TIME-DEPENDENT CROSS-PROBABILITY MODEL FOR FEATURE VECTOR NORMALIZATION

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

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

ABSTRACT

In previous works, Multi-Environment Model based LInear Normalization, MEMLIN, and Phoneme-Dependent MEMLIN, PD-MEMLIN, were presented and they were proved to be effective to compensate environment mismatch. Both are empirical feature vector normalization techniques which model 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 both algorithms. In the previous works the cross-model probability was approximated as time-independent. However, in this paper, a time-dependent estimation based on GMM is proposed for MEMLIN and PD-MEMLIN. Some experiments with SpeechDat Car database were carried out in order to study the performance of the proposed estimation of the cross-probability model in a real acoustic environment, obtaining important improvements: 78.48% and 76.76% of mean improvement in Word Error Rate, WER, for MEMLIN and PD-MEMLIN, respectively (70.21% and 75.44% if time-independent cross-probability model is applied).

1. INTRODUCCIÓN

When training and testing acoustic conditions differ, the accuracy of speech recognition 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, some of the techniques can be combined to generate hybrid solutions, which are effective under certain conditions [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, 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], Multi-Environment Model-based LInear Normalization (MEMLIN) [5], or Phoneme-Dependent MEMLIN [6], use the MMSE estimator to compute the estimated clean feature vector.

The previous works [5] [6] show that MEMLIN and PD-MEMLIN are 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. On the other hand, the main difference of PD-MEMLIN concerning MEMLIN consists on splitting the clean and noisy basic environments into phonemes that are modelled using GMMs.

A critical point in MEMLIN and PD-MEMLIN is the estimation of the cross-probability model: the probability of the clean model Gaussian, given the noisy model one, the noisy feature vector and the phoneme (only in PD-MEMLIN). In [5] [6], a time-independent solution is considered. This work focuses on this term and it is proposed a time-dependent solution, modelling the noisy feature vectors associated to each pair of Gaussians from the clean and the noisy basic environment spaces (and the phoneme in PD-MEMLIN) with a GMM.

This paper is organized as follows: In Section 2, an overview of PD-MEMLIN is detailed. In Section 3, some experiments are presented to show the importance of the cross-probability model estimation. The new proposed cross-probability model based on GMM is explained in Section 4. The results with Spanish SpeechDat Car database [7] are included in Section 5. Finally, the conclusions are presented in Section 6.

2. PD-MEMLIN OVERVIEW

Phoneme Dependent Multi-Environment Models based LInear Normalization 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 for each phoneme. The noisy space is split in several basic acoustic environments and each environment is modelled as a mixture of Gaussians for each phoneme. The transformations are estimated for all basic environments between a clean phoneme Gaussian and a noisy Gaussian of the same phoneme. This can be shown in Fig. 1 for one environment, where s_x^{ph} is the clean model Gaussian associated to the phoneme ph, $s_y^{e,ph}$ is the noisy model Gaussian associated to the basic environment

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

e and the phoneme ph, and finally, $r_{s_x^{ph},s_y^{e,ph}}$, which is the bias vector transformation between s_x^{ph} and $s_y^{e,ph}$.

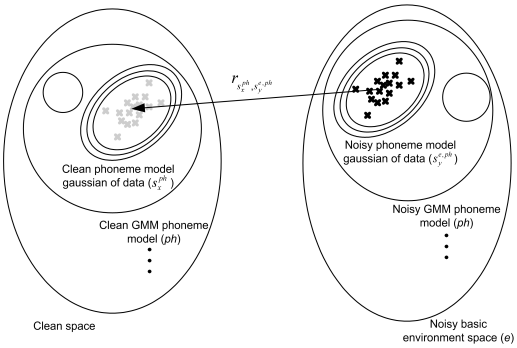

Figure 1. Scheme of PD-MEMLIN transformations for one environment.

2.1. PD-MEMLIN approximations

• Clean feature vectors, $\mathbf{x}_t$, are modelled using a GMM of C components for each phoneme, ph (assuming that all the phonemes are modelled with the same number of components)

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

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

where $\mu_{s_x^{ph}}$, $\mathbf{\Sigma}_{s_x^{ph}}$, and $p(s_x^{ph})$ are the mean vector, the diagonal covariance matrix, and the a priori probability associated with the clean model Gaussian s_x^{ph} of the ph phoneme.

