AN EFFECTIVE MISSING FEATURE COMPENSATION METHOD FOR SPEECH RECOGNITION AT NOISY ENVIRONMENT

Xu-Yan HU, Yue-Xian ZOU, Wei SHI*

ELIP/ADSPLAB, School of Electronic and Computer Engineering, Peking University, Shenzhen, China
*zouyx@pkusz.edu.cn

ABSTRACT

[1]It is a challenge task for maintaining high correct word accuracy rate (WAR) for state-of-art automatic speech recognition (ASR) systems when the SNR goes very low. To deal with such situation, the missing feature technology (MFT) has shown as one of the mainstream algorithms. In principle, conventional MFT firstly separate the unreliable spectral bins from the reliable ones. Then the unreliable bins are reconstructed by missing feature algorithm [7]. When SNR goes low, the performance of the conventional MFT for ASR system is limited since both the reliable and unreliable spectral bins will be corrupted by the noise components. In this paper, a novel missing feature compensation method was developed by considering compensating both unreliable and reliable spectral bins. With the assumption of GMM distribution of the clean speech spectral vector, a dual MFT (DMFT) algorithm is developed, where the reliable spectral bins corrupted by noise have been compensated by removing the noise components. Several experiments have been carried out to evaluate the performance of the proposed DMFT algorithm by using AURORA2 database. From the results, it is clear to see that the proposed DMFT algorithm improves the WAR under all types of noises at different SNR levels compared with the traditional MFT algorithm.

Index Terms**—**speech recognition, missing feature technology, feature compensation, noisy environment, GMM

1. INTRODUCTION

Automatic Speech Recognition (ASR) has gained very wide applications in the wireless mobile communication industry currently due to the fast development of speech technologies. The word accuracy rate (WAR) of the ASR system is able to achieve over 95 percent in the clean environment. However, under noisy condition, the WAR declines drastically due to the mismatch between the training model and the testing speech. Literature study shows that there are many methods have been developed to improve the robustness of the ASR systems in noisy environment, which mainly include the Missing Feature Technology (MFT) [1], Perceptual Linear Predictive Relative Spectral (RASTA-PLP) method [2], the Parallel Model Combination (PMC) method [3] and the Vector Taylor Series (VTS) method [4].

Recent research outcomes have shown that the missing feature technology (MFT) achieved great improvement of the WAR performance of the ASR system, especially in low SNR and non-stationary noisy condition compared with other techniques [1],[7],[8]. In principle, the development of the MFT is based on the observations that the speech signals have a high degree of redundancy and the human listeners are able to comprehend speech that are partly missing [7]. Specifically, the traditional framework of the MFT consists of two main steps: the mask estimation and the missing feature reconstruction.

In the first step, the unreliable spectral bins in the log Mel-spectral (LMS) domain, which are dominated by the noise, are estimated, and the binary mask to remove the unreliable spectral bins and keep the reliable spectral bins is generated [1],[6]. Under MFT framework, a local SNR based binary mask is firstly estimated from the noisy speech, as follows. Let (i,ω) represent one bin in the LMS domain**,** $SNR(i,\omega)$ denote the local SNR at (i,ω), the mask at (i,ω) is defined as:

$$M(i,\omega)=\begin{cases}0 & SNR(i,\omega)<T \quad (i,\omega)\in R_u \\ 1 & SNR(i,\omega)\geq T \quad (i,\omega)\in R_r\end{cases} \tag{1}$$

where T is the SNR threshold which is often set according to the specific condition. R_u and R_r is the reliable and unreliable bin set respectively. Hence, with this binary mask, the reliable spectral bins and the unreliable ones can be extracted separately.

The second step is the missing feature compensation (MFC), which aims at reconstructing the unreliable spectral bins by the estimated reliable spectral bins with the aid of prior knowledge of clean speech provided by the training data. It is clear that the missing feature compensation technology tries to remove the adverse impacts of the noise in the feature extraction level. Hence, the acoustic model trained by the clean data can be kept unchanged although the testing speech is corrupted by noise.

