On the Jointly Unsupervised Feature Vector Normalization and Acoustic Model Compensation for Robust Speech Recognition

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

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

Abstract

To compensate the mismatch between training and testing conditions, an unsupervised hybrid compensation technique is proposed. It combines Multi-Environment Model based LInear Normalization (MEMLIN) with a novel acoustic model adaptation method based on rotation transformations. A set of rotation transformations is estimated between clean and MEMLIN-normalized data by linear regression in a training process. Thus, each MEMLIN-normalized frame is decoded using the expanded acoustic models, which are obtained from the reference ones and the set of rotation transformations. During the search algorithm, one of the rotation transformations is on-line selected for each frame according to the ML criterion in a modified Viterbi algorithm. Some experiments with Spanish SpeechDat Car database were carried out. MEMLIN over standard ETSI front-end parameters reaches 75.53% of mean improvement in WER, while the introduced hybrid solution goes up to 90.54%.

Index Terms: robust speech recognition, feature vector normalization, acoustic model adaptation.

1. Introduction

When training and testing acoustic conditions differ, the accuracy of speech recognition systems rapidly degrades. To compensate this mismatch, classic robustness techniques have been developed along the following two main lines of research: acoustic model adaptation methods, and feature vector normalization methods. In general, acoustic model adaptation methods produce better results when the transcriptions are available [1] because they can model the uncertainty caused by the noise statistics. However, these methods usually require more data and computing time than feature vector normalization methods do, which do not produce as good results but provide more on-line solutions. Hybrid techniques, which are the combination of a feature vector normalization method and an acoustic model adaptation method, also exist [2].

A previous work [3] shows that Multi-Environment Model-based LInear Normalization, MEMLIN, (an empirical feature vector normalization method based on stereo data and the MMSE estimator) is effective to compensate the effects of dynamic and adverse car conditions, improving the performance of techniques based on similar criterions, e.g. Stereo-based Piecewise Linear Compensation for Environments, SPLICE, [4]. However these techniques, which only use a bias vector transformation to compensate the noisy feature vectors, do not take into account all kinds of degradation.

On the other hand, classic acoustic model adaptation methods, e.g. Maximum Likelihood Linear Regression, MLLR, [5] take into account implicitly all kinds of degradations of the feature vectors by mapping the parameters of the reference acoustic models to the noisy space. However, the performance of these techniques degrades when the transcription of the adaptation data is not available (unsupervised methods) [6].

In this work we propose an on-line unsupervised hybrid solution which combines MEMLIN with a novel acoustic model adaptation method based on rotation transformations over an expanded HMM-state space. Hence, clean and MEMLIN-normalized spaces are modelled with GMMs and a set of rotation matrices is obtained, estimating one matrix for each pair of Gaussians (clean-normalized) with stereo normalized and clean data in a previous unsupervised training process using linear regression. In recognition, each MEMLIN-normalized feature vector is decoded with the expanded acoustic models, which are generated from the reference ones and the set of rotation matrices; so that one of the rotation matrices is selected during the search algorithm for each MEMLIN-normalized feature vector by using the ML criterion. Thus, shift and rotation degradations are compensated jointly in an unsupervised way.

This paper is organized as follows: In Section 2, the novel proposed hybrid compensation technique is presented. In Section 3, some considerations about MEMLIN are included. The rotation matrix estimation process is explained in Section 4. The on-line selection of the rotation matrix for each normalized feature vector in the decoding process is presented in Section 5. In Section 6, the results with Spanish SpeechDat Car database [7] are included. Finally, the conclusions are presented in Section 7.

2. Unsupervised Hybrid Compensation

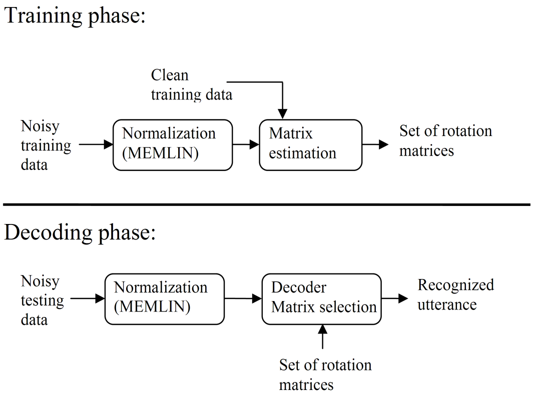

Figure 1: Scheme of the proposed unsupervised hybrid compensation technique.

