Speaker verification and identification using Phoneme Dependent Multi-Environment Models based LInear Normalization in adverse and dynamic acoustic environments

L. Buera, E. Lleida, J.D. Rosas, J. Villalba, A. Miguel, A. Ortega, O. Saz.

University of Zaragoza, Spain
{lbuera, lleida, jdrosas, villalba, amiguel, ortega, oskarsaz}@unizar.es

Abstract

In adverse acoustic conditions, speaker verification and identification system rates degrade very significatively. In order to compensate this effect, several techniques can be applied. In this paper Phoneme Dependent Multi-Environment Models based LInear Normalization, PD-MEMLIN, is presented as a solution in order to clean signal in a early processing step. In this algorithm, clean and noisy spaces are modelled by mixtures of gaussians for each phoneme, and a linear transformation is learnt with stereo data for each phoneme pair of gaussians: one for the clean space and the other one for the noisy space. Some experiments with Spanish SpeechDat Car database were carried out in order to study the behavior of the proposed technique in verification and identification tasks. With an UBM-GMM verification system, important average improvement in Equal-Error Rate (EER) is obtained (70.20%). The improvement in identification task with a GMM system is 48.69%.

1. Introduction

When acoustic conditions are adverse, the accuracy of speaker verification and identification systems rapidly degrades. In order to compensate this effect, several techniques have been developed [1]. Since verification and identification systems are based on Gaussian Mixture Models, GMM, two kinds of robust adaptation techniques can be developed: acoustic models adaptation and feature vector normalization. The first one, which only modifies the gaussian mixture models, can be more specific, whereas, feature compensation, which modifies the feature vectors, needs less data and computation time. In GMM systems, in which each speaker has his own acoustic model, feature normalization and acoustic model adaptation can be simultaneously used: feature normalization clean firstly the noisy signal, and after that, the speaker acoustic model is retrained with normalized signal. However, in real dynamic environments, it can be impossible to retrain speaker models in all situations. In this cases, feature vector normalization techniques are a good option in order to improve the accuracy of the system.

Feature compensation algorithms can be divided into three main kinds of techniques [2]: model-based compensation, empirical compensation by direct cepstral comparison, and compensation via cepstral high pass filtering. The first group uses a mathematical model in order to represent the environment, and the parameters of the supposed environment are estimated with frames of degraded speech (like Vector Taylor Series for normalization, VTS, [3] or Codeword Dependent Cepstral Normalization, CDCN, [4]). Empirical compensation does not assume any defined environment, and uses stereo data in order to determinate the correspondent degradation (like multivariate gaussian-based cepstral normalization, RATZ [3], Stereo based Piecewise Linear Compensation for Environments, SPLICE [5], or Multi-Environment Models based LInear Normalization, MEMLIN [6]). Cepstral high-pass filtering does not obtain as good results as the other kind of algorithms, but the computational cost is almost zero (like Cepstral Mean Normalization, CMN, [2]).

In this paper, an empirical compensation technique is presented in order to improve speaker verification and identification GMM systems: PD-MEMLIN, Phoneme Dependent Multi-Environment Models based LInear Normalization. This stereo data based algorithm uses a Minimum Mean Squared Error, MMSE, estimator and the proposed linear transformations, which are obtained in a training process, depend on clean and noisy phoneme model gaussians. With this algorithm, a quick feature vector adaptation to dynamic environments is obtained, and it can be useful when it is not available an adapted speaker model for each acoustic condition, as cars, for example.

This paper is organized as follows: in Section 2, a PD-MEMLIN is presented. Verification and identification used systems are studied in Section 3. The experiments carried out with Spanish SpeechDat Car database [7] are explained in Section 4, showing and discussing the results obtained. Finally in Section 5, the conclusions are presented.

2. Phoneme Dependent 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 linear transformations in a training process. The clean feature space is modelled as a mixture of gaussians for each phoneme. The noisy one is divided in several basic acoustic environments and each environment is modelled as a mixture of gaussians for each phoneme. The transformations are estimated between a clean phoneme gaussian and a noisy gaussian of the same phoneme, and this, for all basic acoustic environments. This is shown in figure 1 for one environment.

