FALA 2010
VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop

A feature compensation approach using VQ-based MMSE estimation for robust speech recognition

José A. González, Antonio M. Peinado, Angel M. Gómez, José L. Carmona, and Juan A. Morales-Cordovilla

Dpto. de Teoría de la Señal, Telemática y Comunicaciones, University of Granada, Spain

{joseangl,amp,amgg,maqueda,jamc}@ugr.es

Abstract

We describe a novel feature compensation algorithm based on the minimum mean square error (MMSE) estimation and stereo training data for robust speech recognition. The proposed algorithm can be viewed as a piece-wise linear transformation between the noisy and clean feature spaces, where both spaces are modeled by means of vector quantization (VQ) codebooks. By means of this VQ modeling, we show that a very efficient estimator can be obtained in terms of computational cost and recognition accuracy. Also, two approaches are proposed in order to compensate the acoustic noise distortion. First, we propose a novel formulation for the normalization of noisy feature vectors. Second, a novel subregion-based modeling is applied to obtain a better representation of the differences between noisy and clean domains. The experimental results on noisy digit recognition show a relative improvement of 61.49% over the baseline when clean acoustic models are used. Furthermore, important improvements are achieved in comparison with other similar approaches.

Index Terms: robust speech recognition, feature compensation, MMSE estimation, stereo data

1. Introduction

It is well known that the performance of automatic speech recognition (ASR) systems degrades as the mismatch between testing and training conditions increases. Thus, there are several sources of mismatch that directly affect to the ASR performance, such as variety of speakers, accents, channels and noise conditions [1]. Many algorithms have been developed to compensate this mismatch. These algorithms are usually grouped into two categories [3]: feature-based and model-based approaches. Feature-based techniques focus on modifying or enhancing the feature vectors to be closer to the clean training condition or to be less sensitive to the variability introduced by the aforementioned sources of mismatch. On the other hand, model-based approaches adapt the acoustic model parameters to the testing conditions. These approaches often yield better performance than feature-based ones, especially in low SNR conditions. Nevertheless, feature-based techniques have the advantage that can be seamlessly implemented into existing systems, since only a module that pre-process the feature vectors before they are fed into the speech recognizer is needed. In addition, feature compensation is usually less computationally expensive, especially if the acoustic environment is rapidly changing.

Stereo data are widely used in order to achieve noise robustness in ASR systems. In this way, a stereo database including both clean and noisy features can be used to learn the statistical relationship between both domains. The earliest approach based on stereo data was proposed in [3] with the SNR-Dependent Cepstral Normalization (SDCN) and Codeword-Dependent Cepstral Normalization (CDCN). Since then, more sophisticated techniques have appeared, as multivariate Gaussian based cepstral normalization algorithm (RATZ) [4], Stereo based Piecewise LInear Compensation for Environments (SPLICE) [5], Multi-Environment Models based LInear Normalization (MEMLIN) [6] and Stereo-based Stochastic Mapping (SSM) [7]. The later techniques are based on a Minimum Mean Squared Error (MMSE) estimation, where the clean and/or noisy domains are represented by means of Gaussian Mixture Models (GMMs).

In this paper we are also interested in MMSE estimation for feature compensation, although a different approach is followed to represent the clean and noisy domains. Thus, instead of modeling the clean and noisy feature spaces with GMMs, we characterize each of these spaces with a set of cells obtained by means of vector quantization (VQ). As it will be shown, VQ quantization provides much more efficient compensation techniques, but their results are known to be inferior, due to the hard decision involved (a cell is represented by a centroid instead of a probability function). For this reason, in this paper we present a novel MMSE formulation which can cope with this disadvantage. In addition, we show that the recognition accuracy can be significantly improved by considering that every VQ cell contains a set of overlapping subregions with provide a more accurate mapping between the clean and noisy spaces.

This paper is organized as follows. In Section 2, the mathematical formulation for the proposed VQ-based MMSE estimation is derived. The experimental framework is described in Section 3 while the results are presented and discussed in Section 4. Finally, Section 5 presents the conclusions and some directions for future work.

