International Journal of Information and Communication Technology Research

http://www.esjournals.org

Speech compensation using Stereo Based Stochastic Vector Mapping based on Full Covariance Models

Randa Al-Wakeel[1], Mahmoud Shoman[2], Magdy Aboul-Ela [1], Sherif Abdo[2]
[1]Sadat Academy for management Science and Information Systems, Egypt.
[2]Faculty of Computers and Information- Information Technology Department, Cairo University, Egypt.

ABSTRACT

Speech compensation techniques aim to provide speech recognition systems with the robustness against sources of noise existing in the real environments. These sources of noise cause the recognition performance to deteriorate dramatically.

In this paper, we are interested in Stereo based Stochastic Vector Mapping (SSM) speech compensation technique introduced in [1]. In [1], in the experimental work, it was assumed that the speech features are uncorrelated (independent). This assumption simplifies the estimation of the enhancement parameters and it reduces the needed implementation time. In this paper, we will extend the experimental work to the case when the speech features are correlated (dependent). We aim to clarify the effect of considering the correlation on the efficiency of SSM. We considered the two estimators Maximum A Posteriori (MAP) estimation and Minimum Mean Square Error (MMSE) estimation used in [1]. A part of the experimental work was dedicated to test the SSM with Multi Style Trained (MST) recognition models and also with recognition models trained using SSM compensated speech.

However, considering the correlation between features introduces better performance, it cannot be applied in real time applications without a way to reduce the complexity of the implementation and the time needed.

Keywords: *speech compensation, noisy speech recognition, Stochastic vector Mapping.*

1. INTRODUCTION

Although speech recognition systems work reasonably well in quiet conditions, its accuracy degrades severely when the systems operate in adverse acoustical environments. Such degradation is mainly caused by mismatches between training and testing environments due to various sources of noise exist in the testing environments. In this paper, we are interested in robust speech recognition systems in noisy environments. Such systems which keep reasonably satisfied recognition accuracy even in the presence of noise sources. We are interested in SSM [1] speech compensation approach. The main assumption in [1] is that the speech features are statistically independent. This assumption simplifies the compensation algorithm and reduces the implementation time needed. On the other hand, ignoring the correlation between the speech feature limits the improvements achieved in the recognition accuracy.

The aim of our work is to clarify the effect of considering the correlation between speech features on the efficiency of SSM. This work is considered as an extension for the experimental work presented in [1]. We used full covariance GMMs to model the correlation between speech features.

The paper is organized as follows. In section 2.1 we will give an overview about the techniques of speech recognition in noisy environments and in section 2.2 we will give a brief description for modeling the statistical distribution of speech features in such systems. In section 3, we describe how the covariance between the speech features is modeled when the features are statistically dependent and when they are independent. We also will give a brief description for the Discrete Cosine Transform (DCT) which is used in the computation of MFCC speech features in order to remove the dependency between them, making them uncorrelated. Section 4, is dedicated for SSM introduced in [1]. The experimental work and the results are introduced in section 5. Finally, the conclusions and the suggested future are introduced in section 6.

2. SPEECH RECOGNITION IN NOISY ENVIRONMENTS

2.1 Speech Enhancement Techniques

Many techniques have been developed to address the problem of robust speech recognition against the environmental noise [1, 2]. Such techniques can be classified into three classes. The first class of techniques involves the extraction of environmental invariant (robust) features. One of these techniques is Cepstral Mean Normalization (CMN) [3]. CMN is a popular method which makes the speech features robust against the channel noise. CMN is involved in most speech recognition systems.

The second approach is the noisy speech features compensation [4, 1, and 5], in which the clean speech features are estimated from the input noisy speech before entering into the speech recognition system which had been trained using clean data. Examples for this approach include SSM [1], SPLICE [4] and VTS [5].One of the advantages of this class of methods is that they can be implemented independent of the recognizers. This simplifies the operation of speech enhancement and does not require large amount of data.

Speech model adaptation is the third approach for noise robustness. In this approach the recognition models are

manipulated to better fit the noise environment. Several model adaptation techniques have been proposed such as MLLR [6], and MAP adaptation [7].

The first two classes of methods are computationally simpler than the third one and they can be implemented independently from the recognizer. On the other hand the third class of methods need a large amount of adaptation data in order to compensate the probability distributions of the recognizer so they can model the noisy speech more efficiently [8,9]. In many situations, it is difficult to obtain this large amount of data. In this paper we are interested in the second class of methods.

