EFFICIENT VQ-BASED MMSE ESTIMATION FOR ROBUST SPEECH RECOGNITION

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

Dpto. de Teoría de la Señal, Telemática y Comunicaciones, University of Granada
{joseangl,amp,amgg,maqueda,jamc}@ugr.es

ABSTRACT

This paper presents a feature compensation technique based on the minimum mean square error (MMSE) estimation for robust speech recognition. Similarly to other MMSE compensation methods based on stereo data, our approach models the differences between clean and noisy feature spaces, and the resulting MMSE estimate of the clean feature vector is obtained as a piece-wise linear transformation of the noisy one. However, unlike other well-known MMSE techniques such as SPLICE or MEMLIN, which model the feature spaces with GMMs, in our proposal each feature space is characterized by a set of cells obtained by means of VQ quantization. This VQ-based approach allows a very efficient implementation of the MMSE estimator. Also, the possible degradation inherent to any VQ process is overcome by a strategy based on considering different subregions inside each cell and a subregion-based mean and variance compensation. The experimental results show that, along with a a very efficient MMSE estimator, our technique achieves even better recognition accuracies than SPLICE and MEMLIN.

Index Terms— Speech recognition, noise robustness, MMSE, stereo data.

1. INTRODUCTION

Pristine environments and high signal to noise ratio conditions are clearly unrealistic for current speech recognition systems. This has turned out to be yet more evident as new speech-enabled applications has been proposed for portable devices, making almost any adverse acoustical environment plausible [1]. In such scenarios, noise robustness becomes a crucial component of such systems.

Literature about robust speech recognition usually distinguishes between two main families of algorithms devoted to noise robustness [2]: feature compensation, that modifies the speech feature vectors, and model adaptation, where the acoustic model parameters are adjusted. The first approach has the advantage that can be seamlessly implemented into existing systems, since only a module that preprocess the feature vectors before they are fed into the speech recognizer is needed. In addition, feature compensation usually needs less data and time to compensate than model adaptation.

Many feature-domain cepstral de-noising techniques usually involve the use of a stereo database. A stereo database includes both clean and noisy features and can be used to learn the statistical relationship between both domains. While it is generally difficult to directly obtain stereo data, it can be relatively easy to collect for certain scenarios, e.g., in-car environments. In other cases, stereo-data are generated by adding noise sources to the existing "clean" training data [8].

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

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-based probability density functions.

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 Gaussian Mixture Models (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 technique, but their results are known to be inferior, due to the hard decision involved (a cell is represented by a centroid instead of a distribution 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 of the proposed VQ-based MMSE estimation is derived. In Section 3, our proposal is compared with other MMSE-based techniques. The experimental framework is described in Section 4 while the results are presented in Section 5. Finally, the paper is summarized in Section 6.

2. VQ-BASED MMSE ESTIMATION

Let $\boldsymbol{y}$ be the feature vector corresponding to a noisy speech segment distorted by additive noise. The MMSE estimate of the clean feature vector $\boldsymbol{x}$ can be calculated as,

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

The main problem of the above expression is to model the probability density function (pdf) $p(\boldsymbol{x}|\boldsymbol{y})$. In order to solve this problem, we can assume that the clean and distorted feature spaces are represented by means of 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 represent the components (e.g., Gaussians) of the pdf mixtures which model the clean and noisy spaces, respectively.

Applying this decomposition, the conditional probability $p(\boldsymbol{x}|\boldsymbol{y})$ can be approximated as,

$$p(\boldsymbol{x}|\boldsymbol{y}) \approx \sum_{k_x}\sum_{k_y} p\left(\boldsymbol{x}|k_x,k_y,\boldsymbol{y}\right) p\left(k_x|k_y,\boldsymbol{y}\right) p\left(k_y|\boldsymbol{y}\right) \quad (4)$$

Thus, the MMSE estimate can be expressed as,

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

In order to simplify eq. (5), in this work we propose to model the clean and distorted feature spaces by means of vector quantization (VQ) codebooks. Thus, each feature space, clean and distorted, is partitioned into a set of cells. The sets of VQ cells representing the clean space X and noisy space Y will be notated as $\{C_X^{(i)}\ (i = 1,\ldots,M)\}$ and $\{C_Y^{(j)}\ (j = 1,\ldots,N)\}$, respectively. As can be supposed, these cells will play now the role of the pdfs k_x and k_y of eq. (5).

