# Multi-Environment Linear Normalization for Robust Speech Analysis in Cars

*Luis Buera, Eduardo Lleida, Antonio Miguel, and Alfonso Ortega*

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

## Abstract

In this paper Phoneme-Dependent Multi-Environment Models based LInear feature Normalization, PD-MEMLIN, is presented. The target of this algorithm is learning the mismatch between clean and noisy feature vectors associated to a pair of Gaussians of the same phoneme (one for a clean model, and the other one for a noisy model), for each basic defined environment. These differences are estimated in a previous training process with stereo data. In order to compensate some of the problems of the independence assumption of the feature vectors components and the mismatch error between perfect and proposed transformations, two approaches have been proposed too: a multi-environment rotation transformation algorithm, and the use of transformed space acoustic models. The behavior of this technique was studied for speech recognition and speaker verification and identification in a real acoustic environment. The experiments were carried out with SpeechDat Car database and the results show an average improvement in speech recognition of more than 77% using PD-MEMLIN, and more than 85% using transformed space acoustic models and multi-environment rotation transformation. In speaker verification and identification, PD-MEMLIN is applied as a previous phase to clean the signal, with an average improvement in Equal-Error Rate of more than 70%, and 48.69%, respectively.

## 1. Introduction

When testing and training acoustic conditions are different, the accuracy of speech recognition, speaker verification, and speaker identification systems rapidly degrades. In order to compensate this mismatch, several techniques have been developed. They can be grouped into two important categories: acoustic models adaptation, and feature compensation, or normalization. The first one, which only modifies the acoustic models, can be more specific, whereas, normalization, which modifies the feature vectors, needs less data and computation time. The use of one or other kind of algorithms depends on the application. Hybrid techniques also exist, and they have proved to be effective [1]. However, in real dynamic environments, it may be impossible to retrain new acoustic models in all situations. In this cases, feature vector normalization techniques are a good option in order to improve the accuracy of speech recognition, and speaker verification and identification systems.

There are several feature compensation families [2], [3], but one of the most promised research line is based on Minimum Mean Squared Error, MMSE, estimation. Techniques like Stereo based Piecewise LInear Compensation for Environments, SPLICE [4], or Multi-Environment Models based LInear Normalization, MEMLIN [5], are some examples of MMSE based feature compensation. In this paper a Phoneme Dependent Multi-Environment Models based LInear Normalization, PD-MEMLIN, is proposed to clean the noisy signal.

In many cases, normalization techniques assume that the feature vector coefficients are independent. Thus, some kinds of transformations in the feature space, such as translations, can be properly handled, but not others, like rotations. Other problem in normalization techniques is the mismatch between perfect and proposed transformations. In this paper, two approaches are presented in order to compensate these problems. The first one is a multi-environment rotation transformation, which compensates the rotation produced in feature vectors by noisy environments. The second one is using transformed space acoustic models in recognition, which reduces the mismatch error between perfect and proposed normalization transformations. These techniques are treated with MEMLIN and PD-MEMLIN algorithms to study their behaviors in speech recognition.

Since PD-MEMLIN is a noise compensation algorithm, it can be used in a previous phase in order to clean the noisy signal, before speaker verification and identification.

This paper is organized as follows: in Section 2, PD-MEMLIN is presented. The multi-environment rotation technique is introduced in Section 3. The transformed space acoustic models strategy is explained in Section 4. In Section 5, the speaker verification and identification systems are presented. The results for speech recognition and speaker verification and identification, using PD-MEMLIN with SpeechDat Car database [6] are presented and discussed in Section 6. Finally, the conclusions are included in Section 7.

## 2. PD-MEMLIN

Phoneme Dependent Multi-Environment Models based LInear Normalization is an empirical feature vector normalization technique which uses stereo data in order to determine the different compensation linear transformations in a previous training process. 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.

Before obtaining the estimated clean feature vector based on MMSE criterion, some approximations have to be assumed:

### 2.1. Approximations

Three approximations are assumed: firstly, some basic environments are defined in the noisy space, and noisy feature vectors,

This work has been supported by the national project TIC2002-04103-C03-01 and Biosecure Neo.

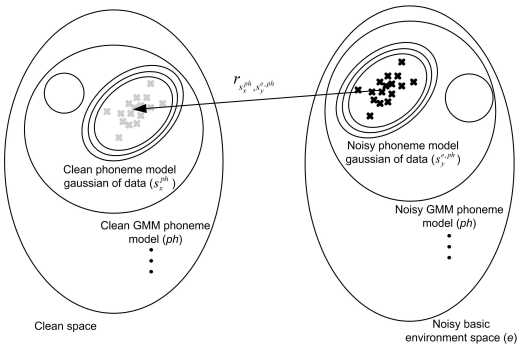


Figure 1: Scheme of PD-MEMLIN transformations for one environment.