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

The scheme of the proposed unsupervised hybrid compensation technique is depicted in Figure 1. It is composed of two phases: training and decoding. In the unsupervised training phase, the available clean and noisy training stereo data are needed to estimate the MEMLIN transformations [3]. Furthermore the noisy training feature vectors are normalized using MEMLIN ("Normalization MEMLIN"), and the clean and MEMLIN-normalized spaces are modelled with GMMs. Also, a set of rotation matrices is estimated by linear regression with the normalized and clean stereo training data ("Matrix estimation"), obtaining one rotation matrix for each pair of Gaussians (clean-normalized). On the other hand, in the decoding phase, each MEMLIN-normalized testing feature vector ("Normalization MEMLIN") is recognized with expanded acoustic models ("Decoder Matrix selection"), which are obtained with the reference acoustic models and the set of rotation matrices. During the search process, a rotation matrix per frame will be selected implicity with the associated expanded state by using the ML criterion in a modified Viterbi algorithm.

3. Feature Vector Normalization

MEMLIN is the selected feature vector normalization technique for the hybrid compensation method in this work, although other algorithms could be used. MEMLIN is an empirical feature vector normalization technique based on MMSE estimator. It is based on three approximations [3]: the clean feature space is modelled as a GMM; the noisy space is split into several basic acoustic environments and each one of them is modelled as a GMM. The third assumption consists on defining a bias vector transformation associated with each pair of Gaussians from the clean and the noisy basic environment spaces.

It can be observed that the clean estimated feature vector that MEMLIN provides for the time index t, $\hat{\mathbf{x}}_t$, is a shifted version of the noisy one $\mathbf{y}_t$: $\hat{\mathbf{x}}_t = \mathbf{y}_t + \mathbf{g}_t$, where $\mathbf{g}_t$ is the corresponding bias vector which depends on the acoustic environment and the noisy and clean GMM modelled spaces [3]. Usually $\hat{\mathbf{x}}_t$ is recognized with clean acoustic models. But it provides the same solution that recognizing the noisy feature vector, $\mathbf{y}_t$, with clean acoustic models, where all the mean vectors, μ, are modified as $\mu - \mathbf{g}_t$ (assuming the acoustic models are composed by HMM with GMMs as observation generation probability density functions, pdfs, of the different states). Thus, the feature vector normalization methods which consist on linear transformations composed only by a bias vector (e. g. Cepstral Mean Normalization (CMN), SPLICE, MEMLIN...) can be seen also as acoustic model adaptation techniques which transform the mean vectors each time index.

4. Rotation Matrix Estimation

Three approximations are considered

- Clean feature vectors, $\mathbf{x}_t$, are modelled using a GMM
$$p(\mathbf{x}_t) = \sum_{s_x} p(\mathbf{x}_t|s_x)p(s_x), \qquad (1)$$
$$p(\mathbf{x}_t|s_x) = \mathcal{N}(\mathbf{x}_t; \mu_{s_x}, \mathbf{\Sigma}_{s_x}), \qquad (2)$$
where μ_{s_x}, $\mathbf{\Sigma}_{s_x}$, and $p(s_x)$ are the mean vector, the diagonal covariance matrix, and the a priori probability associated with the clean model Gaussian s_x.