For the case of the distorted space, the quantization cell which contains the input feature vector $\boldsymbol{y}$ is defined as the cell which minimizes the following weighted Euclidean distance,

$$C_Y^* = \underset{C_Y^{(j)}}{\operatorname{argmin}} \left\{ \left(\boldsymbol{\mu}_Y^{(j)} - \boldsymbol{y}\right)^T \operatorname{diag}\left(\boldsymbol{\Sigma}_Y^{(j)}\right)^{-1} \left(\boldsymbol{\mu}_Y^{(j)} - \boldsymbol{y}\right) \right\} \quad (6)$$

where the operator $\operatorname{diag}(\cdot)$ 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 the cell $C_Y^{(j)}$.

Given that the distorted feature space is VQ quantized, the posterior of the noisy mixture component $P(k_y|\boldsymbol{y})$ in eq. (5) is

$$P\left(C_Y^{(j)}\middle|\,\boldsymbol{y}\right) = \begin{cases} 1 & C_Y^{(j)} = C_Y^* \\ 0 & otherwise \end{cases} \quad (7)$$

This expression involves that the double sum of eq. (5) is reduced to a single sum. Additionally, the conditional probability $P\left(k_x|k_y,\boldsymbol{y}\right)$ is simplified to $P\left(C_X^{(i)}\middle|\,C_Y^{(j)}\right)$. With these modifications, the VQ-based MMSE estimation proposed in this paper finally adopts the following form,

$$\hat{\boldsymbol{x}} = \sum_{i=1}^{M} E\left[\boldsymbol{x}\middle|C_X^{(i)}, C_Y^*, \boldsymbol{y}\right] P\left(C_X^{(i)}\middle|\,C_Y^*\right) \quad (8)$$

where $P\left(C_X^{(i)}\middle|\,C_Y^*\right)$ can be estimated using a stereo database (containing clean feature vectors and its corresponding noisy ones under a certain environmental noise) to compute their relative frequency. On the other hand, the expectation term $E\left[\boldsymbol{x}\middle|C_X^{(i)}, C_Y^*, \boldsymbol{y}\right]$ defines the mapping of feature vectors between the clean cell $C_X^{(i)}$ and the noisy one C_Y^* due to environmental noise.

The use of VQ cells instead of pdfs for MMSE estimation may introduce some performance reduction due to the degrading nature of the VQ process. For this reason, we introduce in this paper the concept of subregions in a VQ-cell. 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 those distorted vectors are grouped. It must be pointed out that these subregions do not involve a hard partition of every cell (as the one due to VQ quantization) since, in fact, they will be overlapped due to the random nature of the noise. A scheme of this mapping is depicted in Figure 1.

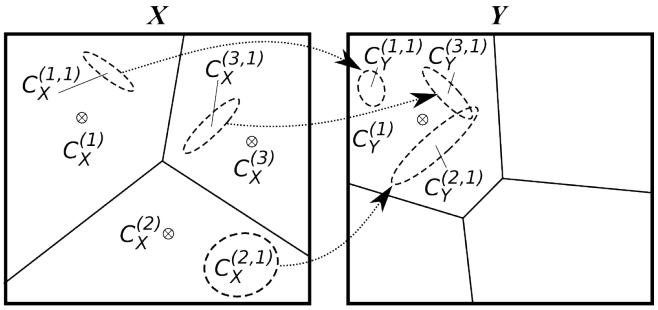

Fig. 1. Mapping between clean and distorted feature vectors applying the VQ-partitioned clean and noisy feature spaces.

In order to model the mapping between each pair of cells, we propose to compensate each feature vector by means of a linear transformation. The proposed transformation assumes that the subregions follow Gaussian pdfs with parameters $C_X^{(i,j)} \sim \mathcal{N}\left(\boldsymbol{\mu}_X^{(i,j)}, \boldsymbol{\Sigma}_X^{(i,j)}\right)$ and $C_Y^{(i,j)} \sim \mathcal{N}\left(\boldsymbol{\mu}_Y^{(i,j)}, \boldsymbol{\Sigma}_Y^{(i,j)}\right)$, where $\boldsymbol{\mu}_X^{(i,j)}$,$\boldsymbol{\mu}_Y^{(i,j)}$ and $\boldsymbol{\Sigma}_X^{(i,j)}$,$\boldsymbol{\Sigma}_Y^{(i,j)}$ are the mean vectors and covariance matrices, respectively. The transformation between both pdfs is achieved by means of the following whitening expression,