Several effective missing feature compensation (MFC) algorithms have been proposed, such as the cluster-based feature compensation method (CBFC) [7], the correlation-based approach [8] and the sparse-coding based method [9]. For the existing MFC methods, we note that the extracted reliable spectral bins are left unchanged. However, at low SNR condition, the extracted reliable spectral bins are simultaneously corrupted by noise and we believe these noise components in the reliable spectral bins also will contribute to the degradation of WAR. To verify our concerns, we have conducted the following experiments.

Let $S_N(i,\omega)$ and $S_C(i,\omega)$ represent the log Mel-spectral (LMS) of the noisy and clean utterance at (i,ω), respectively. To measure the distance between the corrupted and the clean log Mel-spectral of reliable bins in one utterance, we define:

$$D_1=\sum_{i=1}^{M}\sum_{\omega=1}^{D}\mathrm{abs}[S_N(i,\omega)-S_C(i,\omega)]*M(i,\omega)/\mathrm{abs}(S_C(i,\omega)) \tag{2}$$

where M is the number of the frames used and D is the number of the filterbank channels for computing log-Mel spectral.

[1] This work is partially supported by National Natural Science Foundation of China (No: 61271309) and the Shenzhen Science & Technology Fundamental Research Program (No: JC201105170727A)

978-1-4799-5403-2/14/$31.00 ©2014 IEEE

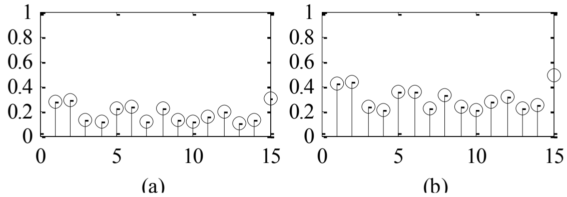

Figure 1: (a) D_1 of fifteen utterances under subway noise condition at 20 dB SNR level (b) D_1 of the same fifteen utterances under subway noise condition at 0dB SNR level.

In Figure 1, D_1 of fifteen randomly selected utterances corrupted by subway noise at 20dB and 0dB is presented. From Figure 1, it is clear to see that D_1 will increase when the SNR goes lower, which indicates that the reliable bins are also corrupted and require to be compensated, especially at the low SNR condition. Compensating the reliable spectral bins will help to reduce D_1 and hence improve the WAR performance of ASR at low SNR noisy condition.

In this paper, we will develop the solution to compensate both reliable and unreliable spectral bins. We propose a dual missing feature compensation technique (DMFT) to reconstruct the reliable spectral bins as well as the unreliable ones in two stages. At the first stage, the reliable bins in a log-Mel spectral vector are compensated based the method proposed in Section 3.1. Next, the unreliable spectral bins in the vector are reconstructed based on the compensated reliable bins in the same vector using the CBFC method proposed in [7]. The re-estimation of the reliable spectral bins can alleviate mismatch between clean log-Mel spectral and the noisy one and thus lead to improvement of the whole ASR system performance. Intensive experiments have been carried out to evaluate the performance of the proposed algorithm on the AURORA2 database. The results are compared to those of the classical cluster-based feature compensation method CBFC [7].

2. ASR SYSTEM DESCRIPTION

In our work, the architecture of the proposed dual missing feature compensation technique (DMFT) ASR system is presented in Figure 2. The design of the ASR system follows the famous HMM-based architecture with Gaussian mixture acoustic models [11]. Specifically, the log-Mel spectrum vectors (LMSVs) and MFCC are determined according to the method proposed in [1].

In the training procedure, the LMSVs of clean speech signals are modelled as GMM distribution and the GMM parameters are obtained by EM method [12]. The HMM models are trained using MFCC as input feature by Baulm-Welch algorithm [10].

As discussed before, in the existing MFT methods, only the unreliable bins have been compensated (or reconstructed) using CBFC [7] but leave the reliable spectral bins unchanged. In our proposed system, as shown in Figure 2, the reliable spectral bins also have been compensated to further reduce the adverse impacts of the noise. The details of our proposed method are described in the following section.

3. THE PROPOSED DUAL MISSING FEATURE COMPENSATION TECHNQIUE (DMFT)

To differentiate our proposed method from the existing MFT methods, we termed our method as dual missing feature compensation technique since we not only reconstruct the unreliable spectral bins but also compensate the reliable ones. It is obvious that different compensation methods should be applied to the reliable spectral and unreliable spectral bins since they have totally different properties. The proposed compensation method for reliable spectral bins will be described in subsection 3.1. The reconstruction method proposed in [7] for unreliable ones will be briefly presented in subsection 3.2 for presentation completeness.