- Normalized feature vectors, $\hat{\mathbf{x}}_t$, are modelled using a GMM
$$p(\hat{\mathbf{x}}_t) = \sum_{s_{\hat{x}}} p(\hat{\mathbf{x}}_t|s_{\hat{x}})p(s_{\hat{x}}), \qquad (3)$$
$$p(\hat{\mathbf{x}}_t|s_{\hat{x}}) = \mathcal{N}(\hat{\mathbf{x}}_t; \mu_{s_{\hat{x}}}, \mathbf{\Sigma}_{s_{\hat{x}}}), \qquad (4)$$
where $\mu_{s_{\hat{x}}}$, $\mathbf{\Sigma}_{s_{\hat{x}}}$, and $p(s_{\hat{x}})$ are the mean vector, the diagonal covariance matrix, and the a priori probability associated with the normalized model Gaussian $s_{\hat{x}}$.

- Normalized feature vectors can be approximated as a linear function of the clean feature vectors which depends on the clean and normalized model Gaussians s_x and $s_{\hat{x}}$: $\hat{\mathbf{x}}_t \approx \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t$, where $\mathbf{A}_{s_x,s_{\hat{x}}}$ is the rotation matrix between the feature vectors $\hat{\mathbf{x}}_t$ and $\mathbf{x}_t$ associated to the pair of Gaussians s_x and $s_{\hat{x}}$.

Let us define the set of rotation matrices
$$\mathcal{A} = \{\mathbf{A}_{s_x,s_{\hat{x}}}\}_{s_x=1,s_{\hat{x}}=1}^{\#s_x,\#s_{\hat{x}}} = \{\mathbf{A}_n\}_{n=1}^{N}, \qquad (5)$$
where there is only one index n for each pair of Gaussians s_x, $s_{\hat{x}}$ and N denotes the pair of Gaussians number: $N = \#s_x \times \#s_{\hat{x}}$. In order to estimate the rotation matrix $\mathbf{A}_n$, stereo data is used in the previous training phase: $(\mathbf{X}^{Tr}, \hat{\mathbf{X}}^{Tr}) = \{(\mathbf{x}_1^{Tr}, \hat{\mathbf{x}}_1^{Tr}); ...; (\mathbf{x}_t^{Tr}, \hat{\mathbf{x}}_t^{Tr}); ...; (\mathbf{x}_T^{Tr}, \hat{\mathbf{x}}_T^{Tr})\}$, with $t \in [1, T]$, where $\hat{\mathbf{X}}^{Tr}$ is obtained applying the corresponding feature vector compensation technique (MEMLIN in this case) to the noisy training data $\mathbf{Y}^{Tr}$. Thus, $\mathbf{A}_n$ is estimated by minimizing the defined mean weighted square error, ξ_n, (6) with respect to $\mathbf{A}_n$ (7), where $Tra[\bullet]$ is the trace, $p(s_x|\mathbf{x}_t^{Tr})$ is the a posteriori probability of the clean model Gaussian s_x, given the clean training feature vector $\mathbf{x}_t^{Tr}$, and $p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr})$ is the a posteriori probability of the normalized model Gaussian $s_{\hat{x}}$, given the normalized training feature vector $\hat{\mathbf{x}}_t^{Tr}$. Both probabilities can be estimated using (1) and (2), for the first case (8), and (3) and (4) for the second one (9)
$$p(s_x|\mathbf{x}_t^{Tr}) = \frac{p(\mathbf{x}_t^{Tr}|s_x)p(s_x)}{\sum_{s_x} p(\mathbf{x}_t^{Tr}|s_x)p(s_x)}, \qquad (8)$$
$$p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr}) = \frac{p(\hat{\mathbf{x}}_t^{Tr}|s_{\hat{x}})p(s_{\hat{x}})}{\sum_{s_{\hat{x}}} p(\hat{\mathbf{x}}_t^{Tr}|s_{\hat{x}})p(s_{\hat{x}})}. \qquad (9)$$