2. Derivation of the proposed MMSE estimator

We denote by $\boldsymbol{y}$ the observed feature vector representation of a noisy speech segment distorted by acoustic noise and by $\boldsymbol{x}$ its corresponding unknown clean version obtained when the segment is not affected by noise. In this work, we seek for a compensation function that provides an estimate of $\boldsymbol{x}$ given $\boldsymbol{y}$, i.e., $\hat{\boldsymbol{x}} = f(\boldsymbol{y})$. Among others, a plausible option to derive this function is by means of the MMSE criterion. In this case, the estimate of clean speech is given by

$$\hat{\boldsymbol{x}} = E[\boldsymbol{x}|\boldsymbol{y}] = \int_{\boldsymbol{x}} \boldsymbol{x} \cdot p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} \qquad (1)$$

where $p(\boldsymbol{x}|\boldsymbol{y})$ is the conditional probability of $\boldsymbol{x}$ given $\boldsymbol{y}$. Different approaches have been proposed to model this distribution. For example, RATZ [4] models the clean feature space by means of a GMM and it assumes an additive effect of the noise on the MFCC domain. On the other hand, SPLICE [5] models the distorted feature space and, as RATZ, also an additive effect of the noise is assumed. A more complex modeling is applied by SSM [7] in which the conditional distribution is derived from the joint distribution of clean and noisy feature vectors $p(\boldsymbol{x}, \boldsymbol{y})$. In this work, however, we follow a different approach. We assume that the clean and noisy feature spaces can be independently represented by means of probability density function (pdf) mixtures in the following way,

$$p(\boldsymbol{x}) = \sum_{k_x} p(\boldsymbol{x}|k_x)P(k_x) \tag{2}$$

$$p(\boldsymbol{y}) = \sum_{k_y} p(\boldsymbol{y}|k_y)P(k_y) \tag{3}$$

where k_x and k_y are components (e.g., Gaussian pdfs) of the mixtures that model the clean and noisy spaces, respectively.

Using the previous models, the conditional probability $p(\boldsymbol{x}|\boldsymbol{y})$ can be expressed as,

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}) &= \sum_{k_x}\sum_{k_y} p\left(\boldsymbol{x}, k_x, k_y|\boldsymbol{y}\right) \\ &= \sum_{k_x}\sum_{k_y} p\left(\boldsymbol{x}|k_x, k_y, \boldsymbol{y}\right) p\left(k_x|k_y, \boldsymbol{y}\right) p\left(k_y|\boldsymbol{y}\right) \end{aligned} \tag{4}$$

Finally, applying (4) to (1), the MMSE estimation takes the following form,

$$\hat{\boldsymbol{x}} = \sum_{k_x}\sum_{k_y} E\left[\boldsymbol{x}|k_x, k_y, \boldsymbol{y}\right] P\left(k_x|k_y, \boldsymbol{y}\right) P\left(k_y|\boldsymbol{y}\right) \tag{5}$$

where $P\left(k_y|\boldsymbol{y}\right)$ and $P\left(k_x|k_y, \boldsymbol{y}\right)$ are obtained using the marginal distributions of eqns. (2)-(3) and stereo training data. In contrast to other methods, such as MEMLIN [6], where these two distributions are modeled by means of GMMs, we propose the use of vector quantization (VQ) codebooks. In this way, every feature vector space is modeled by means of a VQ codebook that partitions its space into a set of disjoint cells. We will notate $\{C_X^{(i)}\ (i = 1, \ldots, M)\}$ as the set of cells corresponding to the clean feature space X and $\{C_Y^{(j)}\ (j = 1, \ldots, N)\}$ as the cells of the noisy space Y. These cells will hereinafter play the role of pdfs k_x and k_y in eqn. (5).

The VQ codebook of the noisy feature space can be used now to compute the *a posteriori* probability $P(k_y|\boldsymbol{y})$ in eqn. (5) as,

$$P\left(C_Y^{(j)}\middle|\,\boldsymbol{y}\right) = \begin{cases} 1 & C_Y^{(j)} = C_Y^* \\ 0 & otherwise \end{cases} \tag{6}$$

where $C_Y^* \equiv C_Y^*(\boldsymbol{y})$ is the cell that contains the input feature vector $\boldsymbol{y}$ according to the following distance,

$$C_Y^*(\boldsymbol{y}) = \underset{j}{\operatorname{argmin}} \left\{ (\boldsymbol{\mu}_Y^{(j)} - \boldsymbol{y})^T \operatorname{diag}(\boldsymbol{\Sigma}_Y^{(j)})^{-1} (\boldsymbol{\mu}_Y^{(j)} - \boldsymbol{y}) \right\} \tag{7}$$

where $\operatorname{diag}(\cdot)$ returns a diagonal matrix with the elements of the main diagonal of the input matrix, and $\boldsymbol{\mu}_Y^{(j)}$ and $\boldsymbol{\Sigma}_Y^{(j)}$ are the mean vector (centroid) and covariance matrix of $C_Y^{(j)}$.