2.1. MMSE estimator

Given the noisy feature vector for each time frame, t, y_t, the clean estimation vector, $\hat{x}_t$, can be calculated by MMSE estimation, where x is the clean feature vector:

This work has been supported by MCyT under contract TIC2002-04103-C03

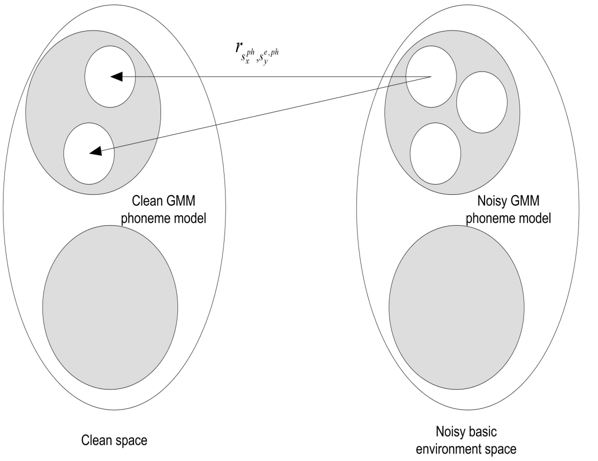
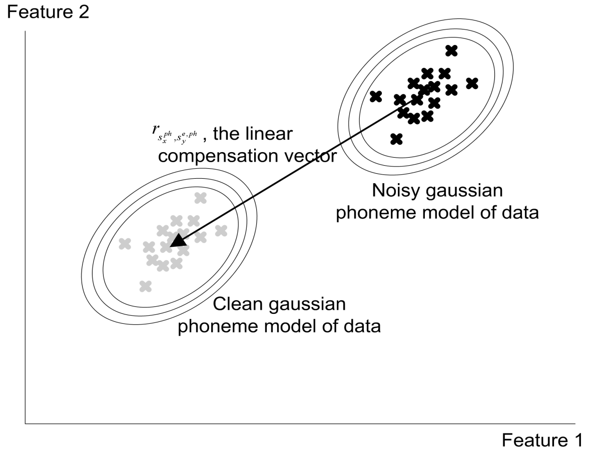


Figure 1: Scheme of PD-MEMLIN transformations for one environment.



Feature 2
$r_{s_x^{ph},s_y^{e,ph}}$, the linear compensation vector
Noisy gaussian phoneme model of data
Clean gaussian phoneme model of data
Feature 1

Figure 2: Scheme of one pair of gaussians transformation for noisy data.

$$\hat{x}_t = E[x|y_t] = \int_x x \cdot p(x|y_t)dx. \tag{1}$$

The problem is how is approximated x and how the probability density function (PDF) of x given y_t, $p(x|y_t)$, can be obtained. In order to calculate them, some approximations can be applied.

PD-MEMLIN supposes that noisy space can be divided into e basic environments. For each one, noisy feature vectors follow a distribution of mixture of gaussians for each phoneme, ph:

$$p_{e,ph}(y_t) = \sum_{s_y^{e,ph}} p(y_t|s_y^{e,ph})p(s_y^{e,ph}), \tag{2}$$

$$p(y_t|s_y^{e,ph}) = N(y_t; \mu_{s_y^{e,ph}}, \Sigma_{s_y^{e,ph}}), \tag{3}$$

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

PD-MEMLIN assumes that clean feature vectors model follows a distribution of mixture gaussians for each phoneme:

$$p_{ph}(x) = \sum_{s_x^{ph}} p(x|s_x^{ph})p(s_x^{ph}), \tag{4}$$

$$p(x|s_x^{ph}) = N(x; \mu_{s_x^{ph}}, \Sigma_{s_x^{ph}}), \tag{5}$$

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

On the one hand, PD-MEMLIN approximates x as a linear function of y_t, s_x^{ph}, and $s_y^{e,ph}$:

$$x \approx \Psi(y_t, s_x^{ph}, s_y^{e,ph}) = y_t - r_{s_x^{ph},s_y^{e,ph}}, \tag{6}$$

where $r_{s_x^{ph},s_y^{e,ph}}$ is the independent term of the PD-MEMLIN linear transformation function associated to clean phoneme-depended gaussian and a noisy environment-phoneme-depended gaussian. This can be shown in figure 2, where the mismatch between noisy and clean data and the linear compensation vector are shown for one pair of gaussians. With this approximation, the equation (1) will be transformed into expression (7), where $p(e|y_t)$ is the probability of the environment, given the noisy feature vector y_t; $p(ph|y_t, e)$ is the probability of the phoneme ph, given y_t and e; $p(s_y^{e,ph}|y_t, e, ph)$ is the probability of $s_y^{e,ph}$, given y_t, e, and ph, and finally, $p(s_x^{ph}|y_t, e, ph, s_y^{e,ph})$ is the probability of s_x^{ph}, given y_t, e, ph, and $s_y^{e,ph}$.