Speech compensation methods may use single channel speech recording [5]. Therefore, the available data to estimate the clean speech are only the clean speech models and the observed noisy speech. No additional information is available. A pre-stage is needed before the estimation of the clean speech. In this pre-stage the environmental noise is estimated based on some assumptions made in order to reduce the complexity of the problem of noisy speech enhancement. For example the noise can be assumed additive. The disadvantage of such methods is that the performance of speech enhancement approach is reasonably good when the noise existing in the environment is the same as it is assumed, however it introduces some weakness when the noise is of different type.

Other methods use microphone arrays [10]. One microphone records the clean speech signal, while the others represent the sources of noise exist in the acoustic environment. Clean speech and noise models can be estimated and used to enhance observed noisy speech. However, in many situations the microphone arrays are not available. In addition, even when multi microphones (more than two microphones) exist, the mounting of them in a particular topology is a difficult process [11].

In most of speech recognition and speech enhancement systems, Mel-Frequency Cepstral Coefficient (MFCC) features are used to train and test HMMs and GMMs. Among the factors that have contributed to the popularity of MFCCs are low computational complexity and high recognition performance in clean conditions [12]. In addition, HMM is a well-known and widely used statistical approach to characterize the spectral properties of frames of speech features as it has an advantage of providing a natural and highly reliable way of recognizing speech for a wide variety of applications and it integrates well into systems incorporating information about both acoustics and semantics [5]. The HMM states output probabilities are usually represented by Gaussian Mixture Densities [13].

In the majority of speech enhancement techniques the main assumption made is that the speech features in both training and testing stages are uncorrelated (independent). This assumption comes from the fact that the Discrete Cosine Transform (DCT) is applied as a main step in the computation of MFCCs. DCT aims to uncorrelate the speech features (making them statistically independent). This assumption reduces the computations and the time needed in both speech recognition and speech enhancement systems.

In this work, we focus on speech enhancement techniques which utilize stereo databases. In particular, we are interested in the Stereo based Stochastic Vector Mapping (SSM) approach which was introduced in [1]. The use of stereo data to build feature mappings was popular in earlier noise robustness research [14,15,7], and in SPLICE [4]. In [1], the experimental work was based on the assumption that speech features are independent. In this work, we will extend the experimental work to the case when the MFCCs are considered correlated and see if this improves the performance of the speech enhancement approach. This also will give us an idea about the efficiency of DCT in uncorrelating the MFCCs.

2.2 Modeling the statistical distribution of speech features

Let $x = (x_1, x_2, ..., x_M)$ is the speech feature vector, *M* is the dimension of the vector. In GMM the frame *x* is produced by a probability density function:

$$p(x) = \sum_{k=1}^{K} a_k(t) N(x; \mu_k, \Sigma_k)$$

where, *K* is the number of mixtures, a_k is the mixture weight, and

$$N(x; \mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_k|}} e^{-1/2 (x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}$$

μ_k is the mean, and Σ_k is the covariance of mixture *k*. μ_k and Σ_k are computed using the following equations

$$\mu_k = E[x] = \frac{1}{N} \sum_{n=1}^{N} x^n$$

$$\Sigma_k = E[(x-\mu_k)(x-\mu_k)^T]$$

$$= \frac{1}{N} \sum_{n=1}^{N} (x^n - \mu_k)(x^n - \mu_k)^T$$

N denotes the number of frames in the training data. *T* denotes the transpose operation. In the following section we will be interested in the covariance modeling.

3. MODELING THE COVARIANCE MATRIX

In the general case when the features are correlated covariance is a *M*M* symmetric matrix and it looks as follows:

$$\begin{bmatrix} r(1,1) & r(1,2) & \cdots & r(1,M) \\ r(2,1) & r(2,2) & \cdots & r(2,M) \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & & \cdots & \vdots \\ r(M,1) & \cdot & \cdots & r(M,M) \end{bmatrix}$$

One of the motivations for using the DCT transform is to decorrelate the feature vector so that the diagonal matrix approximation becomes reasonable [16]. Only the diagonal elements of the covariance matrix will have values and the other elements will take value of zero. So it will be as follows:

$$\begin{bmatrix} r(1,1) & 0 & \cdot & \cdot & 0 \\ 0 & r(2,2) & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & r(M,M) \end{bmatrix}.$$

The diagonal elements are the variances of the coefficients. They will be denoted in this paper as σ_{ii}^2 to denote the variance of the coefficient *i*, where $1 \le i \le M$.