$$E\left[\boldsymbol{x}\middle|C_X^{(i)}, C_Y^{(j)}, \boldsymbol{y}\right] = \boldsymbol{\mu}_X^{(i,j)} + \\ + \left(\boldsymbol{\Sigma}_X^{(i,j)}\right)^{1/2} \left(\boldsymbol{\Sigma}_Y^{(i,j)}\right)^{-1/2} \left(\boldsymbol{y} - \boldsymbol{\mu}_Y^{(i,j)}\right) \quad (9)$$

where the exponent $1/2$ which raises both covariance matrices denotes the square root of the matrix. The square root of a matrix $\boldsymbol{\Sigma}$ is the matrix $\boldsymbol{\Sigma}^{1/2}$ which fulfills $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$. If the matrix $\boldsymbol{\Sigma}$ is symmetric and positive definite, which is the case of the covariance matrices, the square root can be computed as follows,

$$\boldsymbol{\Sigma}^{1/2} = \boldsymbol{V}\,\operatorname{sqrt}(\boldsymbol{D})\boldsymbol{V}^T \quad (10)$$

where $\boldsymbol{V}$ is the matrix of eigenvectors of $\boldsymbol{\Sigma}$, and $\boldsymbol{D}$ is the diagonal matrix of eigenvalues of $\boldsymbol{\Sigma}$. The notation $\operatorname{sqrt}(\boldsymbol{D})$ is introduced to denote the element-wise square root of the matrix $\boldsymbol{D}$.

The mean vectors and covariance matrices required by eq. (9) can be easily computed from a stereo database using the feature vectors assigned to each subregion.

3. COMPARATIVE DISCUSSION

We can find in the literature several MMSE-based compensation techniques similar to the proposed VQ-based one. In this section we analyze the relationship between our proposal and two well-known MMSE techniques: SPLICE and MEMLIN.

Stereo-based Piecewise LInear compensation for Environments (SPLICE) [5] models the distorted feature space by means of a Gaussian mixture model (GMM). For each Gaussian k_y in the noisy feature space, SPLICE computes a correction vector $\boldsymbol{r}_{k_y}$ using stereo data. Thus, the clean feature vector estimate for SPLICE is,

$$\hat{\boldsymbol{x}} = \sum_{k_y} \left(\boldsymbol{y} - \boldsymbol{r}_{k_y}\right) P(k_y|\boldsymbol{y}) \quad (11)$$

where $P(k_y|\boldsymbol{y})$ is the *a posteriori* probability of the noisy Gaussian k_y given the input feature vector $\boldsymbol{y}$. This term is computed by means of eq. (3) assuming Gaussian pdfs.

As we can see, two are the main differences between SPLICE and our proposal. First, SPLICE only models the noisy feature space using GMMs, while we model both spaces, clean and noisy, using VQ codebooks. Second, SPLICE only takes into account the shift introduced by the noise in the means of the GMM, not the rotations or scales in the covariance matrices.

In Multi-Environment Models based LInear Normalization (MEMLIN) [6] the clean and noisy feature spaces are modeled by means of GMMs. This technique result in a piece-wise linear compensation as follows,

$$\hat{\boldsymbol{x}} = \sum_{k_y} \sum_{k_x} \left(\boldsymbol{y} - \boldsymbol{r}_{k_x,k_y}\right) P(k_x|k_y) P(k_y|\boldsymbol{y}) \tag{12}$$

where k_y and k_x are Gaussians components of the noisy and clean features spaces, respectively; the posterior $P(k_y|\boldsymbol{y})$ is again computed using eq. (3), $P(k_x|k_y)$ is the probability of the clean Gaussian given the noisy one, and $\boldsymbol{r}_{k_x,k_y}$ is the correction vector that maps the data from the noisy Gaussian to the clean one. Those vectors are computed using stereo data for each possible environmental noise.