• 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 and phoneme (assuming that all the phonemes of the all the basic environments are modelled with the same number of components)

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

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

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

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

2.2. PD-MEMLIN enhancement

With those approximations, PD-MEMLIN transforms the MMSE estimation expression, $\hat{\mathbf{x}}_t = E[\mathbf{x}|\mathbf{y}_t]$, into (5), where $p(e|\mathbf{y}_t)$ is the a posteriori probability of the basic environment; $p(ph|\mathbf{y}_t, e)$ is the a posteriori probability of the phoneme, given the noisy feature vector and the environment; $p(s_y^{e,ph}|\mathbf{y}_t, e, ph)$ is the a posteriori probability of the noisy model Gaussian, $s_y^{e,ph}$, given the feature vector, $\mathbf{y}_t$, the basic environment, e, and the phoneme, ph. To estimate those terms: $p(e|\mathbf{y}_t)$, $p(ph|\mathbf{y}_t, e)$ and $p(s_y^{e,ph}|\mathbf{y}_t, e, ph)$, equations (3) and (4) are applied as described in [6]. Finally, the cross-probability model, $p(s_x^{ph}|\mathbf{y}_t, e, ph, s_y^{e,ph})$, is the probability of the clean model Gaussian, s_x^{ph}, given the feature vector, $\mathbf{y}_t$, the basic environment, e, the phoneme, ph, and the noisy model Gaussian, $s_y^{e,ph}$. The cross-probability model is estimated in a training phase using stereo data for each basic environment and phoneme $(\mathbf{X}_{e,ph}, \mathbf{Y}_{e,ph}) = (\mathbf{x}_1^{e,ph}, \mathbf{y}_1^{e,ph}); ...; (\mathbf{x}_{t_{e,ph}}^{e,ph}, \mathbf{y}_{t_{e,ph}}^{e,ph}); ...; (\mathbf{x}_{T_{e,ph}}^{e,ph}, \mathbf{y}_{T_{e,ph}}^{e,ph})$, with $t_{e,ph} \in [1, T_{e,ph}]$ [6]. The cross-probability model is computed avoiding the time dependence given by the noisy feature vector as (6) (time-independent cross-probability model). On the other hand, the bias vector transformation, $\mathbf{r}_{s_x^{ph},s_y^{e,ph}}$, is also computed using the stereo data in the previous training phase [6].

The expressions for MEMLIN can be obtained directly from the PD-MEMLIN ones if only one phoneme is considered [5].

3. CROSS-PROBABILITY MODEL PERFORMANCE

To study the performance of the cross-probability model in a qualitative way, the histograms and 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 clean and noisy in real car conditions feature vectors, 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).

Since we only want to observe the importance of the cross-probability model, and PD-MEMLIN performance is highly dependent of the probability of the phoneme, given the environment and the noisy feature vector, $p(ph|\mathbf{y}_t, e)$, we present results with MEMLIN in Fig. 2.b and 2.c. So, clean and normalized coefficients with MEMLIN are represented in the Fig. 2.b. MEMLIN is applied with 128 Gaussians. 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 with MEMLIN feature vectors where the cross-probability model is computed with the corresponding clean feature vector as (7). MEMLIN is also applied with 128 Gaussians. 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. Similar experiments were carried out with PD-MEMLIN, obtain a similar satisfactory performance.

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

4. CROSS-PROBABILITY MODEL BASED ON GMM

To improve the time-independent cross-probability model, we propose to model the noisy feature vectors associated to a pair of Gaussians (s_x and s_y) with a GMM of C'' components (as-

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

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

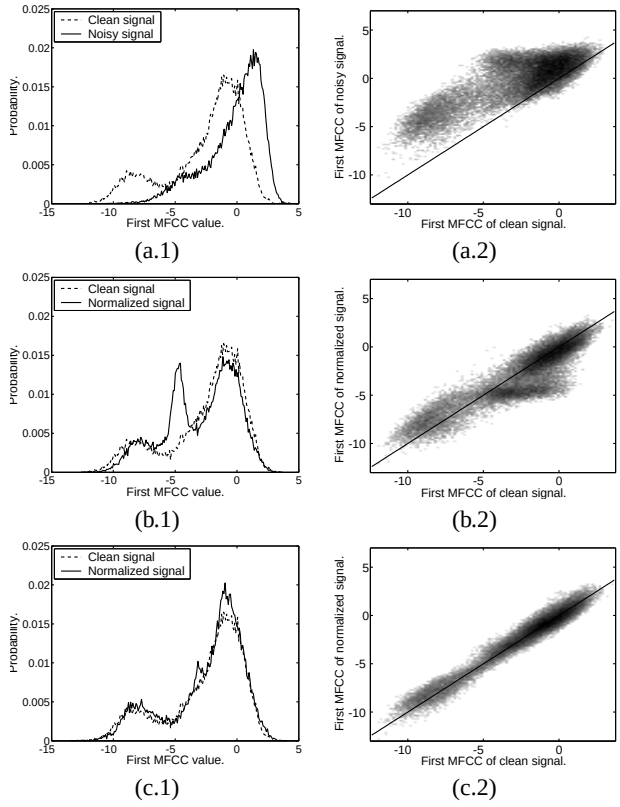

Figure 2. Scattegrams and histograms between the first MFCC in non-silence frames for different signals. The line in the scattergrams represents the function $x = y$.

suming that all the pair of Gaussians, s_x and s_y, are modelled with the same number of Gaussians

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

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

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, 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 [8] is applied. The basic environments and the phonemes are not indexed for clarity, but they can be considered independently.