2.2. MMSE Parameters estimation

In order to calculate the estimator vector, $\hat{x}_t$, different variables have to be obtained. Since some of them are dependent on noisy feature vector each time, they are computed during the verification or identification processes. The other ones have to be calculated in a previous training process with stereo data.

The variables which need to be obtained are: $p(e|y_t)$, $p(ph|y_t, e)$, $p(s_y^{e,ph}|y_t, e, ph)$, $p(s_x^{ph}|y_t, e, ph, s_y^{e,ph})$, and $r_{s_x^{ph},s_y^{e,ph}}$. The first three expressions need to be calculated in speaker verification and identification, and the other ones are learnt in a training process.

The probability of the environment, $p(e|y_t)$, is calculated with an iterative solution. Each frame, t ($t \in 1, ..., T$), a noisy feature vector is available, y_t. The calculation of the environment weight in this moment will be, using (2), (3), and Bayes theorem:

$$p(e|y_t) = \beta \cdot p(e|y_{t-1}) + (1-\beta)\frac{\sum\limits_{ph} p_{e,ph}(y_t)}{\sum\limits_{e}\sum\limits_{ph} p_{e,ph}(y_t)}, \tag{8}$$

where β is the memory constant. $p(e|y_0)$ is considered uniform for all environments. Also, $p(ph|y_t, e)$ can be calculated using (2), (3), and Bayes theorem:

$$p(ph|y_t, e) = \frac{p_{e,ph}(y_t)}{\sum\limits_{ph} p_{e,ph}(y_t)}. \tag{9}$$

$p(s_y^{e,ph}|y_t, e, ph)$ can be calculated using (2), (3), and Bayes theorem:

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

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

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 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 approximated by $p(s_x^{ph}|s_y^{e,ph})$. It may be estimated with stereo training data using (2), (3), (4), (5) and Bayes theorem: expression (11).

The estimate of $r_{s_x^{ph},s_y^{e,ph}}$ can be obtained by minimizing the weighted square error, $E_{s_x^{ph},s_y^{e,ph}}$: (expressions (12), and (13)), where $p(s_x^{ph}|x_{t_{e,ph}}^{e,ph},e,ph)$ is the probability of s_x^{ph} given the clean feature vector, e environment and ph phoneme. It can be calculated with (4), (5), and Bayes theorem:

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

The use of clean and noisy phoneme dependent gaussian linear transformations makes PD-MEMLIN be a more specific normalization technique than other algorithms based on stereo data as RATZ [3], which only uses clean gaussian models, SPLICE [5], which uses noisy gaussian models, or MEMLIN, which represents clean and noisy spaces as mixture of gaussians, but does not use phoneme dependence [6].

3. Verification and identification systems

For the verification task, an independent text Universal Background Model GMM was developed, UBM-GMM. 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 sentences used in verification and identification tasks is 3 seconds.

The universal background model is trained by Expectation-Maximization, EM, algorithm [8], with four iterations. The speakers gaussian models are retrained from UBM by Maximum A Posteriori, MAP, algorithm [9].

Given a sequence of feature vectors of speaker i, Y_i, an UBM, λ_{UBM}, and the correspondent speaker model, λ_i, the decision to determine if the speaker is right will be:

$$\text{if } \frac{p(Y_i|\lambda_i)}{p(Y_i|\lambda_{UBM})} \begin{cases} < \theta \Rightarrow \text{reject } \lambda_i, \\ \geq \theta \Rightarrow \text{accept } \lambda_i, \end{cases} \qquad (15)$$

where $p(Y_i|\lambda_i)$ is the score of Y_i, given the model λ_i, $p(Y_i|\lambda_{UBM})$ is the score of Y_i, given the universal background model, and finally, θ 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 λ_i speaker models are used, and for each speech utterance Y_i, the highest model score, $p(Y_i|\lambda_i)$, will determinate the estimation speaker, $\hat{i}$:

$$\hat{i} = arg\,\max_{i} p(Y_i|\lambda_i). \qquad (16)$$

4. Experiments

In order to study speaker verification and identification in different acoustic conditions, a set of experiments have been carried out using the Spanish SpeechDat Car database [7], which has stereo data. Although this database is not properly to verification ans identification tasks, because there is only a long continuous session per speaker, however, the acoustic situations are so different and dynamic that makes it interesting to study the PD-MEMLIN behavior. Seven environments are 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).

All the utterances are 16 KHz sampled. The clean signals are recorded with a close talk microphone (Shune SM-10A), which is called Ch0, and the noisy signals are recorded by a microphone placed on the car ceiling in front of the driver (Peiker ME15/V520-1): it is called Ch2. The SNR range for the clean signals goes from 20 to 30 dB, and for the noisy signals goes from 4 to 14 dB. 12 MFCC and energy are computed each 10 ms using a 25 ms Hamming window.