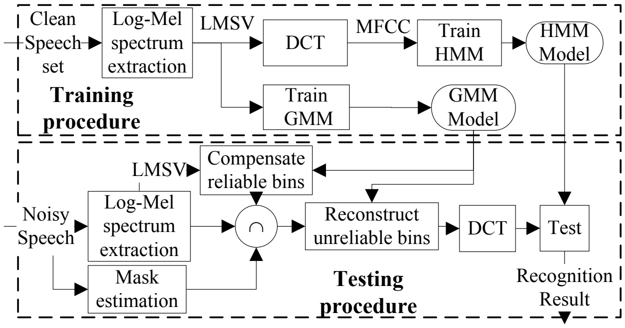

Figure 2: The DMFT speech recognition system architecture

3.1. The compensation method for reliable spectral bins

Let's define **Y**, **X** and **N** represent the $D\times1$ Mel spectral vector of a frame of noisy speech, clean speech and noise, respectively. Since the input speech is corrupted by uncorrelated additive noise, we have the following relation:

$$|\mathbf{Y}|^2 = |\mathbf{X}|^2 + |\mathbf{N}|^2 \tag{3}$$

Taking logarithm on Eq.(3), we have:

$$\log|\mathbf{Y}|^2 = \log(1+|\mathbf{N}|^2/|\mathbf{X}|^2) + \log|\mathbf{X}|^2 \tag{4}$$

To make the presentation clear, let's define $\mathbf{y}=\log|\mathbf{Y}|^2$ (the LMSV of the noisy speech,), $\mathbf{x}=\log|\mathbf{X}|^2$ (the LMSV of the clean speech) and $\mathbf{n}=\log|\mathbf{N}|^2$ (the LMSV of the additive noise). With some manipulations, Eq.(4) can be derived as:

$$\mathbf{y} = \mathbf{x} + \log(1+\exp(\mathbf{n}-\mathbf{x})) \tag{5}$$

Moreover, we define:

$$f(\mathbf{x},\mathbf{n}) = \log(1+\exp(\mathbf{n}-\mathbf{x})) \tag{6}$$

then Eq.(5) becomes:

$$\mathbf{y} = \mathbf{x} + f(\mathbf{x},\mathbf{n}) \tag{7}$$

By expanding Taylor series of the second term in (7) at an initial point ($\mathbf{x}_0$, $\mathbf{n}_0$), which is randomly selected, and taking only up to the first-order Taylor Vector series, Eq. (7) can be approximated as:

$$\mathbf{y} \approx \mathbf{x} + \mathbf{x}\nabla_{\mathbf{x}} f(\mathbf{x}_0,\mathbf{n}_0) + \mathbf{n}\nabla_{\mathbf{n}} f(\mathbf{x}_0,\mathbf{n}_0) + g(\mathbf{x}_0,\mathbf{n}_0) \tag{8}$$

where the functions $\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{n})$ and $\nabla_{\mathbf{n}} f(\mathbf{x},\mathbf{n})$ represent the partial derivative of $f(\mathbf{x},\mathbf{n})$ with respect of **x** and **n**, respectively:
The function $g(\mathbf{x},\mathbf{n})$ is defined as follows:

$$g(\mathbf{x},\mathbf{n}) = -\mathbf{x}\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{n}) - \mathbf{n}\nabla_{\mathbf{n}} f(\mathbf{x},\mathbf{n}) + f(\mathbf{x},\mathbf{n}) \tag{9}$$

As described before, **n** is modeled as a single Gaussian distribution, mathematically, we have:

$$\mathbf{n} \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Theta}_n), \mathbf{n} \in R^D, \boldsymbol{\mu}_n \in R^D, \boldsymbol{\Theta}_n \in R^{D\times D} \tag{10}$$

where $\boldsymbol{\mu}_n$ and $\boldsymbol{\Theta}_n$ denotes the mean vector and the covariance matrix of the Gaussian noise, respectively. D denotes the dimension of the LMSV.