In general the use of full covariance Gaussians in Large Vocabulary systems is impractical due to the sheer size of the model set. Even with small systems, training data limitations often preclude the use of full covariance matrices is $O(d^2)$ compared to $O(d)$ in case of diagonal case where d is the dimensionality of the feature vector[16].

In the following section we will clarify briefly the operation of DCT.

3.1 Discrete Cosine Transform (DCT)

DCT is the last step in the process of calculating the MFCCs. It is applied to the log filter bank amplitudes. As the filterbanks are all overlapping, the filterbank energies are quite correlated with each other. The DCT is needed for two reasons:

The DCT tends to decorrelate the energies which means diagonal covariance matrices can be used to model the features in e.g. a HMM classifier. Whereas the number of filters in the filterbank is usually more than twenty, it is common to use only the first 13 coefficients resulting from the DCT. the higher DCT coefficients represent fast changes in the filterbank energies and it turns out that these fast changes actually degrade ASR performance, so we get a small improvement by dropping them [12].

Let $s = [s_1, s_2, \ldots\ldots, s_N]^T$ is a frame of the input speech signal. *s* is transformed into $x = [x_1, x_2, \ldots\ldots, x_M]^T$ using the following equation [22]:

$$x_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} s_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right)$$

where $\{s_i\}$ are the log filter bank amplitudes, *N* is number of filterbank channels, $\{x_i\}$ are the resulted cepstral coefficients and M is the number of them.

4. STEREO-BASED STOCHASTIC VECTOR MAPPING (SSM)

The stochastic vector mapping (SVM) based methods were explored in [17]. This approach performs a frame-dependent bias removal to compensate for "environmental" variabilities. The objective is to estimate the compensated feature vector $\hat{x}$ from the original feature vector y by applying an environment-dependent transformation $F(y; \Theta^{(ey)})$ where $\Theta^{(ey)}$ represents the trainable parameters of the transformation and *ey* denotes the corresponding environment class to which *y* belongs. SVM introduced in [17] doesn't rely on the availability of the stereo recordings of both clean and noisy speech for the estimation of SVM function parameters. Here in this paper we are interested in Stochastic Vector mapping approach that uses a stereo database (i.e. data that consists of simultaneous recordings of both the clean and noisy speech).

The relationship between the clean and noisy features is modeled by a joint statistical distribution (GMM). In order to train this model, both the clean and noisy channels are stacked to form a large augmented space. This new data space is used to train the Stereo GMM. In testing stage, this joint model and the observed noisy speech will be used to estimate the clean speech. In [1] two estimators were developed to obtain the clean speech, Maximum A-Posteriori (MAP) estimation and Minimum Mean Square Error (MMSE) estimation. The assumption is that the spectral coefficients are independent so that, the ith noisy coefficient is used to predict the ith clean coefficient or alternatively using a time window around the ith noisy coefficient to predict the ith clean coefficient.

In this paper, we will extend the experiments to the case when the features are assumed dependent and we will consider the correlation between them and its effect on the error reduction rate achieved.

In the following discussion we will clarify how SSM works. In section 4.1, the structure of S SSM is shown and in section 4.2 MAP and MMSE clean speech estimators introduced in [1] will be clarified in detail.

4.1 The joint probability distribution

The joint distribution is built using the stereo database. Let $x = (x_1, x_2, \ldots\ldots, x_M)$ be the clean feature vector, *M* is the dimension of the vector. The corresponding simultaneously recorded noisy representation is $y = (y_1, y_2, \ldots\ldots, y_M)$.

Define z ≡ (x, y) as the concatenation of the two channels. The first step in constructing the mapping is training the joint probability model for p (z). We use Gaussian mixtures for this purpose, and hence write

$$p(z) = \sum_{k=1}^{K} c_k N(z; \mu_{z,k}, \Sigma_{zz,k})$$

where *K* is the number of mixture components, c_k, $\mu_{z,k}$ and $\Sigma_{zz,k}$ are the mixture weight, mean and covariance of *k* th component, respectively. Also both the mean and covariance can be partitioned as

$$\mu_{z,k} = \begin{pmatrix} \mu_{x,k} \\ \mu_{y,k} \end{pmatrix}$$

$$\Sigma_{zz,k} = \begin{pmatrix} \Sigma_{xx,k} & \Sigma_{xy,k} \\ \Sigma_{yx,k} & \Sigma_{yy,k} \end{pmatrix}$$

The GMM model can be trained using EM algorithm [18]. During testing, both the observed noisy speech and the augmented model are used to estimate the clean speech.