$y$, follow the distribution of Gaussian mixture for each basic environment and phoneme

$$p_{e,ph}(y) = \sum_{s_y^{e,ph}} p(y|s_y^{e,ph})p(s_y^{e,ph}), \quad (1)$$

$$p(y|s_y^{e,ph}) = N(y; \mu_{s_y^{e,ph}}, \Sigma_{s_y^{e,ph}}), \quad (2)$$

where $s_y^{e,ph}$ denotes the correspondent Gaussian of the noisy model for the $e$ environment and $ph$ phoneme, and $\mu_{s_y^{e,ph}}$, $\Sigma_{s_y^{e,ph}}$, and $p(s_y^{e,ph})$ are the mean vector, the diagonal covariance matrix, and the weight associated to $s_y^{e,ph}$.

Secondly, clean feature vectors, $x$, are modelled following the distribution of Gaussian mixture

$$p_{ph}(x) = \sum_{s_x^{ph}} p(x|s_x^{ph})p(s_x^{ph}), \quad (3)$$

$$p(x|s_x^{ph}) = N(x; \mu_{s_x^{ph}}, \Sigma_{s_x^{ph}}), \quad (4)$$

where $s_x^{ph}$ denotes the correspondent Gaussian of the clean model and phoneme, and $\mu_{s_x^{ph}}$, $\Sigma_{s_x^{ph}}$, and $p(y|s_x^{ph})$ are the mean, diagonal covariance matrix, and the weight associated to $s_x^{ph}$.

Thirdly, for each time frame, $t$, $x$ is approached as a function, $\Psi$, of the noisy feature vector, $y_t$, clean model Gaussians, $s_x^{ph}$, and noisy environment model Gaussians, $s_y^{e,ph}$

$$x \simeq \Psi(y_t, s_x^{ph}, s_y^{e,ph}) = y_t - r_{s_x^{ph}, s_y^{e,ph}}, \quad (5)$$

where $r_{s_x^{ph}, s_y^{e,ph}}$ is the independent term of the linear transformation, and it depends on each pair of Gaussians, $s_x^{ph}$ and $s_y^{e,ph}$.

### 2.2. Cepstral enhancement

Given the noisy vector, $y_t$, the clean one is estimated by MMSE criterion

$$\hat{x}_t = E[x|y_t] = \int_x xp(x|y_t)dx, \quad (6)$$

where $p(x|y_t)$ is the Probability Density Function (PDF) of $x$ given $y_t$. With the three previous approximations, (6), can be approximated as expression (7).

In (7), $p(e|y_t)$ is the environment weight, $p(ph|y_t, e)$ is the probability of the phoneme $ph$, given the noisy feature vector and the environment, $p(s_y^{e,ph}|y_t, e, ph)$ is the probability of the noisy Gaussian given $y_t$, the environment, and the phoneme, and finally $p(s_x^{ph}|y_t, e, ph, s_y^{e,ph})$ is the probability of the clean Gaussian given $y_t$, $e$, $ph$ and $s_y^{e,ph}$.

$r_{s_x, s_y^{e,ph}}$ and $p(s_x^{ph}|y_t, e, ph, s_y^{e,ph})$ are computed through a previous training process. The other probabilities in (7) are estimated on line for each time frame in the recognition phase.

The probability of the environment, $p(e|y_t)$, is estimated using a recursive solution as

$$p(e|y_t) = \beta \cdot p(e|y_{t-1}) + (1-\beta)\frac{\sum_{ph} p_{e,ph}(y_t)}{\sum_e \sum_{ph} p_{e,ph}(y_t)}, \quad (8)$$

where $\beta$ is the memory constant, close to 1 (0.98 in this paper), and $p(e|y_0)$ is considered uniform for all environments. Also, $p(ph|y_t, e)$ and $p(s_y^{e,ph}|y_t, e, ph)$, are estimated as

$$p(ph|y_t, e) = \frac{p_{e,ph}(y_t)}{\sum_{ph} p_{e,ph}(y_t)}, \quad (9)$$

$$p(s_y^{e,ph}|y_t, e, ph) = \frac{p(y_t|s_y^{e,ph})p(s_y^{e,ph})}{\sum_{s_y^{e,ph}} p(y_t|s_y^{e,ph})p(s_y^{e,ph})}. \quad (10)$$