Moreover, the LMSVs of clean speech are molded as a mixture Gaussian distribution, then the prior probability of **x** can be denoted as w_k:

$$P(\mathbf{x}) = \sum_{k=1}^{Q} w_k P(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Theta}_k \mid k) \tag{11}$$

where $\boldsymbol{\mu}_k$, $\boldsymbol{\Theta}_k$ and w_k are the mean vector, covariance matrix and the priori probability of the k-th component of the GMM. Q is the number of Gaussian components.

Therefore, according to the central limit theorem, **y** also can be modelled as a mixture Gaussian distribution given as:

$$P(\mathbf{y}) = \sum_{k=1}^{Q} w_k P(\mathbf{y}; \bar{\boldsymbol{\mu}}_k, \bar{\boldsymbol{\Theta}}_k \mid k) \qquad (12)$$

where $\bar{\boldsymbol{\mu}}_k$, $\bar{\boldsymbol{\Theta}}_k$ and w_k are the mean vector, covariance matrix and the priori probability of the *k*-th component of the GMM of the noisy speech. The priori probability of the noisy speech GMM is assumed to be the same with that of the clean speech GMM.

With the MMSE criteria, the estimation of **x** from **y** can be derived as follows (More details can be referred to [4]):

$$\begin{aligned}\bar{\mathbf{x}} &= E\{\mathbf{x} \mid \mathbf{y}\} = \int_X \mathbf{x} p(\mathbf{x} \mid \mathbf{y}) d\mathbf{x} \\ &= \int_X \sum_{k=1}^{Q} [\mathbf{y} - c(\mathbf{x}, \mathbf{n}, \mathbf{x}_0, \mathbf{n}_0)] p(\mathbf{x}, \mathrm{k} \mid \mathbf{y}) d\mathbf{x} \qquad (13) \\ &\approx \mathbf{y} - \sum_{k=1}^{Q} \mathbf{c} p(k \mid \mathbf{y})\end{aligned}$$

In Eq.(13), $c(\mathbf{x}, \mathbf{n}, \mathbf{x}_0, \mathbf{n}_0)$ and **c** is defined as [4]:

$$\mathbf{c} = c(\mathbf{x}, \mathbf{n}, \mathbf{x}_0, \mathbf{n}_0) = \mathbf{x} \nabla_{\mathbf{x}} f(\mathbf{x}_0, \mathbf{n}_0) + \mathbf{n} \nabla_{\mathbf{n}} f(\mathbf{x}_0, \mathbf{n}_0) + g(\mathbf{x}_0, \mathbf{n}_0) \qquad (14)$$

and $p(k|\mathbf{y})$ represents the posterior probability that **y** belongs to the *k*-th Gaussian component given as [4]:

$$p(k \mid \mathbf{y}) = w_k P(\mathbf{y}_r \mid \boldsymbol{\mu}_k, \boldsymbol{\Theta}_k) \Big/ \sum_{k=1}^{Q} w_k P(\mathbf{y}_r \mid \boldsymbol{\mu}_k, \boldsymbol{\Theta}_k) \qquad (15)$$

By taking expectation of both sides of Eq.(8), we can get:

$$\bar{\boldsymbol{\mu}}_k = (\nabla_{\mathbf{x}} \mathbf{f} + 1) \boldsymbol{\mu}_k + \boldsymbol{\mu}_n \nabla_{\mathbf{n}} \mathbf{f} + g(\boldsymbol{\mu}_k, \boldsymbol{\mu}_n) \qquad (16)$$

In order to compute $\boldsymbol{\mu}_n$, we need to estimate the Gaussian parameters of the background noise $\boldsymbol{\mu}_n$ and $\boldsymbol{\Theta}_n$ when a set of **y** is given. This estimation can be done iteratively using Maximum likelihood (ML) method. Specifically, the likelihood function is:

$$h(\mathbf{y}_1, \cdots, \mathbf{y}_M \mid \lambda_n^{(t)}, \lambda_n^{(t+1)}) = \sum_{k=1}^{Q} p(k \mid \mathbf{y}_i, \lambda_n^{(t)}) \log[(\prod_{i=1}^{M} p(\mathbf{y}_i \mid \lambda_n^{(t+1)})] \qquad (17)$$