5. Rotation Matrix Selection in Decoding

In order to select the rotation matrix, $\mathbf{A}_t$, associated to each normalized testing feature vector, $\hat{\mathbf{x}}_t$, from the set of estimated rotation matrices, $\mathbf{A}_n$, ML maximization criterion is applied in the decoding process. To do that, the acoustic models are modified in a similar way as described in [8], where the set of linear transformations in this case are the matrices $\mathbf{A}_n$ previously estimated. Hence, each state of the clean space HMM acoustic models, ($q \in [1, Q]$), is expanded into N states (q, n) considering the linear approximation $\hat{\mathbf{x}}_t \approx \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t = \mathbf{A}_n\mathbf{x}_t$. The goal of the state expansion is to reduce the mismatch between the clean space acoustic models and the normalized feature vectors for each rotation transformation. Thus, each expanded state is specialized in one of the rotation transformations previously estimated. Assuming that a component s_q in the pdf mixture of the original state q follows a normal distribution: $\mathcal{N}(\mathbf{x}_t; \mu_{s_q}, \mathbf{\Sigma}_{s_q})$, the corresponding expanded state component $s_{q,n}$ is assumed to follow the distribution $\mathcal{N}(\hat{\mathbf{x}}_t; \mathbf{A}_n\mu_{s_q}, \mathbf{A}_n\mathbf{\Sigma}_{s_q}\mathbf{A}_n^T)$. So, the pdf for the expanded state (q, n), $p(\hat{\mathbf{x}}_t|q, n)$, is a GMM composed by the defined expanded components where the a priori component weights remain unaltered: $p(s_{q,n}) = p(s_q)$.

$$p(\hat{\mathbf{x}}_t|q, n) = \sum_{s_q} p(s_q)\mathcal{N}(\hat{\mathbf{x}}_t; \mathbf{A}_n\mu_{s_q}, \mathbf{A}_n\mathbf{\Sigma}_{s_q}\mathbf{A}_n^T). \qquad (10)$$

$$\xi_n = \frac{1}{T}\sum_t p(s_x|\mathbf{x}_t^{Tr})p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr}) \cdot Tra\left[(\hat{\mathbf{x}}_t^{Tr} - \mathbf{A}_n\mathbf{x}_t^{Tr})(\hat{\mathbf{x}}_t^{Tr} - \mathbf{A}_n\mathbf{x}_t^{Tr})^T\right]. \tag{6}$$

$$\mathbf{A}_n = \mathbf{A}_{s_x,s_{\hat{x}}} = \underset{\mathbf{A}_n}{arg\,min}\{\xi_n\} = \left[\sum_t p(s_x|\mathbf{x}_t^{Tr})p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr})(\hat{\mathbf{x}}_t^{Tr}(\mathbf{x}_t^{Tr})^T)\right] \cdot \left[\sum_t p(s_x|\mathbf{x}_t^{Tr})p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr})(\mathbf{x}_t^{Tr}(\mathbf{x}_t^{Tr})^T)\right]^{-1}. \tag{7}$$

Note that the proposed expanded acoustic models, from a generative point of view, can be seen as a more flexible speech production process in adverse environment conditions, since they can generate sequences of rotated feature vectors more suitable to the normalized space.

Once the clean acoustic models have been expanded, the classic search algorithm (Viterbi) for decoding unlabeled sequences has to be modified. Given a normalized testing utterance, the sequence of expanded states which maximizes the likelihood determines implicity the rotation matrix $\mathbf{A}_t$ for each normalized feature vector. Thus, the search algorithm under this framework can be performed by computing recursively the score state variable, $\phi_{q,n}(t)$, for the state (q, n) and the time index t

$$\phi_{q,n}(t) = \max_{q',n'}\{\phi_{q',n'}(t-1) \cdot \pi_{q',n',q,n} \cdot p(\hat{\mathbf{x}}_t|q,n)\}, \tag{11}$$