The feature normalization technique is applied over the 12 MFCC and delta energy. Mixtures of 16 gaussians are used for each phone dependent model.

The universal background model in verification is calculated with the training corpus of Spanish SpeechDat Car (218 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 218 training corpus ones. The results can be seen in table 1 and table 2, where E1,...E7 represent the different environments, 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 indicates 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 %.

$$p(s_x^{ph}|s_y^{e,ph}) = \frac{\sum\limits_{t_{e,ph}} p(x_{t_{e,ph}}^{e,ph}|s_x^{ph})p(y_{t_{e,ph}}^{e,ph}|s_y^{e,ph})p(s_x^{ph})p(s_y^{e,ph})}{\sum\limits_{t_{e,ph}}\sum\limits_{s_x^{ph}} p(x_{t_{e,ph}}^{e,ph}|s_x^{ph})p(y_{t_{e,ph}}^{e,ph}|s_y^{e,ph})p(s_x^{ph})p(s_y^{e,ph})}. \tag{11}$$

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

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

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 1: 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 2: Identification results with PD-MEMLIN for each environment

The number of sentences 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 sentence, one of the 91 possible speakers is considered as author of the sentence 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 1 that noise produces an important degradation in the behavior of the system: from near 1%, EER falls down until 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 figure 3, 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.

In identification, which success rate results in % are in table 2, it can be observed that the use of noisy signal 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 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. With verification and identification results using PD-MEMLIN, it can be seen that the learnt transformations in order to project from noisy space to clean one are very general, loosing the speaker specificity. Since it is very important in verification and identification tasks, speaker clustering techniques can be used in order to define speaker-dependent transformations. In this sense, similar projections would be used for the same kind of speakers, the speaker

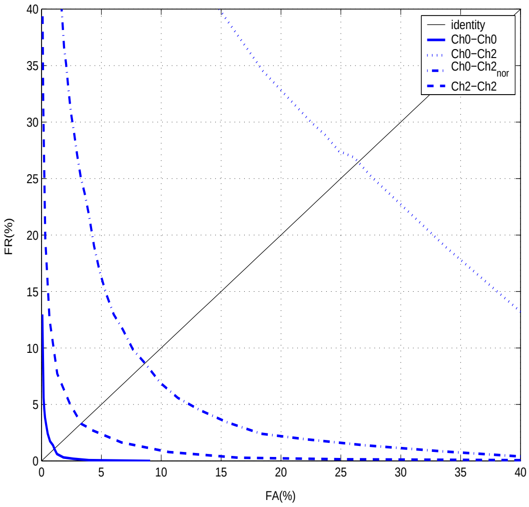


Figure 3: Total verification results with all environments and different thresholds.

specificity would not lose, and the results could be improved.

5. Conclusions

In this paper a stereo data based Cepstral normalization technique, PD-MEMLIN, has been presented in order to improve the verification and identification results with adverse and dynamic acoustic conditions. The algorithm learns different linear transformations for each pair of phoneme gaussians: one of clean space and the other for noisy space. In order to study the behavior of the technique, some experiments have been carried out with Spanish SpeechDat Car database. An UBM-GMM system is developed for verification, and a GMM system for identification. The results show that noise degrades seriously the accuracy of the systems. Pre-processing the noisy signal with PD-MEMLIN in verification, a mean 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.

6. References

[1] D. A. Reynold , T. F. Quatieri, R. B. Dunn ”Speaker verification using adapted gaussian mixture models”, in Digital Signal Processing, vol. 10, pp 19-41. 2000.

[2] R. M. Stern, and 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, pp. 33-42, Apr. 1997.

[3] P. Moreno, ”Speech Recognition in Noisy Environments”, Ph.D. Thesis, ECE Department, Carnegie-Mellon University. Apr. 1996.

[4] A. Acero, ”Acustical and environmental robustness in automatic speech recognition”, Ph.D. Thesis, ECE Department, Carnegie-Mellon University, Sep 1990.

[5] J. Droppo, L. Deng, A. Acero, ”Evaluation of the SPLICE algorithm on the Aurora2 database”, in Proc. Eurospeech, vol. 1, Sep 2001.

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

[7] A. Moreno, and A. Noguiera, and A. Sesma, ”SpeechDat-Car: Spanish”, Technical Report SpeechDat

[8] J. Bilmes, ”A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models”, Technical Report, University of Berkeley, ICSI-TR-97-021, 1997.

[9] 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, pp. 291-298, vol. 2,

No. 2, Apr, 1994.