In order to compute $p(s_x^{ph}|y_t, e, ph, s_y^{e,ph})$, and $r_{s_x^{ph}, s_y^{e,ph}}$, a previous training process with available stereo data for each environment and phoneme is needed: $X_{e,ph} = \{x_1^{e,ph}, ..., x_{t_{e,ph}}^{e,ph}, ..., x_{T_{e,ph}}^{e,ph}\}$, for clean feature vectors and $Y_{e,ph} = \{y_1^{e,ph}, ..., y_{t_{e,ph}}^{e,ph}, ..., y_{T_{e,ph}}^{e,ph}\}$ for noisy ones, with $t_{e,ph} \in [1, T_{e,ph}]$.

The conditional probability, $p(s_x^{ph}|y_t, e, ph, s_y^{e,ph})$, can be considered time independent, and it may be estimated using (1), (2), (3), and (4): expression (11).

Finally, $r_{s_x^{ph}, s_y^{e,ph}}$ can be obtained by minimizing the weighted square error, $E_{s_x^{ph}, s_y^{e,ph}}$ (expressions (12) and (13)).

In (12) and (13), $p(s_y^{e,ph}|y_{t_{e,ph}}^{e,ph}, e, ph)$ is the probability of $s_y^{e,ph}$, given the noisy feature vector, $y_{t_{e,ph}}^{e,ph}$, the environment, and the phoneme. It can be obtained as (10). Also, in the same expressions, $p(s_x^{ph}|x_{t_{e,ph}}^{e,ph}, e, ph)$ is the probability of $s_x^{ph}$ given the clean feature vector, $e$ and $ph$, and it is estimated as

$$p(s_x^{ph}|x_{t_{e,ph}}^{e,ph}, e, ph) = \frac{p(x_{t_{e,ph}}^{e,ph}|s_x^{ph})p(s_x^{ph})}{\sum_{s_x^{ph}} p(x_{t_{e,ph}}^{e,ph}|s_x^{ph})p(s_x^{ph})}. \quad (14)$$

## 3. Multi-environment rotation transformation

The goal of rotation transformation [7] is to obtain a transformation matrix ($U_1$) in order to normalize the feature vector

$$\hat{x}_t = U_1 y_t, \quad (15)$$

where index 1 means that the rotation modifies only the direction of the biggest variance feature space axes. With the stereo database training corpus, a transformation matrix can be obtained, $U_{e,1}$, for each basic environment, $e$. Principal Component Analysis (PCA) of the covariance matrixes of clean, and noisy feature vectors for each environment, ($\tilde{\Sigma}_e$, $\Sigma_e$, respectively) is performed in order to determine the most important axes of clean and noisy data spaces. The corresponding orthonormal eigenvectors and eigenvalues are: $\tilde{v}_{e,i}$, and $\tilde{\lambda}_{e,i}$ for clean space, and $v_{e,i}$, and $\lambda_{e,i}$, for the noisy one, where $i = 1...D$, $\tilde{\lambda}_{e,1} \geq \tilde{\lambda}_{e,2} \geq ... \geq \tilde{\lambda}_{e,D}$, $\lambda_{e,1} \geq \lambda_{e,2} \geq ... \geq \lambda_{e,D}$, and $D$ is the dimension of the feature vectors. The rotation

$$\hat{x}_t \simeq y_t - \sum_{e}\sum_{ph}\sum_{s_x^{ph}}\sum_{s_y^{e,ph}} r_{s_x^{ph},s_y^{e,ph}} p(e|y_t)p(ph|y_t,e)p(s_y^{e,ph}|y_t,e,ph)p(s_x^{ph}|y_t,e,ph,s_y^{e,ph}). \quad (7)$$

$$p(s_x^{ph}|y_t,e,ph,s_y^{e,ph}) \simeq p(s_x^{ph}|s_y^{e,ph}) = \frac{\sum_{t_{e,ph}} p(x_{t_{e,ph}}^{e,ph}|s_x^{ph})p(y_{t_{e,ph}}^{e,ph}|s_y^{e,ph})p(s_x^{ph})p(s_y^{e,ph})}{\sum_{t_{e,ph}}\sum_{s_x^{ph}} p(x_{t_{e,ph}}^{e,ph}|s_x^{ph})p(y_{t_{e,ph}}^{e,ph}|s_y^{e,ph})p(s_x^{ph})p(s_y^{e,ph})}. \quad (11)$$