being $\pi_{q',n',q,n}$ the transition probability from expanded state (q', n') to (q, n), which is considered equiprobability for all (q, n) in this work. It can be observed that the presented hybrid solution can be seen as recognizing each MEMLIN-normalized feature vector, $\hat{\mathbf{x}}_t = \mathbf{y}_t + \mathbf{g}_t$, with the corresponding expanded acoustic models, where the mean vectors and covariance matrices are adapted as $\mathbf{A}_t\mu$ and $\mathbf{A}_t\mathbf{\Sigma}\mathbf{A}_t^T$. This solution provides the same results that recognizing the noisy feature vector, $\mathbf{y}_t$, with acoustic models where the mean vectors and covariance matrices are: $\mathbf{A}_t\mu - \mathbf{g}_t$ and $\mathbf{A}_t\mathbf{\Sigma}\mathbf{A}_t^T$, respectively. Note that this point of view is conceptually similar to MLLR, where shift and rotation are included in acoustic models. However, the shift and rotation transformations for the proposed hybrid technique, which are selected for each feature vector, are estimated with a different criterion than MLLR. Also, the unsupervised MLLR version needs a previous step to provide an estimation of the transcription of the adaptation data (usually a recognition process). Thus, the performance of the unsupervised MLLR solution can degrade dramatically when the adaptation data are highly noise corrupted or the adaptation data tasks are complex (e.g. large vocabulary, spontaneous speech...) so that the estimation of the transcription would not be precise enough. These problems do not affect to the proposed hybrid technique, which does not precise the transcription of the adaptation data.

6. Empirical Results

To observe the performance of the proposed unsupervised hybrid compensation technique 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 (Shune 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.

The recognition task is isolated and continuous digits recognition. As feature set, the standard ETSI front-end features plus the energy and the corresponding delta and delta delta coefficients are used in all the experiments. Cepstral mean normalization is applied to testing and training data in all cases. On the other hand, in this work, MEMLIN and SPLICE with Environmental Model selection (SPLICE EM) [4] are applied to the 12 MFCCs and energy, whereas the derivatives are computed over the normalized static coefficients. The acoustic models are composed of 16 state HMM for each digit, a 3 state begin-end silence HMM and a 1 state inter-word silence HMM. In all cases, each pdf state is composed by a mixture of three Gaussians.

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 utterances in each basic environment. "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 models are retrained for each basic environment. All acoustic models are obtained with ML algorithm. "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 and ML algorithm (Train HF), MWER decreases, 4.63%. Finally, the most competitive results (3.42% MWER) are obtained when specific acoustic models are retrained for each basic environment with ML algorithm, (Train † HF), despite the poor WER reached with E7 due to the reduced amount of data for that condition. However, this option is not possible in a real situation because the basic environment can not be known for each testing utterance.

Figure 2 shows the mean improvement in WER (MIMP) in % for SPLICE EM, MEMLIN and the proposed hybrid technique based on MEMLIN (MEMLIN A) when different number of Gaussians per basic environment are considered for the feature vector normalization techniques (4, 8, 16, 32, 64 and 128). In case of MEMLIN, clean feature space is modelled with the same number of Gaussians than the basic environments. Also, 16 rotation matrices (N) are estimated in all cases for MEMLIN A ($\#s_x = \#s_{\hat{x}} = 4$). MIMP is computed with MWER as

$$MIMP = \frac{100(MWER - MWER_{CLK-HF})}{MWER_{CLK-CLK} - MWER_{CLK-HF}}, \tag{12}$$

where $MWER_{CLK-CLK}$ is the mean WER obtained with clean conditions (0.91 in this case), and $MWER_{CLK-HF}$ is

Train	Test	E1	E2	E3	E4	E5	E6	E7	MWER (%)
CLK	CLK	0.95	2.32	0.70	0.25	0.57	0.32	0.00	0.91
CLK	HF	3.05	13.29	15.52	27.32	31.36	35.56	53.06	21.48
HF	HF	3.81	6.86	3.50	3.76	4.96	4.44	3.06	4.63
† HF	HF	1.14	4.37	1.68	2.13	2.10	2.06	23.13	3.42
HF MLLR	HF	1.33	4.55	2.52	3.63	7.34	5.24	26.19	5.28
CLK A	HF MEMLIN	2.19	3.95	2.10	3.26	3.24	1.90	2.38	2.86

Table 1: WER baseline results, in %, from the different basic environments (E1,..., E7), where MWER is the Mean WER.