Applying (6) to (5), the MMSE estimation can be rewritten as,

$$\begin{aligned} \hat{\boldsymbol{x}} &= \sum_{i=1}^{M}\sum_{j=1}^{N} E[\boldsymbol{x}|C_X^{(i)}, C_Y^{(j)}, \boldsymbol{y}] P(C_X^{(i)}|C_Y^{(j)}, \boldsymbol{y}) P(C_Y^{(j)}|\boldsymbol{y}) \\ &\approx \sum_{i=1}^{M} E\left[\boldsymbol{x}\middle|C_X^{(i)}, C_Y^*, \boldsymbol{y}\right] P\left(C_X^{(i)}\middle|\,C_Y^*\right) \end{aligned} \tag{8}$$

We will refer to this estimation as VQ-based MMSE estimation (VQ-MMSE). As can be observed, the conditional probability $P(k_x|k_y, \boldsymbol{y})$ of eqn. (5) is simplified to $P(\,C_X^{(i)}\Big|\,C_Y^*)$ in VQ-MMSE. This probability can be estimated using stereo data. This simplification, along with the application of VQ for the computation of the noisy component posterior in eqn. (6), leads to a very efficient implementation of the MMSE estimation. It is important to note that the original input vector $\boldsymbol{y}$ remains in the expected value $E[\boldsymbol{x}|C_X^{(i)}, C_Y^*, \boldsymbol{y}]$ of eqn. (8) in spite of the VQ modeling. That is, the VQ modeling is applied to compute the probabilities required by the MMSE estimation, but it does not necessarily involve a quantization of the input that could lead to a performance reduction.

The term $E[\boldsymbol{x}|C_X^{(i)}, C_Y^*, \boldsymbol{y}]$ in eqn. (8) defines the transformation of feature vectors between cells $C_X^{(i)}$ and C_Y^* due to acoustic noise. In order to accurately model this transformation, we introduce in the following the concept of subregion of a VQ cell. We will consider that every clean cell $C_X^{(i)}$ is composed by a set of subregions $\{C_X^{(i,j)}, (j = 1, \ldots, N)\}$, where $C_X^{(i,j)}$ represents all the clean feature vectors whose corresponding distorted ones belong to the noisy cell $C_Y^{(j)}$. Similarly, $C_Y^{(i,j)}$ represents the subregion of $C_Y^{(j)}$ where the feature vectors of $C_X^{(i)}$, once transformed by noise, are mapped. It is interesting to point out that this refined modeling can also be seen as a cross-modeling between the clean and noisy domains. Thus, a subregion defined in a given feature space can be considered as a part of the projection of a cell defined in the other space.

In order to compensate the noise distortion, we propose to apply a linear transformation to every feature vector. To do so, we assume that the subregions in the clean and noisy feature spaces are Gaussian distributed, i.e., $C_X^{(i,j)} \sim \mathcal{N}(\boldsymbol{\mu}_X^{(i,j)}, \boldsymbol{\Sigma}_X^{(i,j)})$ and $C_Y^{(i,j)} \sim \mathcal{N}(\boldsymbol{\mu}_Y^{(i,j)}, \boldsymbol{\Sigma}_Y^{(i,j)})$, where $\boldsymbol{\mu}_X^{(i,j)}$,$\boldsymbol{\mu}_Y^{(i,j)}$ are the mean vectors and $\boldsymbol{\Sigma}_X^{(i,j)}$,$\boldsymbol{\Sigma}_Y^{(i,j)}$ the corresponding covariance matrices. Then, the proposed transformation takes the following form,

$$E\left[\boldsymbol{x}\middle|C_X^{(i)}, C_Y^{(j)}, \boldsymbol{y}\right] = \boldsymbol{A}^{(i,j)}\boldsymbol{y} + \boldsymbol{b}^{(i,j)} \tag{9}$$