where $\mathbf{y}_i$ is the LMSV extracted for the *i*-th frame noisy input. The parameter set $\lambda_n^{(t)} = \{\boldsymbol{\mu}_n^t, \boldsymbol{\Theta}_n^t\}$ describe the the Gaussian distribution of the noise in the *t*-th step and $\lambda_n^{(t+1)}$ describe the Gaussian parameter vector of the noise in the (*t*+1)-th step respectively in the ML estimation procedure. The initial value for $\boldsymbol{\mu}_n^0$ is estimated from the first several noise frames of the input. By taking derivation of Eq.(17) in respect of $\boldsymbol{\mu}_n$ and set derivation to zero, we can get:

$$\begin{aligned}&\boldsymbol{\mu}_n^{(t+1)} = [M \sum_{k=1}^{Q} \nabla_{\mathbf{n}} f(\boldsymbol{\mu}_k, \boldsymbol{\mu}_n^{(t)}) g(\boldsymbol{\mu}_k, \boldsymbol{\mu}_n^{(t)})]^{-1} \\ &\sum_{k=1}^{Q} \sum_{i=1}^{M} [p(k \mid \mathbf{y}_i, \lambda_n^{(t)}) \mathbf{y}_i - \boldsymbol{\mu}_k (\nabla_{\mathbf{x}} f(\boldsymbol{\mu}_k, \boldsymbol{\mu}_n^{(t)}) + 1) - g(\boldsymbol{\mu}_k, \boldsymbol{\mu}_n^{(t)})]\end{aligned} \qquad (18)$$

When Eq.(18) converges, $\boldsymbol{\mu}_n$ is estimated.

To estimate its clean LMSV of a specified noisy LMSV **y** in Eq.(13), by replacing $\mathbf{x}_0$ with **x**, and $\mathbf{n}_0$ with its mean value estimated by Eq.(18), Eq.(13) can be approximated as:

$$\bar{\mathbf{x}} = \mathbf{y} - \sum_{k=1}^{Q} p(k \mid \mathbf{y}) [\boldsymbol{\mu}_k \nabla_{\mathbf{x}} f(\boldsymbol{\mu}_k, \boldsymbol{\mu}_n) + \boldsymbol{\mu}_n \nabla_{\mathbf{n}} f(\boldsymbol{\mu}_k, \boldsymbol{\mu}_n) + g(\boldsymbol{\mu}_k, \boldsymbol{\mu}_n)] \qquad (19)$$

To make the presentation clear, let's assume *D*-by-1 vector **y** consists of the reliable spectral bins and unreliable bins, denoted as sub-vectors $\mathbf{y}_r$ and $\mathbf{y}_u$, respectively. Correspondingly, let $\mathbf{x}_r$, $\mathbf{x}_u$ denote the counterparts of $\mathbf{y}_r$ and $\mathbf{y}_u$ in **x** (LMSV of clean speech), respectively. Since $\bar{\mathbf{x}}$ is the estimated LMSV of clean speech **x**, and $\bar{\mathbf{x}}$ also consists of reliable and unreliable spectral bins, named as sub-vectors $\bar{\mathbf{x}}_r$ and $\bar{\mathbf{x}}_u$. Let's define $\mathbf{c}_r$ as the counterparts of $\mathbf{y}_r$ in **c** (computed in Eq.(14)), then $\bar{\mathbf{x}}_r$ is computed as:

$$\bar{\mathbf{x}}_r = \mathbf{y}_r - \sum_{k=1}^{Q} \mathbf{c}_r p(k \mid \mathbf{y}) \qquad (20)$$

3.2. Reconstruction of unreliable spectral bins

As presented in Eq.(19), we have obtained the optimal MMSE estimation of the LMSVs for the clean speech by giving the testing noisy speech. In [7], the reconstruction of the unreliable bins is completed by using the prior knowledge of the GMM model of all clean speech. In our proposed solution, we will use the compensated reliable spectral bins given in Eq.(20) to help to reconstruct the unreliable bins.