In section 4.2 we will present MAP estimation and MMSE estimation briefly and the reader can refer to [1] for details.

4.2 Clean speech estimation in SSM

In [1] two clean speech estimators were introduced, Maximum A Posteriori (MAP) estimation and Minimum Mean Square Error (MMSE) estimation based on the joint probability distribution.

4.2.1. MAP clean speech estimation

The clean speech can estimated using MAP estimation. This will be defined as:

$$\hat{x} = \arg\max_x p(x \mid y)$$

the equation can be further decomposed as

$$\begin{aligned}\hat{x} &= \arg\max_x p(x \mid y) \\ &= \arg\max_x \sum_k p(x,k \mid y) \\ &= \arg\max_x \sum_k p(k \mid y)\, p(x \mid k,y)\end{aligned}$$

Using EM algorithm [18], where the clean speech can be iteratively estimated e objective function will be as follows;

$$\begin{aligned}\hat{x} &= \arg\max_x \sum_k p(k \mid \bar{x},y) \log p(y \mid x)\, p(x \mid k,y) \\ &= \arg\max_x \sum_k p(k \mid \bar{x},y) \log[p(k \mid y) + \log p(x \mid k,y)] \\ &\equiv \arg\max_x \sum_k p(k \mid \bar{x},y) \log p(x \mid k,y)\end{aligned}$$

$$\equiv \arg\max_x \frac{-1}{2} \sum_k p(k \mid \bar{x},y) \cdot \left[\log|\Sigma_{x|y,k}| + (x-\mu_{x|y,k})^T \Sigma^{-1}_{x|y,k} (x-\mu_{x|y,k})\right] \quad (1)$$

where $\bar{x}$ is the value of x from previous iteration, and $x|y$ is used to indicate the statistics of the conditional distribution $p(x|y)$. By differentiating equation (1) with respect to x , setting the resulting derivative to zero, and solving for x, we arrive at the clean feature estimate given by:

$$\sum_k p(k \mid \bar{x},y) \Sigma^{-1}_{x|y,k} \hat{x} = \sum_k p(k \mid \bar{x},y) \Sigma^{-1}_{x|y,k} \mu_{x|y,k} \quad (2)$$

which is basically a solution of linear system equations, $p(k \mid \bar{x},y)$ are the usual posterior probabilities that can be calculated using the original mixture model and Bayes rule and the conditional statistics $\mu_{x|y,k}$ and $\Sigma_{x|y,k}$ are known to be:

$$\mu_{x|y,k} = \mu_{x,k} + \Sigma_{xy,k} \Sigma^{-1}_{yy,k} (y - \mu_{y,k}), \quad (3)$$

and

$$\Sigma_{x|y,k} = \Sigma_{xx,k} - \Sigma^{-1}_{yy,k} \Sigma_{yx,k} \quad (4)$$

The following discussion shows the equation when the features are assumed independent and when they are considered dependent.

a) The speech features are independent

In [1], the assumption is that the coefficients are independent so, the *ith* clean coefficient can be estimated using the *ith* noisy coefficient or alternatively using a time window around the ith noisy coefficient. Applying this assumption to equations will reduce the solution of the linear system in equation (2) to the following simple calculation for every vector dimension.

$$\hat{x} = \frac{\sum_k p(k \mid \bar{x},y) \mu_{x|y,k} / \sigma^2_{x|y,k}}{\sum_k p(k \mid \bar{x},y) / \sigma^2_{x|y,k}}$$

So, in this equation, x is a scalar, and $\sigma^2_{x|y,k}$ is used instead of $\Sigma_{x|y,k}$ to indicate it is also a scalar. Limiting our attention to a single feature dimension, the clean speech x is 1-dimensional, while the noisy speech y has the dimension of the window say L_n, and accordingly the mean and the variance will be 1-dimensional. In this paper we set L_n =1 in the experimental work.

