Robust Speech Recognition in Cars using Phoneme Dependent Multi-Environment Linear Normalization

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 a Phoneme-Dependent Multi-Environment Models based LInear feature Normalization, PD-MEMLIN, is presented. The target of this algorithm is to learn the difference 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. Some experiments with SpeechDat Car database were carried out in order to study the behavior of the proposed techniques in a real acoustic environment. The experimental results show an average improvement of more than 77% using PD-MEMLIN, and more than 85% using transformed space acoustic models and multi-environment rotation transformation, concerning the baseline.

1. Introduction

When testing and training acoustic conditions are different, the accuracy of speech recognition 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, feature compensation, 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 [1], and they have proved to be effective.

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 and compared against SPLICE and MEMLIN.

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 treated, 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 used with MEMLIN and PD-MEMLIN algorithms.

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. The results for MEMLIN and PD-MEMLIN with SpeechDat Car database [6] are presented and discussed in Section 5. Finally, the conclusions are included in Section 6.

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 training process. The clean feature space is modelled as a mixture of gaussians for each phoneme. The noisy one 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.

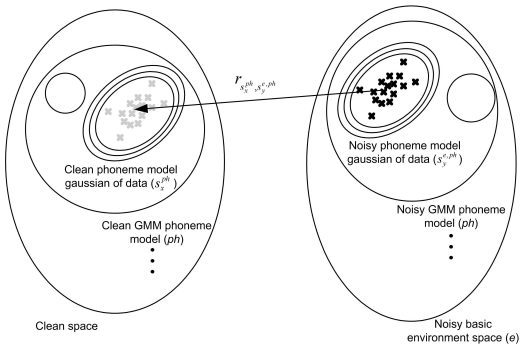

Figure 1: Scheme of PD-MEMLIN transformations for one environment.

This work has been supported by the national project TIC2002-04103-C03-01

2.1. Approximations

Three approximations are assumed: firstly, some basic environments are defined in the noisy space, and noisy feature vectors, y, follow the distribution of gaussians 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; $\mu_{s_y^{e,ph}}$, $\Sigma_{s_y^{e,ph}}$, and $p(s_y^{e,ph})$ are the mean vector, the diagonal covariance matrix, and the weight associated to $s_y^{e,ph}$.

Second, clean feature vectors, x, are modelled following the distribution of gaussians 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; $\mu_{s_x^{ph}}$, $\Sigma_{s_x^{ph}}$, and $p(y|s_x^{ph})$ are the mean, diagonal covariance matrix, and the weight associated to s_x^{ph}.

Third, for each time frame, t, x is approached as a function, Ψ, 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:

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

where $p(x|y_t)$ is the Probability Density Function (PDF) of x given y_t. Using 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 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 β 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, and 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. Principal Component Analysis (PCA) of the covariance matrixes of clean, and noisy feature vectors for each environment, ($\tilde{\Sigma}_e$, Σ_e, respectively) is used 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 $i = 1...D$, $\tilde{\lambda}_{e,1} \geq \tilde{\lambda}_{e,2} \geq ... \geq \tilde{\lambda}_{e,D}$, and $\lambda_{e,1} \geq \lambda_{e,2} \geq ... \geq \lambda_{e,D}$, and D is the dimension of the feature vectors. The rotation angle between the two principal directions of clean and noisy spaces is calculated 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:

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

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

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

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

	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.

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

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 mean 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 used 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 used 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.
- Normalization of noisy training data normalization.
- New acoustic models are trained with normalized noisy training data.

	MWER (%)	IMP (%)
PD-MEMLIN	5.30	77.67
PD-MEMLIN + rot	5.37	76.82
MEMLIN	6.06	72.24
MEMLIN + rot	5.65	76.32
SPLICE	7.57	57.92
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.

5. RESULTS

A set of experiments have been carried out using the Spanish SpeechDat Car database [6]. 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).

The task used is isolated and continuous digits. 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 are applied over the 12 MFCC and delta 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 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 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.

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 %.

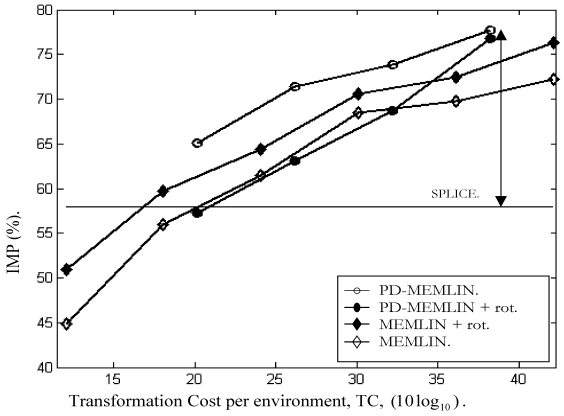

Figure 2: Improvement, in %, for different techniques, where rot indicates that multi-environment rotation transformation is used after normalization techniques.

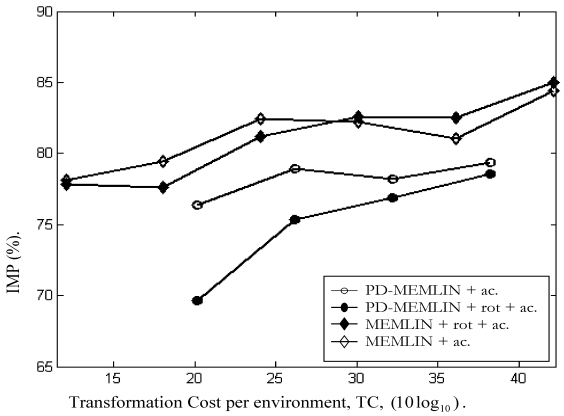

Figure 3: Improvement, in %, for different techniques, where rot and ac indicate that multi-environment rotation transformation or transformed space acoustic models are respectively used.

In order to compare the presented techniques, the transformation 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 gaussian 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 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. 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 that 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. 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: 85.02%.

6. CONCLUSIONS

In this paper we have presented a feature vector normalization, PD-MEMLIN, and two approaches in order to compensate the 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 with PD-MEMLIN (77.67%), better than other techniques as MEMLIN or SPLICE. When multi-environment rotation transformation and transformed space acoustic models are used with MEMLIN an improvement of 85.02% is obtained. Since it is not always available stereo data, in a future work, a non stereo data PD-MEMLIN approximation will be studied.

7. 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.