard, Khalid Choukri, Stephan Euler, and Jeffrey Allen, "Speechdat-car. a large speech database for automotive environments," in *Proceedings of LREC*. Athens, Greece, 2000, vol. 2, pp. 895–900.

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

[9] A. P. Dempster, N.P. Laird, and D.B. Rubin, "Maximum likelihood from imcomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 9, no. 1, pp. 1–37, 1977.

[10] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," .

[11] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," .

[12] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.