where $\boldsymbol{A}^{(i,j)}$ and $\boldsymbol{b}^{(i,j)}$ are computed in order to eqn. (9) firstly normalizes the noisy feature vectors regarding the mean and covariance of the noisy subregion, and then transforms them to the clean domain. Thus, eqn. (9) can be seen as a whitening and mapping transformation whose parameters are computed as,

$$\boldsymbol{A}^{(i,j)} = \left(\boldsymbol{\Sigma}_X^{(i,j)}\right)^{1/2} \left(\boldsymbol{\Sigma}_Y^{(i,j)}\right)^{-1/2} \tag{10}$$

$$\boldsymbol{b}^{(i,j)} = \boldsymbol{\mu}_X^{(i,j)} - \left(\boldsymbol{\Sigma}_X^{(i,j)}\right)^{1/2} \left(\boldsymbol{\Sigma}_Y^{(i,j)}\right)^{-1/2} \boldsymbol{\mu}_Y^{(i,j)} \tag{11}$$

where these terms can be precomputed offline for every pair of cells $(C_X^{(i)}, C_Y^{(j)})$. In addition, the mean vectors and covariance matrices can be easily computed from a stereo database using the feature vectors assigned to each subregion.

Finally, the proposed VQ-MMSE estimation in (8) becomes

$$\begin{aligned}\hat{\boldsymbol{x}} &= \sum_{i=1}^{M} E\left[\boldsymbol{x}\left|C_X^{(i)}, C_Y^*, \boldsymbol{y}\right.\right] P\left(C_X^{(i)}\middle| C_Y^*\right) \\ &= \sum_{i=1}^{M}\left(\boldsymbol{A}^{(i,*)}\boldsymbol{y} + \boldsymbol{b}^{(i,*)}\right) P\left(C_X^{(i)}\middle| C_Y^*\right) \\ &= \underbrace{\left(\sum_{i=1}^{M} P\left(C_X^{(i)}\middle| C_Y^*\right)\boldsymbol{A}^{(i,*)}\right)}_{\boldsymbol{A}^*}\boldsymbol{y} + \underbrace{\sum_{i=1}^{M} P\left(C_X^{(i)}\middle| C_Y^*\right)\boldsymbol{b}^{(i,*)}}_{\boldsymbol{b}^*} \\ &= \boldsymbol{A}^*\boldsymbol{y} + \boldsymbol{b}^* \end{aligned} \quad (12)$$

where $\boldsymbol{A}^*$ and $\boldsymbol{b}^*$ can be precomputed offline. Thus, we can see that the proposed compensation depends only on the noisy cell C_Y^* which the input feature vector $\boldsymbol{y}$ belongs to.

3. Experimental framework

Experiments are performed under the framework proposed by ETSI STQ-Aurora working group using the Aurora-2 database [8]. This database consists of utterances of connected digits spoken by American English speakers. For our purposes, we have extracted the speech data from the clean training set and the clean utterances from the test *set A* of this database. The European Telecommunication Standards Institute front-end (ETSI FE, ES 201 108) [9] is used in this work. It provides a 13-dimension feature vector containing 12 Mel-Frequency Cepstral Coefficients (MFCCs) (the 0th order one is discarded), plus the log-energy feature. The recognizer is the one provided by Aurora-2 using whole word acoustic models trained on clean speech. Each digit is modeled by means of a 16-state continuous HMM with 3 Gaussians per state. On the other hand, the silence and short pause models are modeled by means of HMMs with 3 and 1 states, respectively, and 6 Gaussians per state.

The speech features extracted by ETSI FE are directly processed by VQ-MMSE. After the compensation, the dynamic speech features are computed. VQ codebooks are trained for every available acoustic condition using a k-means algorithm which applies the weighted Euclidean distance defined in eqn. (7). Through this set of codebooks, the compensation parameters are estimated for the proposed technique using stereo data. These compensation parameters account for the possible transformations due to acoustic noise between the clean feature space and the noisy one, both modeled by means of VQ codebooks with the same number of cells. In order to compare our proposal with other MMSE-based estimators, GMMs are also trained. Thus, one GMM with diagonal covariance matrices is estimated for every available training condition using the Expectation-Maximization (EM) algorithm.