Motivated by the cluster-based feature compensation (CBFC) [7], PDFs of partly missing components are used. The posterior probability that the noisy speech LMSV **y** belongs to the *k*-th Gaussian component is determined by [7] :

$$P(k \mid \mathbf{y}) = w_k P(\bar{\mathbf{x}}_r, \mathbf{x}_u \le \mathbf{y}_u \mid \boldsymbol{\mu}_k, \boldsymbol{\Theta}_k) \Big/ \sum_{k=1}^{Q} w_k P(\bar{\mathbf{x}}_r, \mathbf{x}_u \le \mathbf{y}_u \mid \boldsymbol{\mu}_k, \boldsymbol{\Theta}_k) \qquad (21)$$

After obtaining the posterior probability of the *k*-th Gaussian component, the clean estimates of $\mathbf{y}_u$ using the *k*-th Gaussian component is calculated as:

$$\bar{\mathbf{x}}_u^k = \underset{\mathbf{x}_u}{\arg\max}\, p(\mathbf{x}_k \mid k, \bar{\mathbf{x}}_r, \mathbf{x}_u \le \mathbf{y}_u) \qquad (22)$$

Eq.(22) can be solved using iterative Bounded Maximum a posterior (BMAP) procedure described in [7]. $\bar{\mathbf{x}}_u$ is computed as:

$$\bar{\mathbf{x}}_u = \sum_{k=1}^{Q} P(k \mid \mathbf{y}) \bar{\mathbf{x}}_u^k \qquad (23)$$

3.3. The proposed DMFT Algorithm

To make the presentation clear, the proposed DMFT algorithm is summarized in Table 1 The proposed DMFT algorithmTable 1.

Table 1 The proposed DMFT algorithm

	For each utterance of noisy speech
1.	Convert the speech signal into LMSVs
2.	Estimate spectral mask or calculate oracle mask
3.	Calculate the initial value of $\boldsymbol{\mu}_n^0$
4.	Estimate $\boldsymbol{\mu}_n$ according to Eq.(18)
5.	Re-estimate $\boldsymbol{\mu}_k$ according to Eq.(16)
6.	Redo Step.4 to Step.5 until $\boldsymbol{\mu}_n$ converged
7.	Compute $\bar{\mathbf{x}}_r$ referring Eq.(19)
8.	Compute $\bar{\mathbf{x}}_u$ using CBFC method according to Eq.(23)

4. EXPERIMENTS AND ANALYSIS

In our experiments, the AURORA2 database [13] was used. We formed the training data set by using 8440 clean utterances from 55 male and 55 female adult speakers. Then, the acoustic HMM models are trained using this training data set. Speech data in testing sets A and testing sets B (termed as Sets A and Sets B in short) are used for evaluation. There are eight types of background noise in the AURORA2 database. In Sets A, the noise includes subway, babble, car and exhibition noise. In Sets B, the noise includes the restaurant, street, airport and station noise. Noisy speech data set are generated by artificially adding the noise data at a variety of SNR levels.

A conventional 36-dimensional Mel-frequency cepstral coefficient (MFCC) feature vector is used. The specific parameters used to compute MFCC vector is as follows: The number of Mel-filterbanks is 23. The analysis window is of 25ms duration and 10ms step rate. The number of cepstrum coefficients is 12 (i.e., c1–c12). The first and second-order time derivatives of the cepstrum coefficients are used.

We employed the toolbox HTK [7] to train the HMMs. Each HMM represents a word consisting of 16 states with 6 Gaussian components per state. The LMSVs of the clean speech is modeled as a GMM with 32-components and EM algorithm is used to obtain the GMM parameters. The first and the last 5 frames of the input are used to estimate the initial value of $\boldsymbol{\mu}_n^0$ in Eq.(17).

The word accuracy rate (WAR) is taken as the performance measure. The performance of the baseline system [10] and the CBFC method [7] is compared. The experiment settings of baseline system and CBFC are exactly same as DMFT.

Experiment 1: The WAR performance under different SNR levels. This experiment aims at evaluating and comparing the WAR performance of the MFT methods with two binary masks, which are oracle mask [8] and SS-mask [6]. It is noted that the oracle mask is obtained by assuming the knowledge of the clean speech is exactly known, therefore, it is impractical and we take it as reference. The SS-mask is extracted by the spectral subtraction method [6]. According to the extraction of binary mask in Eq.(1), the SNR threshold *T* is set to 0dB when oracle masks are computed and -6dB when SS-masks are estimated. Besides, six SNR levels are evaluated which are clean, 20dB, 15dB, 10dB, 5dB, 0dB and -5dB respectively. For each SNR level and each type of noise in Sets A and Sets B, there are 1001 utterances from 52 male and 52 female adult speakers. Experimental results of are given in Table 2.