b) The speech features are dependent

In this case the noisy speech and the clean speech are vectors. By using equation (3), we will reach equation (5) where the clean speech estimate $\hat{x}$ is expressed as a mixture of linear transformations weighted by components posteriors:

$$\hat{x} = \sum_k p(k \mid \bar{x},y)(A_k y + b_k) \quad (5)$$

where:

$$A_k = CD_k, b_k = Ce_k, \text{ and}$$

$$C = \left(\sum_k p(k \mid \bar{x},y) \Sigma^{-1}_{x|y,k}\right)^{-1}$$

$$e_k = \Sigma^{-1}_{x|y,k} (\mu_{x,k} - \Sigma_{xy,k} \Sigma^{-1}_{yy,k} \mu_{y,k})$$

$$D_k = \Sigma^{-1}_{x|y,k} \Sigma_{xy,k} \Sigma^{-1}_{yy,k}$$

$p(k \mid \bar{x},y)$ can be calculated using the original mixture model and Bayes rule,

A reasonable initialization is to set $\bar{x} = y$, (i.e. initialize the clean speech with the noisy speech) [1].

It is important to clarify the dimensions of all the parameters used in the equations. Table 1 shows the dimensions of the parameters in this case.

Table 1: The dimension of the SSM parameters

The Parameter	Dimension
$\mu_{x,k}$	1 * M
$\mu_{y,k}$	1 * M
$\Sigma_{xx,k}$	M * M
$\Sigma_{yy,k}^{-1}$	M* M
$\Sigma_{xy,k}$	M * M
$\mu_{x\|y,k}$	M * 1
$\Sigma_{x\|y,k}$	M * M
C	M * M
A_k	M * M
e_k	M * 1
b_k	M * 1
D_k	M * M

This is the general case of estimating the clean speech. We see that at each frame a matrix inversion is needed to compute the C matrix. Therefore, the use of full covariance GMMs introduces heavy computational burden in implementation stage of SSM. This is clarified in the experimental work according to a comparison between the two cases on the basis of the implementation time needed.

4.2.2 MMSE-based Estimation

The MMSE estimate of the clean speech feature given the noisy speech feature y is known to be the mean of the conditional distribution $p(x|y)$. This can be written as:

$$\hat{x} = E[x|y] \qquad (6)$$

Considering the GMM structure of the joint distribution, Equation (6) can be further decomposed as:

$$\hat{x} = \int_x p(x|y)x dx = \sum_k \int_x p(x,k|y)x dx$$
$$\hat{x} = \sum_k p(k|y) \int_x p(x,k|y)x dx$$
$$= \sum_k p(k|y) \int_x p(x|k,y)x dx$$
$$= \sum_k p(k|y)E[x|k,y]$$

The posterior probability term $p(k|y)$ can be computed as

$$p(k|y) = \frac{p(k,y)}{p(y)} = \frac{p(y|k)p(k)}{\sum_k p(y|k)p(k)}$$

and the expectation term $E[x|k, y]$ is given in Equation (3).

MMSE predictor can also be written as a weighted sum of linear transformations as follows:

$$\hat{x} = \sum_k p(k|y)(F_k y + g_k) \qquad (7)$$

where

$$F_k = \Sigma_{xy,k} \Sigma_{yy,k}^{-1}$$
$$g_k = \mu_{x,k} - \Sigma_{xy,k} \Sigma_{yy,k}^{-1} \mu_{y,k}$$

Applying the special case to these two equations, F_k will take dimension L_n and g_k will be 1- dimensional. However when the features are correlated F_k is $M * M$ dimensional and g_k is a vector of M –dimension.

It is clear that the MMSE estimate is not performed iteratively and that no matrix inversion is required to calculate the clean speech estimate in Equation (7).

5. EXPERIMENTAL WORK AND RESULTS

The experiments presented in this paper have been implemented using CARVUI database recorded inside a moving car. The data was collected in Bell Labs area, under various driving conditions (highway/city roads) and noise environments (with and without radio/music in the background). About 2/3rd of the recordings contain music or babble noise in the background. A total of 56 speakers participated in the data collection. The speech material from 50 speakers is used for training, and the data from the 6 remaining speakers is used for test. Simultaneous recordings were made using a close-taking microphone and a 16-channel array of 1st order hypercardiod microphones mounted on the visor. Data from 2 channels only are used. The first one is the close-talking microphone (CT). The second one is a single channel from the microphone array, referred to as hands-free data (HF) henceforward. The average SNR is about 21 db for the CT channel and 8 db for the HF channel. The experiments were implemented using the part of the database that contains only the digits utterances.