$$E_{s_x^{ph},s_y^{e,ph}} = \sum_{t_{e,ph}} p(s_x^{ph}|x_{t_{e,ph}}^{e,ph},e,ph)p(s_y^{e,ph}|y_{t_{e,ph}}^{e,ph},e,ph)(x_{t_{e,ph}}^{e,ph} - y_{t_{e,ph}}^{e,ph} + r_{s_x^{ph},s_y^{e,ph}})^2. \quad (12)$$

$$r_{s_x^{ph},s_y^{e,ph}} = arg \min_{r_{s_x^{ph},s_y^{e,ph}}} (E_{s_x^{ph},s_y^{e,ph}}) = \frac{\sum_{t_{e,ph}} p(s_x^{ph}|x_{t_{e,ph}}^{e,ph},e,ph)p(s_y^{e,ph}|y_{t_{e,ph}}^{e,ph},e,ph)(y_{t_{e,ph}}^{e,ph} - x_{t_{e,ph}}^{e,ph})}{\sum_{t_{e,ph}} p(s_x^{ph}|x_{t_{e,ph}}^{e,ph},e,ph)p(s_y^{e,ph}|y_{t_{e,ph}}^{e,ph},e,ph)}, \quad (13)$$

| | Angles ($^0$) |
|---|---|
| Ch0 - Ch2 | 21.02 |
| Ch0 - MEMLIN 128-128 | 6.11 |
| Ch0 - PD-MEMLIN 16-16 | 5.98 |
| Ch0 - MEMLIN 128-128 + rot | 2.45 |
| Ch0 - PD-MEMLIN 16-16 + rot | 4.21 |

Table 1: Angles in degrees between the highest variance axes, where rot indicates that multi-environment rotation transformation is applied after normalization techniques.

angle between the two principal directions of clean and noisy spaces is obtained as: $\eta_{e,1} = \arccos(\tilde{v}_{e,1} \cdot v_{e,1})$. It can be considered that $\tilde{v}_{e,1}$ and $v_{e,1}$ determine an hyperplane, $\pi_{e,1}$. The geometric idea of this normalization technique is to split each vector into two parts: the projection over $\pi_{e,1}$, which will be rotated $\eta_{e,1}$ degrees, and the perpendicular part, which will not be modified.

Since $\tilde{v}_{e,1}$ and $v_{e,1}$ are not orthogonal, Gram-Schmidt is applied to $v_{e,1}$ to obtain an orthonormal basis vector $\hat{v}_{e,1}$, lying in the same rotation hyperplane

$$\hat{v}_{e,1} = \frac{v_{e,1} - (\tilde{v}_{e,1} \cdot v_{e,1}) \cdot \tilde{v}_{e,1}}{\|v_{e,1} - (\tilde{v}_{e,1} \cdot v_{e,1}) \cdot \tilde{v}_{e,1}\|}. \quad (16)$$

$J_{e,1}^T$ is the projection matrix of $\pi_{e,1}$, and $R_{e,1}$ is the rotation transformation for the angle $\eta_{e,1}$

$$J_{e,1}^T = (\hat{v}_{e,1}, \tilde{v}_{e,1})^T. \quad (17)$$

$$R_{e,1} = \begin{pmatrix} \cos(\eta_{e,1}) & -\sin(\eta_{e,1}) \\ \sin(\eta_{e,1}) & \cos(\eta_{e,1}) \end{pmatrix}. \quad (18)$$

Finally, the transformation matrix for the correspondent environment, $U_{e,1}$, can be obtained as

$$U_{e,1} = J_{e,1} R_{e,1} J_{e,1}^T + I + J_{e,1} J_{e,1}^T, \quad (19)$$

where $I$ is the identity matrix. The rotation can be performed in all the axes, not only for the biggest variance one, but it can be shown that with the first vector is enough [7]. In recognition, all frames of each utterance are normalized with the most probable environment, $\hat{e}$, matrix: $U_1 = U_{\hat{e},1}$.