From Table 2, it is clear to see that the proposed DMFT method with oracle mask outperforms CBFC over different SNR under noise conditions in both Set-A and Set-B, especially when SNR is below 10dB. Under noise conditions in Set-A, when the SNR is higher than 10dB, the proposed DMFT outperforms CBFC by 1.61% on average, when the SNR goes low, DMFT outperforms CBFC by over 5%, especially at low SNR and non-stationary noise. Moreover, it is also clear to see the impact of the binary mask estimation where the SS-mask degrades the WAR for all SNR levels compared with those using oracle masks. The best WARs for different SNR levels are highlighted. From Table 2, we also clearly see the impact of the binary mask. The same algorithms with different masks gave different performance. Specifically, when SNR goes lower than 10dB, the impact of the noise is bigger. The proposed DMFT obtains a more significant improvement over the CBFC when adopting the estimated SS-masks. When SNR is below 10dB, the DMFT outperforms CBFC by over 20%. Under the noise condition in Sets-B, the performance is similar. However, the performance improvement of DMFT over CBFC is smaller than that under noise conditions from Sets-A, which indicates the WAR performance using MFT depends on the type of noise. These results in Table 2 validate the necessity of compensating the reliable spectral bins and the effectiveness of our proposed method. Moreover, all algorithms outperform the baseline system, which reflects the effectiveness of the MFT.

Experiment 2: The WAR performances of the proposed DMFT method. This experiment is carried out purposely to further evaluate the proposed DMFT method with oracle mask and SS-mask under 8 types of the noise at different SNR levels. The simulation parameters are the same as those used in Experiment 1. The simulation results are listed in Table 3. Comparing the results in Table 3, we can see clearly that the WAR performance of the proposed DMFT method with SS-masks is inferior to that with the oracle mask. However, carefully evaluating the results, we found the WAR difference between the proposed DMFT method with the oracle mask and SS-mask is small when SNR level larger than 10dB (about 0.35% to 1.74%), but the difference goes larger when SNR is less than 10dB. It can be seen that the largest WAR difference is 10.96%. The extraction of the good binary mask should be further studied.

Table 2: WAR performance averaged across 4 different types of noise in Sets A and Sets B of the AURORA2 database

WAR (Sets-A)						
Method	20dB	15dB	10dB	5dB	0dB	-5dB
Baseline[10]	96.75	91.77	74.54	41.79	22.04	12.09
CBFC [7]	98.19	96.53	91.96	80.56	59.87	44.24
DMFT	**98.37**	**97.53**	**95.61**	**89.82**	**74.65**	**47.11**
CBFC-SS[7]	96.91	93.19	80.4	56.53	35.04	20.88
DMFT-SS	97.88	96.98	94.05	86.80	68.93	42.20
(WAR) Sets-B						
Method	20dB	15dB	10dB	5dB	0dB	-5dB
Baseline[10]	96.855	92.643	83.633	51.165	25.173	12.033
CBFC [7]	98.05	97.76	95.09	86.80	62.10	46.05
DMFT	**98.51**	**98.24**	**96.58**	**92.93**	**80.13**	**53.26**
CBFC-SS[7]	96.77	96.32	81.6	59.87	39.8	24.97
DMFT-SS	97.79	97.88	95.73	88.78	69.42	43.01

Table 3: WAR (%) for DMFT with oracle mask and SS-mask for 8 types of noise in Sets-A and Sets- B of the AURORA2 database ((1) Subway (2) Babble (3) Car (4) exhibition (5) Restaurant (6) Street (7) Airport (8) Train Station)