A set of 9 acoustic noises is used for training purposes, namely: airport, highway, babble, bar, beach, pedestrian street, restaurant, street, and train station. Every noise recording is split into two parts: two-thirds are employed to train the proposed MMSE estimator while the remaining third is reserved for testing. The training part is added to the *clean* training set of Aurora-2 at 6 different SNRs (20, 15, 10, 5, 0, and -5 dB), resulting in 54 environmental noisy training conditions plus a clean condition (55 training conditions in total). In order to evaluate the performance of our proposal, two different test sets are defined. The first set, called *Set A*, is intended to show the performance of the different techniques when considering the same environments used for training. Thus, 55 testing conditions are defined by artificially contaminating the clean test set of Aurora-2 with the testing part of the noises. The second set, called *Set B*, is created in the same way, but using five new different noises (pedestrian square, car, bus station, heavy sea, and heavy traffic avenue) at 5 new different SNRs (17.5, 12.5, 7.5, 2.5, and -2.5 dB). Thus, we can evaluate the influence of considering different environments that the ones used for training.

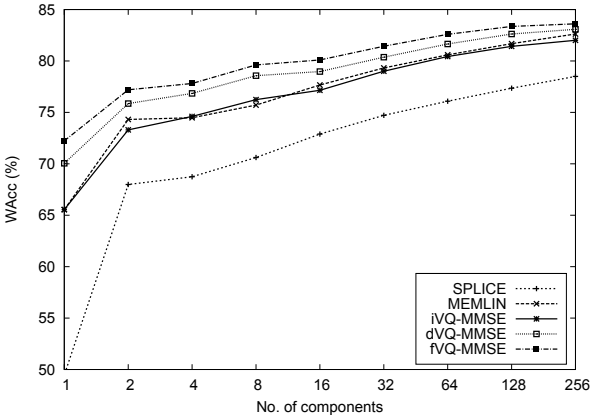

Figure 1: Oracle results for different feature compensation algorithms based on MMSE regarding the number of components (Gaussians or VQ cells) used.

4. Results

In the first part of this section, we give results for the aforementioned digit recognition task in non-mismatch conditions. Later, we provide results when the proposed technique is tested for unknown acoustic noises.

4.1. Oracle experiments

Fig. 1 shows the average word accuracy (WAcc), in percent, achieved by different estimators for Set A. For these experiments, oracle information about the acoustic noise is assumed, i.e., each utterance is compensated using a set of compensation parameters trained under the same noise. It must be pointed out that this information is not available in practice. However, the oracle results provide an estimate of the best performance that could be expected from every technique.

The baseline system applies acoustic models trained with clean speech and no compensation. This configuration achieves a WAcc of 50.83%. Three different versions of the proposed VQ-MMSE estimation are evaluated: iVQ-MMSE, dVQ-MMSE, and fVQ-MMSE. These versions assume identity, diagonal, and full covariance matrices, respectively, for the computation of the expected value in eqn. (9). Also, two well-known MMSE estimators using GMMs are considered: SPLICE [5] and MEMLIN [6]. As can be seen, our 3 proposals greatly improve the results achieved by the baseline system and SPLICE for all GMM and VQ sizes. This improvement shows the benefits of modeling both feature spaces, clean and noisy, instead of only one space such as SPLICE. Thus, the transformation applied by our approach is more accurate than in SPLICE. MEMLIN achieves a performance slightly better than iVQ-MMSE (82.62% vs. 82.02% for 256-component codebooks) . In fact, both techniques are quite similar, although our approach is more computationally efficient. Further improvements can be ob-

	Set A	Set B	Avg.	Imp.
Baseline	50.83	40.28	45.56	–
SPLICE	72.99	59.45	66.22	45.35
MEMLIN	77.21	62.89	70.05	53.75
iVQ-MMSE	77.29	65.44	71.37	56.64
dVQ-MMSE	79.04	67.18	73.11	60.47
fVQ-MMSE	79.54	67.61	**73.57**	61.49

Table 1: Average Word Accuracy (%) achieved by SPLICE, MEMLIN, and VQ-MMSE in the soft-compensation experiments for Set A and Set B (256-components codebooks).

tained when a more complex mapping is applied. This is the case of dVQ-MMSE and fVQ-MMSE, which compensate the shifts and scales in the feature domain due to environmental noise. In this case, our approaches obtain better results than MEMLIN.