The behavior of the multi-environment rotation transformation technique can be observed in Table 1, where Ch0 - Ch2 indicates the average angle between the most important axes of clean (Ch0) and noisy (Ch2) testing signals of SpeeechDat Car database. Ch0 - MEMLIN 128-128 represents the angle between clean and normalized feature vectors axes when MEMLIN technique is used with 128 Gaussians for noisy and clean models. Ch0 - PD-MEMLIN 16-16 indicates the angle between clean and normalized feature vectors axes when PD-MEMLIN is applied with 16 Gaussians for each phoneme and environment. The results show that the normalization technique is not enough in order to compensate the rotation produced by the environment noises. If normalized signal is transformed by multi-environment rotation transformation technique, the angles decrease. The results are better with MEMLIN due to rotation transformation with PD-MEMLIN produces a rough modification in the transformed space because it is only applied only one transformation for environment, without any phoneme dependence.

## 4. Transformed space acoustic models

Normalization techniques map the noisy feature vectors into the clean space. Since they do not generate a perfect transformation, the new transformed space is not the clean one as it should be. This mismatch error can be compensated with the acoustic models in recognition. By transformed space acoustic models we mean new acoustic models trained with normalized features. The new models are obtained through three phases:

- Normalization training process (obtaining $r_{s_x^{ph},s_y^{e,ph}}$ and $p(s_x^{ph}|y_t,e,ph,s_y^{e,ph})$.
- Compensation of the noisy training data used in training process.
- New acoustic models are trained with normalized noisy training data.

## 5. Speaker verification and identification systems

For the verification task, an independent text Universal Background Model GMM was developed, UBM-GMM [8]. The input of the system is composed of the 12 normalized MFCC with cepstral mean substraction, the first and second derivative and the normalized delta energy, given a feature vector of 37 coefficients. A simple VAD based on energy is used in order to verify only with speech signal. The average length of the utterances handled in verification and identification tasks is 3 seconds.

The universal background model is trained by Maximum Likelihood criterion, ML, using Expectation-Maximization, EM, algorithm [9], with four iterations. The speakers Gaussian models are retrained from UBM by Maximum A Posteriori, MAP, algorithm [10].

Given a sequence of feature vectors of speaker $i$, $Y_i$, an UBM, $\lambda_{UBM}$, and the correspondent speaker model, $\lambda_i$, the decision to determine if the speaker is right will be

| Train | Test | E1 | E2 | E3 | E4 | E5 | E6 | E7 | MWER (%) |
|---|---|---|---|---|---|---|---|---|---|
| Ch0 | Ch0 | 1.90 | 2.64 | 1.81 | 1.75 | 1.62 | 0.64 | 0.35 | 1.75 |
| Ch0 | Ch2 | 5.91 | 14.49 | 14.55 | 20.17 | 21.07 | 16.19 | 35.71 | 16.21 |
| Ch2 | Ch2 | 6.67 | 14.24 | 12.73 | 12.91 | 14.97 | 9.68 | 8.50 | 11.81 |

Table 2: WER baseline results, in %.

| | MWER (%) | IMP (%) |
|---|---|---|
| SPLICE | 7.57 | 57.92 |
| PD-MEMLIN | 5.30 | 77.67 |
| PD-MEMLIN + rot | 5.37 | 76.82 |
| MEMLIN | 6.06 | 72.24 |
| MEMLIN + rot | 5.65 | 76.32 |
| PD-MEMLIN + ac | 4.64 | 79.39 |
| PD-MEMLIN + rot + ac | 4.79 | 78.55 |
| MEMLIN + ac | 4.16 | 84.42 |
| MEMLIN + rot + ac | 4.09 | 85.02 |

Table 3: Best mean WER and improvement for different techniques, in %, where rot and ac indicate that multi-environment rotation transformation or transformed space acoustic models are respectively used.

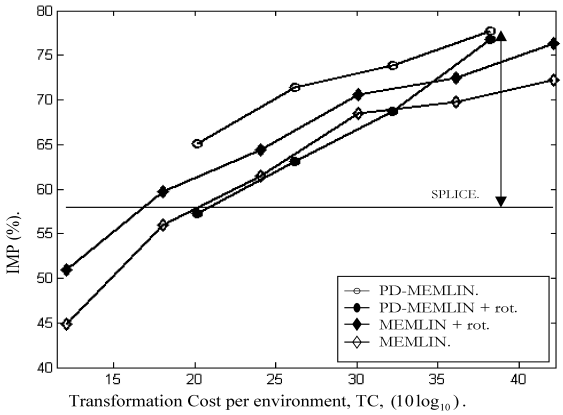


Figure 2: Improvement, in %, for different techniques, where rot indicates that multi-environment rotation transformation is used after normalization techniques.

$$\text{if } \frac{p(Y_i|\lambda_i)}{p(Y_i|\lambda_{UBM})} \begin{cases} < \theta \Rightarrow \text{reject } \lambda_i, \\ \geq \theta \Rightarrow \text{accept } \lambda_i, \end{cases} \quad (20)$$

where $p(Y_i|\lambda_i)$ is the score of $Y_i$, given the model $\lambda_i$, $p(Y_i|\lambda_{UBM})$ is the score of $Y_i$, given the universal background model, and finally, $\theta$ is the threshold, which is empirically obtained when false accept rate and false reject rate are similar.