By comparing the MEMLIN estimate in eq. (12) with the proposed VQ-based MMSE estimate in eq. (8) we can observe the difference between both methods. First, the compensation in MEMLIN is estimated by averaging over all the noisy Gaussians, while in our method only the nearest noisy cell is selected. This leads to a more efficient implementation of the MMSE estimation. In contrast, we will see that this approximation will not lead to a performance degradation in terms of speech recognition accuracy. Second, basic MEMLIN does not model the transformation in the covariance due to additive noise. Under this simplification, the covariance of every subregion in our proposal is the identity matrix and eq. (9) becomes,

$$E\left[\boldsymbol{x} \left| C_X^{(i)}, C_Y^{(j)}, \boldsymbol{y} \right.\right] = \boldsymbol{y} - \left(\boldsymbol{\mu}_Y^{(i,j)} - \boldsymbol{\mu}_X^{(i,j)}\right) = \boldsymbol{y} - \boldsymbol{r}^{(i,j)} \tag{13}$$

4. EXPERIMENTAL FRAMEWORK

The experimental setup is based on the framework proposed by ETSI STQ-Aurora working group using the Aurora-2 database [8]. This database consists of utterances of connected English digits. In addition, two silence models (normal silence and short pause) are used. For our purposes, we have selected the clean training set and the clean utterances from the test *set A* of this database.

A set of 9 environmental noises is employed to test the proposed MMSE methods, namely: airport, highway, babble, bar, beach, pedestrian street, restaurant, street, and train station. Each noise in this set is split into two parts: two-thirds are employed for training purposes while the remaining third is reserved for testing. The training part of the 9 environmental noises has been 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 the clean condition. Similarly, 55 test conditions are defined by artificially contaminating the clean test set of Aurora-2 with the testing part of the 9 environmental noises. The SNRs considered here are the same that in the training stage.

Speech features are extracted according with the European Telecommunication Standards Institute front-end (ETSI FE, ES 201 108) [9]. The final feature vector employed by the recognizer consists of 12 Mel-Frequency Cepstral Coefficients (MFCCs) (the 0th order cepstral coefficient is discarded), plus the log-energy feature and their delta and delta-delta coefficients, resulting in a 39 dimension feature vector.

The recognizer is the one provided by Aurora-2 based on the HTK toolkit [10]. It uses 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.

A different VQ codebook with 256 cells is trained for every available training condition using only the static speech features (12 MFCCs and the log-energy). Training is carried out by a *K-means* algorithm applying the weighted Euclidean distance defined in eq. (6). Through this set of codebooks, the compensation parameters are estimated for the proposed VQ-based MMSE method using stereo data. These compensation parameters take into account the possible transformations due to environmental noise between the clean feature space and the distorted one, both modeled by means of VQ codebooks with the same number of cells. The utterance compensation is performed in a frame-by-frame basis, where an estimate of the clean speech parameters is computed using only the static features from each frame. Finally, the dynamic parameters (delta and delta-delta) are computed for each frame of the utterance using the previously compensated static features. In order to compare our proposal with other MMSE-based techniques (SPLICE and MEMLIN), GMMs with 256 Gaussians are also trained. Thus, one GMM with diagonal covariance is estimated for every available training condition using iterative splitting using only the static speech features.

5. RESULTS

In order to evaluate the proposed MMSE techniques, several tests are developed. First, all the experiments are carried out using oracle information about the environmental noise, i.e., each utterance is compensated using a set of compensation parameters trained under the same environmental noise. Note that in practice this information is not available. However, the oracle results will give us an estimate of the best performance that could be expected from each method, since each utterance is compensated under its same noise condition. Later, a set of soft-compensation experiments are conducted in order to evaluate the methods in a more realistic scenario. In the soft-compensation experiments, the final clean vector estimate is computed as a linear combination of the estimates obtained for several defined environments in the following way,

$$\hat{\boldsymbol{x}} = \sum_{e} \hat{\boldsymbol{x}}_e P(e|\boldsymbol{y}) \tag{14}$$

where $\hat{\boldsymbol{x}}_e$ is the MMSE estimate obtained for environment e and $P(e|\boldsymbol{y})$ is the *a posteriori* probability the environment. This probability is computed using the GMM trained for this environment.