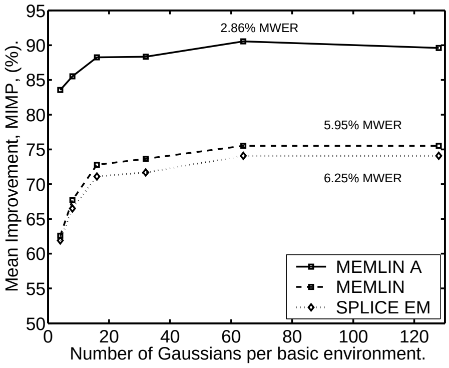

Figure 2: Mean improvement in WER, MIMP, in % for different normalization techniques: SPLICE with environmental model selection (SPLICE EM), MEMLIN and the proposed hybrid technique based on MEMLIN and acoustic model adaptation based on rotation transformations (MEMLIN A).

the baseline (21.48). So, A 100% MIMP would be achieved when MWER equals the one obtained under clean conditions.

It can be verified in Figure 2 the important improvement that the presented hybrid solution obtains when it is applied over MEMLIN for any number of Gaussians per basic environment concerning SPLICE EM and MEMLIN. In fact, the performance with 64 components per basic environment (90.54% MIMP, 2.86% MWER) is significantly better than SPLICE EM (74.08% MIMP, 6.25% MWER) and MEMLIN (75.53% MIMP, 5.95% MWER); even if matched training condition (81.93% MIMP, 4.63% MWER) or specific acoustic models for each basic environment are considered (87.81% MIMP, 3.42% MWER), the performance is slightly inferior with respect to the proposed hybrid solution due to the noisy space is more heterogenous than the normalized one. The complete best WER results obtained with the hybrid solution are also included in Table 1 (Train CLK A, Test HF MEMLIN). Also the performance of unsupervised MLLR, where the transcription of the noisy data is assumed as the true one, is presented in Table 1 to complete the comparison (Train HF MLLR, Test HF). Note that the obtained performance (MWER 5.28%, 78.77% MIMP) is inferior than the match training condition results and the ones obtained with the proposed hybrid technique.

7. Conclusions

In this paper we have presented an unsupervised on-line hybrid compensation solution which combines Multi-Environment Model based LInear Normalization (MEMLIN) with a novel acoustic model adaptation technique based on rotation transformations which depend on GMMs. The purpose of the hybrid solution is to compensate jointly the shift and rotation introduced by the acoustic environment. Some results with Spanish SpeechDat Car database show the effective performance of the proposed technique (90.54% of mean improvement with 64 Gaussians per basic environment) with respect to classic feature vector normalization techniques: SPLICE EM (74.08%), and MEMLIN (75.53%), or acoustic model adaptation techniques: unsupervised MLLR (78.77%). Even match training condition, which is a supervised solution, does not reach the performance of the proposed technique (81.93%) due to the variability of the noisy space, which is higher than the normalized space. As future line we propose to use the hybrid solution with other feature vector normalization techniques.

8. References

[1] L. Neumeyer and M. Weintraub, "Robust Speech Recognition in Noise Using Adaptation and Mapping Techniques," in *Proceedings of ICASSP* , vol. 1, 1995, pp. 141–144.

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

[3] L. Buera, E. Lleida, A. Miguel, A. Ortega, and O. Saz, "Cepstral vector normalization based on stereo data for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 15, pp. 1098–1113, March 2007.

[4] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the AURORA2 database," in *Eurospeech*, 2001.

[5] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continous-density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[6] M. Padmanabhan, G. Saon, and G. Zweig, "Lattice-based unsupervised mllr for speaker adaptation," in *ASR*, vol. 2, 2000, pp. 128–132.

[7] H. van den Heuvel, J. Boudy, R. Comeyne, S. Euler, A. Moreno, and G. Richard, "The speechdat-car multilingual speech databases for in-car applications: some first validation results," in *Eurospeech*, 1999.

[8] A. Miguel, E. Lleida, A. Juan, L. Buera, A. Ortega, and O. Saz, "Local transformation models for speech recognition," in *ICSLP*, Pittsburgh, USA, 2006.