To identify, a GMM system is developed. The same $\lambda_i$ speaker models are used, and for each speech utterance $Y$, the highest model score, $p(Y|\lambda_i)$, will determinate the estimation speaker, $\hat{i}$

$$\hat{i} = arg\max_i p(Y|\lambda_i). \quad (21)$$

## 6. RESULTS

A set of experiments have been carried out using the Spanish SpeechDat Car database [6] in order to study the behavior of the presented techniques in speech recognition, and speaker verification and identification. Noisy space is split in seven basic environments: 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).

All the utterances are 16 KHz sampled. The clean signals (Ch0) are recorded with a close talk microphone (Shune SM-10A), and the noisy signals (Ch2) are recorded by a microphone placed on the car ceiling in front of the driver (Peiker ME15/V520-1). The SNR range for the clean signals goes from 20 to 30 dB, and for the noisy signals goes from 5 to 20 dB. 12 MFCC and energy are computed each 10 ms using a 25 ms hamming window.

The feature normalization techniques (PD-MEMLIN and MEMLIN to compare) are applied over the 12 MFCC and delta

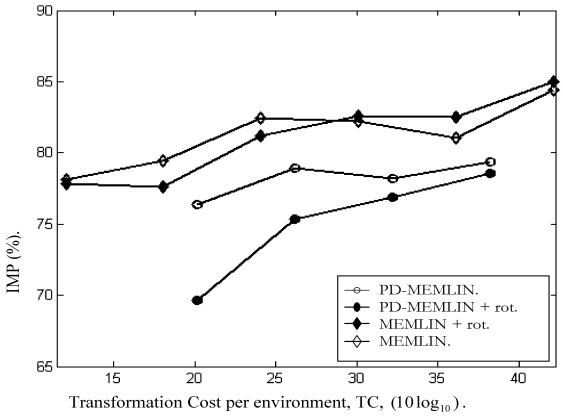


Figure 3: Improvement, in %, for different techniques with transformed space acoustic models, where rot indicates that multi-environment rotation transformation is used.

energy, and the different used models have 4, 8, 16, 32, 64 and 128 Gaussians for MEMLIN, and 26 Spanish phonemes with 2, 4, 8, or 16 Gaussians for each one in PD-MEMLIN.

For speech recognition, the feature vector is composed of the 12 normalized MFCC with cepstral mean substraction, the first and second derivative and the normalized delta energy, given a feature vector of 37 coefficients. The phonetic acoustic models are composed of 25 three state continuous density HMM with 16 Gaussians per state to model Spanish phonemes and 2 silence models for long and interword silences. The task consists on isolated and continuous digits.

The Word Error Rate, WER, baseline results for each environment are presented in Table 2. MWER represents the Mean WER, computed proportionality to the number of utterances of each environment.

In order to compare the presented techniques, the trans-

formation cost per environment is defined as: $TC = 10log_{10}(N_{ph}N_{s_x^{ph}}N_{s_y^{ph}})$, where $N_{ph}$ is the number of phonemes, $N_{s_x^{ph}}$ is the number of clean Gaussians for each phoneme, and $N_{s_y^{ph}}$ is the number of noisy Gaussians for each phoneme and environment. For MEMLIN, the number of phonemes can be considered as 1.

The comparative results in speech recognition between MEMLIN and PD-MEMLIN, with or without multi-environment rotation transformation, are shown in Fig. 2. It is presented the improvement, IMP, which has been calculated with the improvement of each environment and proportionality to the number of utterances of each environment. The best IMP and MWER are included in Table 3 for the different cases. In order to compare, the values for SPLICE [4] with 128 Gaussians for noisy model are included, too. It can be observed that multi-environment rotation transformation produces an improvement when it is applied with MEMLIN, but not when it is applied with PD-MEMLIN. The reason is the difference between normalized training data, which is used in order to obtain the rotation transformations, and normalized testing data is higher in PD-MEMLIN than in MEMLIN, and that rotation transformation with PD-MEMLIN produces a rough modification in the transformed space because it is used only one transformation for environment, without any phoneme dependence. In any case, PD-MEMLIN obtains the highest results, obtaining an improvement of 77.67%, almost 20% more than SPLICE.