5.1. Oracle experiments

Table 1 shows the performance in terms of word accuracy (WAcc; WAcc=1-Word Error Rate), in percent, achieved by different techniques for the oracle experiments. For each technique, the average WAcc (%) for each SNR, i.e., from clean condition to -5 dB, of the 9 environmental noises is shown. In addition, the average WAcc (%) between 0 dB and 20 dB is considered in column Avg.

The baseline results are obtained applying acoustic models trained with clean speech and no compensation. As expected, the

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.
Baseline	99.02	90.79	75.53	50.70	25.86	11.27	6.18	50.83
SPLICE	99.02	98.09	95.87	88.88	70.62	39.04	15.99	78.50
MEMLIN	99.02	98.36	97.01	92.43	78.26	47.03	18.76	82.62
iVQ-MMSE	99.02	98.23	96.79	91.60	76.82	46.60	20.02	82.01
dVQ-MMSE	99.02	98.33	97.06	92.43	78.70	48.88	20.26	83.08
fVQ-MMSE	99.02	98.37	97.15	92.88	79.61	50.04	20.89	83.61

Table 1. Recognition results achieved by the different MMSE-techniques for the oracle experiments.

addition of noise seriously degrades the performance of the recognizer. Three different versions of the proposed VQ-based MMSE estimation are evaluated. The first version, iVQ-MMSE, assumes identity covariance matrices for the calculation of the expected value in eq. (9). The second one, dVQ-MMSE, employs diagonal covariance matrices. In the last one, fVQ-MMSE, full covariance matrices are used. As can be seen, the three versions greatly improve the results achieved by the baseline system and SPLICE. This improvement shows the benefits of modeling both feature spaces, clean and distorted, instead of only one such as SPLICE. MEMLIN achieves a performance slightly better than our proposal iVQ-MMSE. In fact, both techniques are quite similar, although our approach is more computationally efficient. Further improvements can be obtained 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 VQ-based approach obtains better results than MEMLIN.

5.2. Soft-compensation experiments

The average recognition results obtained for the soft-compensation experiments are shown in Table 2. As can be seen, all methods suffer a performance degradation regarding the oracle experiments. This degradation is produced by mismatches in the environment identification. Nevertheless, the results presented here again show the superior performance of our proposal. This is reflected in relative improvements of 8.97 % and 3.02 % of fVQ-MMSE compared to SPLICE and MEMLIN, respectively. Furthermore, now MEMLIN and iVQ-MMSE behave equal.

	WAcc (%)
SPLICE	72.99
MEMLIN	77.21
iVQ-MMSE	77.29
dVQ-MMSE	79.04
fVQ-MMSE	79.54

Table 2. Average recognition results achieved by the different MMSE-techniques for the soft-compensation experiments.

6. CONCLUSIONS

In this paper we have presented a feature compensation technique based on MMSE estimation for robust speech recognition. Our proposal is shown to be a piece-wise linear function between the noisy feature space and the clean one, both modeled by means of vector quantization codebooks. We show that the use of VQ codebooks allows an efficient implementation of an MMSE estimator. In addition, a novel subregion-based approach is proposed in order to reduce the degradations introduced by the VQ quantization. The relationship between our technique and other MMSE-based methods, such as SPLICE and MEMLIN, has also been discussed.

Two sets of experiments are conducted. First, oracle experiments are carried out in order to evaluate the upper performance of the proposed compensation algorithms under non-environmental noise mismatch. Second, soft-compensation experiments test the methods in a more realistic scenario. In these last experiments, the final compensated vector is obtained as a linear combination of the estimates for several defined environments. The experimental results show the importance of modeling both feature spaces (clean and noisy) in order to obtain a more accurate probability model for the MMSE estimation. In addition, the compensation of feature vectors taking into account the transformations in mean and covariance introduced by the noise, leads to further improvements. As a result, important improvements over SPLICE and MEMLIN have been obtained.

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*, pp. 217–220, Aalborg, Denmark, 2001.

[6] 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 Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1098–1113, March 2007.

[7] X. Cui, M. Afify, and Y. Gao, "MMSE-based stereo feature stochastic mapping for noise robust speech recognition", in *Proc. ICASSP*, pp. 4077–4080, Las Vegas, USA, 2008.

[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*, Paris, France, 2000.

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

[10] S.Young, et al., "The HTK Book - Version 3.4", Cambridge University Engineering Department, December 2006, http://htk.eng.cam.ac.uk