The data are recorded at 24kHz sampling rate and are down sampled to 8kHz. They are then windowed with 12.5msec frame rate, and 25msec frame size. Filter bank analysis is then applied. The filter bank has 26 channels. MFCCs are calculated by applying DCT to log filter bank amplitudes to produce 12 cepstral coefficients. The energy coefficients are also computed and appended to the cepstral vectors. Therefore, each cepstral vector contains 13 coefficients. During recognition, delta and delta delta coefficients are computed on the fly resulting in vectors of 39 coefficients.

We have implemented speech compensation experiments using Gaussian mixtures models with sizes 16, 64, and 256 mixtures.

The recognition models for digits are trained using about 6500 training clean speech files collected from the CT microphone and tested using about 800 utterances. For each digit from 0 to 9 there is a HMM model. The digit 0 has an additional model for the utterance ‘oh’. A twelfth model is considered to model silence. Each of these models consists of 6 states. Each state contains 8 mixtures. The clean speech files are also used to build the clean speech Gaussian mixture models that were used in the experiments. Training and recognition is done using HTK [13].

International Journal of Information and Communication Technology Research

http://www.esjournals.org

A baseline set of results for this task are given in table 2.

Table 2: The base Line Recognition Results

Condition	SER
Clean/Clean	12.94
Noisy/Noisy	16.79
Clean/Noisy	31.72

This table shows the evaluation of the recognition system in the different test/train conditions. In terms of Sentence Error Rate (SER). Clean refers to the CT and noisy refers to the HF.

The first result shows the SER when clean speech is recognized using systems trained using clean database. The second result shows the result when the recognition system is trained and tested using noisy speech. The error rate is low in the two cases. This is because the environment of testing is similar to the environment of training. However, when the system is trained using clean speech and tested using noisy speech the recognition performance degrades severely. This is clear in the third experiment of table 1 which shows how the system will perform in practical situations when the system is trained using clean speech and tested using observed speech. We can see the dramatic degradation in the performance due to the noise, 31.72 in noisy environment vs. 12.94 in clean environment.

Applying compensation techniques to the noisy speech improves the recognition results. Table 3 shows the results obtained by applying SPLICE and first order Vector Taylor Series (VTS) technique.

Table 3: SER after SPLICE and VTS Compensation

	16	64	256
SPLICE	29.85	28.48	27.49
VTS-1st order	31.84	30.47	28.86

We see that SPLICE represents about 6% improvement to the baseline, and VTS represent about 4%.

The next set of experiments aims to test the SSM performance as a compensation technique.The mapping is applied to the MFCC coefficients before CMN. After applying the compensation, CMN is performed followed by calculating the delta and delta-delta coefficients. In [1], the SSM compensation was implemented based on the assumption that speech features are independent (diagonal covariance), so the clean speech estimate is scalar (equation 2). Here we will extend the experimental work to the case when the speech features are correlated (full covariance). So the clean speech frame is estimated using the noisy frame. In these experiments, DiagC refers to GMMs with diagonal covariance matrices, while Full refers to GMM with full covariance matrices.

Tables 4 and 5 show the results obtained in case of implementing one iteration of MAP estimation and MMSE estimation, respectively.

Table 4: SER after SSM with MAP Estimation

	16	64	256
SSM-DiagC	25.5	26.12	24.13
SSM-Full	21.14	23.26	24.5

Table 5: SER after SSM with MMSE Estimation

	16	64	256
SSM-DiagC	31.1	29.48	31.84
SSM-Full	24.01	22.26	22.51

In table, 4 we see that the increase in the number of mixtures in the stereo GMM does not increase the efficiency of SSM, however in general we can see that the SSM performs better when full covariance matrices are used than the case of diagonal covariance. In particular, the use of Full covariance results in enhancement by a percentage of 15.5% compared to only 9% in case of Diagonal covariance when Map estimation is applied with 16 mixture GMM. In addition, in table 5 when MMSE estimation is used, we can see that the use of full covariance achieves about 14% enhancement percentage in case of full covariance compared to about 3% achieved when diagonal covariance is used, in case of 64 mixtures GMMs.

We also implemented one iteration of MAP estimation of the clean speech initialized by MMSE estimates. The obtained results are shown in table 6.

Table 6: SER after MMSE-MAP Compensation

	16	64	256
SSM-DiagC	40.3	36.82	36.82
SSM-Full	38.31	26.37	26.8

Table 6 shows that implementing MAP and MMSE estimation separately achieves better recognition results than MAP estimation initialized by MMSE estimation.