The comparative results between MEMLIN and PD-MEMLIN, with or without multi-environment rotation transformation, and with transformed space acoustic models are shown in Fig. 3. Also the best values are presented in Table 3. The results are better than those obtained without the transformed acoustic models, specially in WER because the biggest improvements are in more noisy environments, which have the highest WERs. Another advantage of using transformed-space acoustic models is that the results are less dependent on the number of transformation Gaussians. The higher difference between normalized training data and normalized testing data for PD-MEMLIN is the reason of results with MEMLIN are better. The best improvement is obtained with MEMLIN + rot + ac: 85.02%.

In order to study speaker verification and identification in different acoustic conditions, the universal background model in verification is obtained with the training corpus of Spanish SpeechDat Car (182 speakers and 16108 sentences) and it is composed of 512 Gaussians. Testing corpus of the database is used to prove the verification and identification systems. There are 91 speakers with approximately 112 sentences: 50 selected from all environments to train the 512 Gaussian speaker models and approximately 62 from all environments to test the systems. These 91 speakers are different from the 182 training corpus ones. The results can be seen in Table 4 and Table 5, where EER is Equal-Error Rate, in %, Ch0-Ch0 indicates the results when clean signal is used to test and train the speakers and UBM models (clean models), Ch0-Ch2 means the results when noisy signal is used to verify with clean models, Ch2-Ch2 uses noisy signal to test and train the models, $Ch0 - Ch2_{nor}$ tests with normalized signal and clean models, and IMP is the improvement obtained with the performance of $Ch0 - Ch2_{nor}$ compared to the Ch0-Ch0 and Ch0-Ch2 margin in %.

The number of utterances used for each environment is: 254 for E1, 290 for E2, 235 for E3, 238 for E4, 254 for E5, 247 for E6 and 47 for E7. In verification, for each utterance, one of the 91 possible speakers is considered as author of it each time; so, the system has to detect in each case if the speaker is the right one, or not.

It can be observed in Table 4 that noise produces an important degradation in the behavior of the system: EER falls down from near 1%, to 26%. If noisy signal is treated with PD-MEMLIN, the improvement is significant, obtaining 8.64% in false accept and false reject rates: this is, an improvement of near 70%. Global results with all environments and different thresholds are presented in Fig. 4, where Ch0-Ch0 is represented with a solid line, Ch2-Ch2 is printed with a dash line, $Ch0 - Ch2_{nor}$ with dash and dot line, and finally, Ch0-Ch2 is printed with a dot line. The threshold is varied with a step of 0.05.

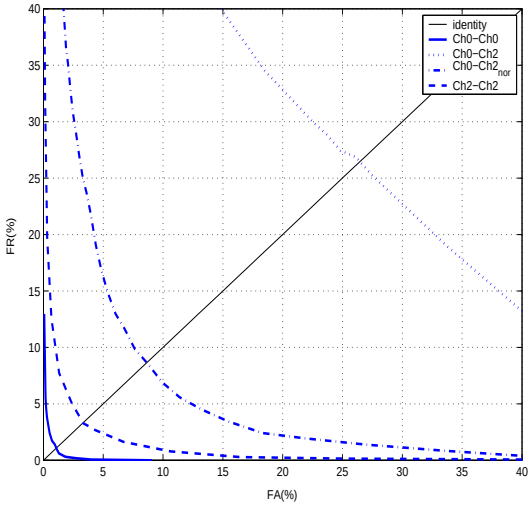


Figure 4: Total verification results with all environments and different thresholds.

In identification task, which success rate results in % are presented in Table 5, it can be observed that noise degrades the behavior of the system and the results are very poor concerning the ones obtained with clean signal: 99.69% versus 22.02% (average results). Since PD-MEMLIN is used, the success rate increases until 59.84%: this is an improvement of 48.69%.

Although the improvements, the results in speaker verification and identification obtained with normalized signal are far away from Ch2-Ch2. Anyhow, in many cases noisy speaker models are not available because it is not possible to retrain the models in all acoustic conditions. In this sense, normalization techniques are a good approximation to Ch0-Ch0 results. The reason of the results obtained when PD-MEMLIN is applied, is that the learnt transformations project from noisy space to a very general clean one, loosing the speaker specificity. Since it is very important in verification and identification tasks, speaker clustering techniques can be advantageous in order to define speaker-dependent transformations. In this sense, similar projections would be used for the same kind of speakers, the speaker specificity would not lose performance, and the results could be improved.