Noise type	Mask	20dB	15dB	10dB	5dB	0dB	-5dB
(1)	Oracle	98.28	97.36	94.87	87.75	74.76	51.24
	SS	**97.73**	**96.61**	**93.76**	**84.66**	**69.19**	**45.98**
(2)	Oracle	98.58	97.82	96.95	93.23	79.9	51.18
	SS	**98.23**	**97.22**	**95.21**	**89.99**	**73.84**	**46.33**
(3)	Oracle	98.63	97.91	95.94	89.5	71.67	39.1
	SS	**98.25**	**97.33**	**94.20**	**86.61**	**65.64**	**34.08**
(4)	Oracle	97.99	97.04	94.69	88.8	72.26	46.9
	SS	**97.29**	**96.75**	**93.03**	**85.93**	**67.06**	**42.39**
(5)	Oracle	98.56	98.53	97.7	94.29	84.74	60.7
	SS	**98.03**	**97.92**	**96.92**	**89.61**	**74.14**	**49.74**
(6)	Oracle	98.04	97.34	94.95	89.24	74.15	46.83
	SS	**97.28**	**97.18**	**93.73**	**85.31**	**63.41**	**36.88**
(7)	Oracle	98.78	98.81	97.76,	95.32	84.67	58.28
	SS	**98.18**	**98.63**	**96.66**	**91.34**	**74.05**	**48.40**
(8)	Oracle	98.67	98.27	97.1	92.87	76.95	47.21
	SS	**97.66**	**97.77**	**95.61**	**88.85**	**66.07**	**37.00**

In conclusion, from the simulation results using the AURORA2 database, we are confident that, under MFT framework, the proposed DMFT method is able to further improve the WAR performance of speech recognition system by compensating the reliable spectral bins as well as the unreliable spectral bins, especially under low SNRs and non-stationary noise environment.

5. CONCLUSION

In this paper, we proposed a new approach that not only reconstruct the unreliable spectral bins but also compensate the reliable bins under the missing feature compensation technique for improving the performance of the ASR system. For compensating the reliable spectral bins, an algorithm is derived with the assumption that clean speech is modeled as GMM distribution. Intensive experiments have been conducted to evaluate the WAR performance. An increase of WAR above 10% over that of CBFC [7] at SNR below 10dB shows that the proposed DMFT method is effective in further improving WAR performance of the ASR system in non-stationary and low-SNR noisy conditions.

REFERENCE

[1] Cooke, Martin, Phil D. Green, and Malcolm Crawford. "Handling missing data in speech recognition." *Proc. ICSLP*, pp. 1555 -1558, 1994.

[2] Hermansky H. "Perceptual linear predictive (PLP) analysis of speech." *The Journal of the Acoustical Society of America*, 87, pp. 1738 -1752, 1990.

[3] Gales, Mark JF, and Stephen J. Young. "Robust continuous speech recognition using parallel model combination." *Speech and Audio Processing, IEEE Transactions on,* 4(5) , pp. 352-359,1996.

[4] Moreno, Pedro J., Raj, Bhiksha, and Richard M. Stern. "A vector Taylor series approach for environment-independent speech recognition." *Proc. ICASSP*, 2, pp. 733 -736, 1996.

[5] Raj, Bhiksha, R. Singh, and Richard M. Stern. "Inference of missing spectrographic features for robust speech recognition." *Proc. ICSLP*, 98, pp. 1491-1494, 1998.

[6] Vizinho, Ascension, et al. "Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: an integrated study." *Proc. Eurospeech*, 99, pp. 2407-2410, 1999.

[7] Raj, Bhiksha, Michael L. Seltzer, and Richard M. Stern. "Reconstruction of missing features for robust speech recognition." *Speech Communication*, 43(4), pp. 275-296, 2004.

[8] Raj, Bhiksha, and Richard M. Stern. "Missing-feature approaches in speech recognition." *Signal Processing Magazine*, 22(5), pp.101 - 116, 2005.

[9] Gemmeke, Jort F., and Tuomas Virtanen. "Noise robust exemplar-based connected digit recognition.", *Proc. ICASSP*, pp.4546 -4549, 2010

[10] Rabiner, Lawrence. "A tutorial on hidden Markov models and selected applications in speech recognition." *Proceedings of the IEEE* 77.2, pp. 257-286, 1989.

[11] Young, Steve, et al. "The HTK book (for HTK version 3.4)." *Cambridge university engineering department* 2.2,2006:

[12] Bailey, Timothy L., and Charles Elkan. "Fitting a mixture model by expectation maximization to discover motifs in bipolymers." , pp. 28-36, 1994.

[13] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condidions," in *ISCA ITRW ASR 2000 Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, September 2000.