In the next experiment, we tested the performance of SSM using recognition models trained using noisy speech compensated by SSM. The results of this experiment are shown in table 7. The type of covariance and the SSM estimator precede the word "test" to denote the type of processing implemented on the test noisy data and also precede the word "Rec" to denote the type of processing performed on the noisy database used to train the recognition model.

The GMMs used in this experiment were 16 mixtures and the structure recognition model was the same as the recognition models trained using clean database (i.e. each digit model has 6 states and each state has 8 mixtures.

Table 7: SER after SSM evaluated by recognition models trained by SSM compensated noisy speech

	SER
Full MAP test with Full MAP_Rec	19.40
Full MMSE test with Full MMSE_Rec	19.28
DiagC MAP with test Full MAP_Rec	21.14
DiagC MMSE test with Full MMSE_Rec	25.75
DiagC MAP test with DiagC MAP_Rec	20.9
DiagC MMSE test with DiagC MMSE_Rec	23.01

We also tested SSM with Multi Style Trained (MST) recognition models. This is done in two steps:

1. Applying SSM to the noisy training data to yield SSM enhanced speech.
2. Constructing the training database by merging the clean speech database with the SSM enhanced speech database. The new database is used to train the recognition models.

The same notation, used in the last table, is used in this experiment.

Table 8: SER after SSM evaluated by Multi-Style trained recognition models

	SER
Full MAP test with Full MAP_Rec	18.41
Full MMSE test with Full MMSE_Rec	17.16
DiagC MAP test with Full MAP_Rec	19.14
DiagC MMSE testwith Full MMSE_Rec	28.48
DiagC MAP test with DiagC MAP_Rec	21.39
DiagC MMSE test with DiagC MMSE_Rec	23.76
Clean test with DiagC MMSE_Rec	12.19
Clean test with DiagC MAP_Rec	12.44
Clean test with Full MMSE_Rec	9.58
Clean test with Full MAP_Rec	10.95

From tables 7 and 8 we see that employing SSM to enhance noisy speech database and use it to train the recognition models achieves better recognition when the enhanced noisy database is used separately and when it is used with the clean speech database to train the recognition models in MST. However the MST achieves the best recognition results. For example when the recognition models are trained using MAP SSM compensated data, the improvement was about 18% in full covariance case (test and train). This percentage increases to be 19.4% in MST.

In general, we see that considering the correlation between speech features makes the recognition models more efficient. However, full covariance modeling is a very time consuming operation making the application of SSM with full covariance impractical in real time applications. The following table shows a comparison between the implementation time needed for SSM with diagonal covariance and implementation time needed for SSM with full covariance to compensate 804 noisy test files in cases of 16, 64 and 256 GMM models.

Table 9: Comparison between SSM-DiagC and SSM-Full in terms of implementation time.

	16	64	256
Full MAP	18 min	50 min	190 min
DiagC MAP	1 min	4 min	17 min/ L
Full MMSE	6 min	12 min	29 min
DiagC MMSE	30 sec	1 min	4 min

6. CONCLUSIONS AND FUTURE WORK

From the last section we see that the SSM approach is superior over SPLICE approach and also over VTS approach [1, 5]. We also saw that using full covariance in SSM achieved lower SERs in most of the cases than the cases when we use diagonal covariance matrices in the stereo GMMs. So, we can say that the assumption that the MFCC coefficients are statistically independent is not optimal but it is used to simplify the implementation.

On the other hand, the use of full covariance matrices imposes a heavy computational burden, making it difficult to achieve real-time recognition, specifically with the increase in the number of mixtures in the stereo GMM. Moreover, one rarely has enough data to (reliably) estimate full covariance matrices. Some of these disadvantages can be overcome by parameter tying (e.g. sharing the covariance matrices across different states or models) [19].

Factor analysis [19, 20] represents a compromise between full covariance and diagonal covariance. Factor analysis is a strategy for dimensionality reduction. It enables one to express the covariance matrices in terms of small number of parameters that model the most significant correlations without incurring much overhead in time or memory.

Another direction to solve the correlation problem is to use a linear decorrelation transform [21, 22]. The objective is to transform the coefficients into a new space in which they are decorrelated so they can be modeled by diagonal covariance matrices. After estimating the clean speech, the coefficients are re-transmitted again into the original space and can be decoded using the recognition system. This experiment is our suggested future work.

REFERENCES

[1] M. Afify, X. Cui, and Y. Gao “Stereo-Based Stochastic Mapping For Robust Recognition,”Audio, Speech, and Language Processing, IEEE Transactions, Vol. 17, No. 7, pp. 1325-1334, Sept. 2009.