## 7. CONCLUSIONS

In this paper we have presented a feature vector normalization, PD-MEMLIN, and two approaches in order to compensate the

| EER (%) | Ch0-Ch0 | Ch0-Ch2 | Ch2-Ch2 | $Ch0 - Ch2_{nor}$ | IMP (%) |
|---|---|---|---|---|---|
| E1 | 1.55 | 10.50 | 0.79 | 5.13 | 60.00 |
| E2 | 1.21 | 26.73 | 4.70 | 9.58 | 67.20 |
| E3 | 0.87 | 24.61 | 3.91 | 11.80 | 53.96 |
| E4 | 0.89 | 27.42 | 2.08 | 6.15 | 70.17 |
| E5 | 0.91 | 26.93 | 2.02 | 7.17 | 75.94 |
| E6 | 1.08 | 35.00 | 2.71 | 9.71 | 74.56 |
| E7 | 0.29 | 41.45 | 0.46 | 9.05 | 78.72 |
| Total | 1.06 | 26.50 | 3.29 | 8.64 | 70.20 |

Table 4: Verification results with PD-MEMLIN for each environment

| Success rate (%) | E1 | E2 | E3 | E4 | E5 | E6 | E7 | Total (%) |
|---|---|---|---|---|---|---|---|---|
| Ch0-Ch0 | 99.6 | 99.65 | 99.57 | 99.15 | 100 | 100 | 100 | 99.69 |
| Ch0-Ch2 | 65.35 | 13.44 | 27.65 | 12.60 | 10.72 | 11.67 | 0 | 22.02 |
| Ch2-Ch2 | 98.03 | 95.52 | 91.06 | 97.89 | 96.52 | 87.55 | 100 | 94.89 |
| $Ch0 - Ch2_{nor}$ | 86.22 | 61.37 | 49.36 | 68.90 | 44.34 | 56.03 | 48.93 | 59.84 |
| IMP (%) | 60.93 | 55.60 | 30.19 | 65.05 | 37.66 | 50.22 | 48.93 | 48.69 |

Table 5: Identification results with PD-MEMLIN for each environment

feature vector rotation generated by noise (multi-environment rotation transformation) and the mismatch between the perfect and proposed normalization transformations (transformed space acoustic models). Important improvements are obtained in speech recognition with PD-MEMLIN (77.67%), better than other techniques as MEMLIN or SPLICE. When multi-environment rotation transformation and transformed space acoustic models are applied with MEMLIN, an improvement of 85.02% is obtained. In speaker verification and identification tasks, PD-MEMLIN is used as a previous phase to clean noisy feature vectors to improve the results in adverse and dynamic acoustic conditions. An UBM-GMM system has been developed for verification, and a GMM system for identification. The results show that noise degrades seriously the accuracy of the systems, but pre-processing the noisy signal with PD-MEMLIN in verification, an average improvement of 70.20% in EER is obtained. In identification, the improvement using PD-MEMLIN reaches until 48.69%, using clean speaker models. As a future work line to improve these results, speaker dependent transformations are proposed.

## 8. References


[1] A. Sankar and C. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, pp. 190–202, May 1996. [Online]. Available: citeseer.nj.nec.com/181474.html

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

[3] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language, Vol. 12*, 1998.

[4] J. Droppo, L. Deng, and A. Acero, "Evaluation of the splice algorithm on the aurora2 database," *in Proc. Eurospeech*, vol. 1, Sep. 2001.

[5] L. Buera, E. Lleida, A. Miguel, and A. Ortega, "Multi-environment models based linear normalization for speech recognition in car conditions," *in Proc. ICASSP*, May. 2004.

[6] A. Moreno, A. Noguiera, and A. Sesma, "Speechdat-car: Spanish," *Technical Report SpeechDat*.

[7] S. Molau, "Normalization in the acoustic feature space for improved speech recognition," *Ph. D. Thesis*, Computer Science Department, RWTH Aachen. Feb. 2003.

[8] R. B. D. D. A. Reynold, T. F. Quatieri, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing, Vol. 10, pp. 19-41*, 2000.

[9] J. Bilmes, "A gentle tutorial on the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," University of Berkeley, ICSI-TR-97-021, 1997. [Online]. Available: citeseer.nj.nec.com/bilmes98gentle.html

[10] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriory estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. on Speech and Audio Processing, Vol. 2, pp. 291-298*.