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 generating $\mathbf{y}_n$ 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 generating $\mathbf{y}_n$ 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 generating $\mathbf{x}_n$ 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 indicator vectors 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.

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 $E[\bullet]$ is the expected value, k is the iteration index and $\mathbf{\Theta}$ includes the unknown parameters of the cross-probability GMM. It is expressed as (12), where

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

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

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 phoneme. 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 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

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

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

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

Train	Test	E1	E2	E3	E4	E5	E6	E7	MWER (%)
CLK	CLK	1.90	2.64	1.81	1.75	1.62	0.64	0.35	1.75
CLK	HF	5.91	14.49	14.55	20.17	21.07	16.19	35.71	16.21
HF	HF	6.67	14.24	12.73	12.91	14.97	9.68	8.50	11.81
†HF	HF	2.86	7.12	4.34	4.39	7.63	4.60	4.76	5.30

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

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

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

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

Once the cross-probability GMM parameters are estimated for each basic environment, $p(s_x|\mathbf{y}_t,e,s^e_y)$ can be obtained for MEMLIN with (8) as (19). Note that the time-independent assumption has been avoided.

$$p(s_x|\mathbf{y}_t,e,s^e_y) = \frac{p(\mathbf{y}_t|s_x,s^e_y)}{\sum_{s_x}p(\mathbf{y}_t|s_x,s^e_y)}. \quad (19)$$

For PD-MEMLIN, the cross-probability GMM parameters are estimated for each basic environment and phoneme independently, and $p(s^{ph}_x|\mathbf{y}_t,e,ph,s^{e,ph}_y)$ is computed in a similar way as (19).

5. RESULTS

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, first and second derivatives and the delta energy, giving a final feature vector of 37 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. Three-state 16 Gaussian continuous density HMM to model the 25 Spanish phonemes and 2 silence models for long and interword silences are used in this task.

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, 5.30% of MWER is obtained for environment match condition.

Figure 3 shows the mean improvement in WER (MIMP) in % for MEMLIN with Time-Independent cross-probability model (MEMLIN TI) and with Time-Dependent cross-probability GMM (MEMLIN TD). Also the results with SPLICE with Environmental Model Selection (SPLICE EMS) [4] are included. A 100% MIMP would be achieved when MWER equals the same of clean conditions. The cross-probability GMMs are composed by 2 Gaussians. It can be observed the important improvement of MEMLIN TD concern-

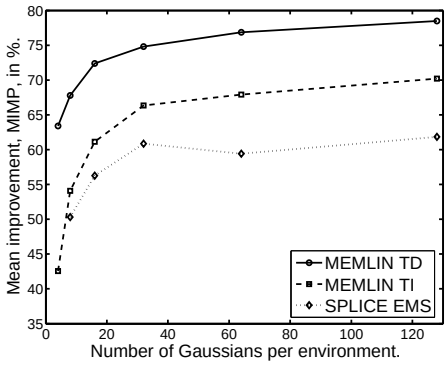

Figure 3. Mean improvement in WER, MIMP, in % for MEMLIN with Time-Dependent cross-probability model, MEMLIN TD, MEMLIN with Time-Dependent GMM cross-probability model, MEMLIN TI, and SPLICE with Environmental Model Selection, SPLICE EMS.

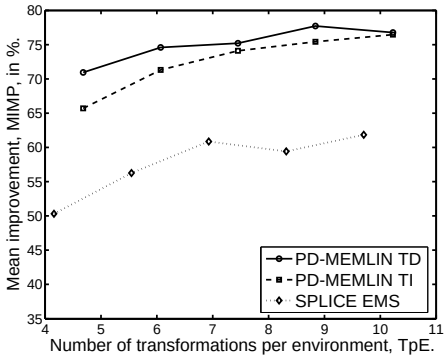

Figure 4. Mean improvement in WER, MIMP, in % for PD-MEMLIN with Time-Dependent cross-probability model, PD-MEMLIN TD, PD-MEMLIN with Time-Dependent GMM cross-probability model, PD-MEMLIN TI, and SPLICE with Environmental Model Selection, SPLICE EMS.

ing MEMLIN TI: from 42.55% to 63.38% with 4 Gaussians per basic environment and from 70.58% to 78.47% with 128 Gaussians.