[2] N. Garg1, Jyoti Gupta," Review On Speech Enhancement Using Signal Subspace Method 2," International Journal of Application or Innovation in Engineering & Management, Vol. 2, Issue 5, pp. 215-221, May 2013.

[3] F. Liu, R.M. Stern, X. H. Huang, and A. Acero, "Efficient Cepstral Normalization for Robust speech Recognition," In: Proc, of ARPA workshop on human language technology; pp 69-74, 1993.

[4] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE Algorithm on the AURORA 2 Database," in Proc. Eurospeech, Denmark, pp. 217-220, September, 2001.

[5] P.J. Moreno, "Speech Recognition in Noisy Environments," PhD thesis, Carnegie Mellon University, Pittsburgh, Pensilvania, April 1996.

[6] X. Cui, A.Alwan, "Noise Robust Speech Recognition Using FeatureCompensation Based on Polynomial," IEEE Transactions on Speech and Audio Processing, Vol. 13, NO. 6, Nov 2005.

[7] J. Gauvain, C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,"Speech and Audio Processing, IEEE Transactions on Vol. 2, Issue: 2, pp. 291 – 298, Apr. 1994.

[8] H. Y. Jung, B. O. kang , B. Kangj,and Y. Lee, "Model Adaptation using Discriminative Noise Adaptive Training Approach for New Environments" ETRI Journal, Vol. 30, No. 6, pp. 865-867, Dec. 2008.

[9] X. Wang and D. O'Shaughnessy "Environmental Independent ASR Model Adaptation/Compensation by Bayesian Parametric Representation" IEEE Transactions Audio, Speech & Language Processing - TASLP, vol. 15, no. 4, pp. 1204-1217, 2007.

[10] H. Attias, L. Deng, " A new approach to speech enhancement by microphone array using EM and Mixture Models" InterSpeech,pp. 785-788, 2002.

[11] I.A. McCowan. "Robust Speech Recognition using Microphone Arrays," PhD Thesis, Queensland University of Technology, Australia, 2001.

[12] S. G. S. Pettersen. "Speech Recognition in the presence of additive noise, " PhD Thesis, Norwegian University of Science and Technology, 2008.

[13] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland. The HTK Book. Revised for HTK Version 3.4 Dec. 2006. http://htk.eng.cam.ac.uk/

[14] A. Acero and R. Stern, " Robust Speech Recognition By Normalization Of The Acoustic Space," in Proc. Int. Conf. on Acoustics, Speech and Signal Processing, Toronto, Canada, pp. 893-869, 1991.

[15] F. H. Liu, R. M. Stern, A. Acero, and P.J.Moreno, "Environment Normalization for Robust Speech Recognition using Direct Cepstral Comparison," Proc. of the ICASSP, Adelaide, Australia, pp. 61-64, 1994.

[16] E. W. Robert, York, "Modeling Longitudinal Data, " P. Springer New pp 243-301, 2005.

[17] Q. Huo, and D. Zhu, "A maximum likelihood training approach to irrelevant variability compensation based on piecewise linear transformations," in Proc. Interspeech'06, Pittsburgh, Pennsylvania, pp.1129-1132, Sep., 2006.

[18] J. Bilmes, "A Gentle Tutorial On The EM Algorithm And Its Application To Parameter Estimation For Gaussian Mixture And Hidden Markov Models," Technical Report ICSI-TR-97-021, University of Berkeley, 1998.

[19] L. Saul, and M. Rahim, "Maximum Likelihood and Minimum Classification Error Factor Analysis for Automatic Speech Recognition," IEEE Transactions Audio, Speech & Language Processing, vol. 8, no. 2, pp. 115-125, 2000.

[20] K. Yao, K. Paliwal, and T. Lee, "Generative factor analyzed hmm for automatic speech recognition," Speech communication, vol. 45, no. 4, pp. 435–454, 2005.

[21] M. Mastriani, and J. Gambini, "Fast Cosine Transform to Increase Speed-up and Efficiency of Karhunen-Loève Transform for Lossy Image Compression," International Journal of Information and Mathematical Sciences 6, pp. 79–89, 2010.

[22] J.V. Psutka, L. Müller "Comparison Of Various Feature Decorrelation Techniques In Automatic Speech Recognition," Journal of Systemics, Cybernetics and Informatics, vol. 5, no. 1, pp. 18-26, 2007.