4.2. Soft-compensation experiments

The proposed techniques are also evaluated in a more realistic scenario in which the acoustic noise that distorts the speech is unknown. In such a scenario, the clean feature vector estimate is obtained by means of a soft-compensation approach [10]. Thus, an estimate $\hat{\boldsymbol{x}}_e$ is obtained for every possible environmental condition e. The final estimate is computed as a linear combination of the estimates obtained for all environments. To do so, GMMs trained on every environmental condition are employed as environment classifiers to obtain the required probabilities $P(e|\boldsymbol{y})$. It must be pointed out that these GMMs are the same as those employed by SPLICE and MEMLIN, although a more sophisticated environment modeling could be applied.

Table 1 shows the recognition results achieved in the soft-compensation experiments for Set A and Set B when codebooks (GMMs or VQ codebooks) with 256 components are used. The average word accuracy (Avg.) and the relative improvement over the baseline in percent (Imp.) are also shown. As can be seen, all techniques suffer a performance degradation regarding the oracle experiments for Set A. This degradation is produced by mismatches in the environment identification. Furthermore, all methods yield poorer results for Set B. This is one of the lacks of the soft-compensation approach: the performance drops in mismatch situations. Nevertheless, these results demonstrate again the superior performance of our proposal. Thus, fVQ-MMSE achieves relative improvements of 11.10% and 5.03% in comparison with SPLICE and MEMLIN, respectively. Furthermore, now iVQ-MMSE outperforms MEMLIN.

5. Conclusions

In this paper, we have presented a novel feature compensation technique based on MMSE estimation and stereo training data for robust speech recognition. As a result, a piece-wise linear function between the noisy feature space and the clean one is obtained. We show that the application of VQ codebooks for the modeling of the feature spaces allows an efficient implementation of the MMSE estimator. Also, a novel subregion modeling is applied in order to accurately represent the acoustic noise distortion.

Two sets of experiments are carried out. Firstly, oracle experiments are conducted to obtain an upper bound of the performance that could be expected under non-mismatch. Secondly, the proposed techniques are also tested with unknown noises. In these experiments, we follow a soft-compensation approach in which the clean feature vector estimate is obtained as a linear combination of the estimates obtained for several defined environments. A relative improvement of 61.49% regarding the baseline is achieved for these experiments. Furthermore, relative improvements of 11.10% and 5.03% are obtained in comparison with two other well-known MMSE-based compensation algorithms: SPLICE and MEMLIN.

The experimental results show the importance of modeling both feature spaces (clean and noisy) in order to obtain an accurate probability model for the MMSE estimation. Furthermore, the proposed normalization of noisy feature vectors and the more accurate representation of the noise distortion by means of the proposed subregion modeling, lead to further improvements. Finally, we think that the application of the proposed feature compensation algorithm to scenarios where stereo data is unavailable is an interesting issue that deserves more research in the future.

6. Acknowledgments

This work has been supported by an FPU grant from the Spanish Ministry of Science and Innovation and by project MEC-FEDER TEC2007-66600.

7. References

[1] A.M. Peinado and J.C. Segura, "Speech Recognition over digital channels. Robustness and Standards", *Wiley*, 2006.

[2] X. Huang, A. Acero, and H. Hon, "Spoken language processing: A guide to theory, algorithm, and system development", *Prentice Hall*, 2001.

[3] A. Acero, "Acoustical and environmental robustness in automatic speech recognition", *Kluwer Academic Publishers*, Norwell, MA, U.S.A, 1993.

[4] P.J. Moreno, "Speech recognition in noisy environments", Ph.D. thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, 1996.

[5] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database", in *Proc. Eurospeech 2001*, pp. 217–220.

[6] L. Buera, E. Lleida, A. Miguel, A. Ortega, and O. Saz, "Cepstral vector normalization based on stereo data for robust speech recognition", *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 3, pp. 1098–1113, Mar. 2007.

[7] M. Afify, X. Cui, and Y. Gao, "Stereo-based stochastic mapping for robust speech recognition", *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 7, pp. 1325–1334, Sep. 2009.

[8] H.G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluations of the speech recognition systems under noisy conditions", in *ISCA ITRW ASR2000*, 2000.

[9] "ETSI ES 201 108 - Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms", ETSI.

[10] José A. González, A.M. Peinado, A.M. Gómez, José L. Carmona, and Juan A. Morales-Cordovilla, "Efficient VQ-based MMSE estimation for robust speech recognition", in *Proc. ICASSP 2010*, pp. 4558–4561.