On the other hand, the mean improvement in WER for PD-MEMLIN with Time-Independent cross-probability model (PD-MEMLIN TI) and with Time-Dependent cross-probability GMM (PD-MEMLIN TD) are depicted in Figure 4. The results obtained with SPLICE with Environmental Model Selection (SPLICE EMS) are also included to compare. The cross-probability GMMs are composed by 2 Gaussians for each pair of clean and noisy Gaussians of the same phoneme. To make a fair comparison between the methods, the results have been plotted as a function of the number of Transformations per basic Environment, TpE, which each method has to compute for each frame in normalization, in log_{10}

$$TpE = log_{10}(n_{s_y^{ph}} n_{s_x^{ph}} n_{ph}), \quad (20)$$

where $n_{s_y^{ph}}$ and $n_{s_x^{ph}}$ are the number of noisy and clean model Gaussians for ph phoneme, respectively, and n_{ph} is the number of phonemes ($n_{ph} = 1$, for SPLICE EMS).

It can be observed a slight improvement of PD-MEMLIN TD concerning PD-MEMLIN TI when TpE is higher: from 65.71% to 70.94% with 2 Gaussians per phoneme, and a bigger improvement when the phonemes are modelled with few Gaussians (lower TeP): from 75.43% to 77.72% with 16 Gaussians. Note that the critical point in PD-MEMLIN is not only the cross-probability model but also the probability of the phoneme, given the noisy feature vector and the environment, $p(ph|\mathbf{y}_t, e)$. So, there is still one term that needs to be improved in PD-MEMLIN.

Although the number of Gaussians to model the basic environments could be the same for MEMLIN TI and MEMLIN TD or for PD-MEMLIN TI and PD-MEMLIN TD, the computing time is not the same. To reduce it, only the cross-probability GMMs of the most probable pairs of Gaussians can be computed in normalization. In this case, for each noisy feature vector, the most probable PHonemes and Noisy model Gaussians (♯PH and ♯NG) can be obtained with (3) and (4), and for each one, the corresponding most probable Clean model Gaussians (♯CG) are obtained with (6) (No phoneme dependence for MEMLIN).

Table 2 shows the results for MEMLIN TD for different ♯NG and ♯CG. In all cases the cross-probability GMMs are composed by 2 Gaussians.

	♯NG	♯CG	MWER	MIMP
MEMLIN TD 4-4	4	4	7.04	63.40
MEMLIN TD 8-8	4	4	6.87	64.40
MEMLIN TD 16-16	8	8	5.67	72.87
MEMLIN TD 32-32	8	8	5.62	73.23
MEMLIN TD 64-64	16	16	5.44	74.46
MEMLIN TD 128-128	32	32	5.11	76.77

Table 2. Mean WER (MWER) and mean improvement in WER (MIMP) in % for MEMLIN TD when different Gaussians of cross-probability GMM are computed.

Table 3 shows the results for PD-MEMLIN TD for different ♯PH, ♯NG and ♯CG. In all cases the cross-probability GMMs are composed by 2 Gaussians. It can be observed that in MEMLIN TD and PD-MEMLIN TD it is not necessary to compute the cross-probability for all the Gaussians to obtain satisfactory results due to all of them has not the same importance.

6. CONCLUSIONS

In this paper we have presented an approach of MEMLIN and PD-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 70.21% with 128 Gaussians per environment, whereas MEMLIN with cross-probability GMM reaches 78.47% for the same number of Gaussians to model each basic environment. On the other hand, PD-MEMLIN with 16 Gaussians per phoneme obtains an improvement in WER of 75.43%, whereas PD-MEMLIN with cross-probability GMM reaches 77.72%. Since the computing cost for the proposed approach is higher, an alternative is considered: only the cross-probability GMM of the most probable pair of Gaussians are computed. So, only with the 1024 most probable pair of Gaussians, an improvement of 76.77% is obtained, when 128 Gaussians per basic environment are used in MEMLIN; and if the cross-probability model GMM is computed over the 8125 most probable pair of Gaussians, an improvement of 77.36% is obtained in PD-MEMLIN when each phoneme is modelled wih 32 Gaussians.

	♯PH	♯NG	♯CG	MWER	MIMP
PD-MEMLIN TD 2-2	8	2	2	6.00	70.58
PD-MEMLIN TD 4-4	8	4	4	5.58	73.49
PD-MEMLIN TD 8-8	8	6	6	5.39	74.82
PD-MEMLIN TD 16-16	13	12	12	5.04	77.25
PD-MEMLIN TD 32-32	13	25	25	5.02	77.36

Table 3. Mean WER (MWER) and mean improvement in WER (MIMP) in % for PD-MEMLIN when different phonemes and Gaussians of cross-probability GMM are computed.

7. BIBLIOGRAPHY

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

[2] Leonardo Neumeyer and Mitchel Weintraub, "Robust Speech Recognition in Noise Using Adaptation and Mapping Techniques," 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, "Robust speech recognition in cars using phoneme dependent multi-environment linear normalization," in *Proceedings of Interspeech*. Lisbon, Portugal, 2005, pp. 381–384.

[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] 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.