Spectral-Domain Speech Enhancement for Speech Recognition

Chang Huai YOU, Bin MA
Institute for Infocomm Research, A*STAR, Singapore
Email: {echyou, mabin}@i2r.a-star.edu.sg

***Abstract*—Speech recognition performance deteriorates in face of unknown noise. Speech enhancement offers a solution by reducing the noise in speech at runtime. However, it also introduces artificial distortion to the speech signal. In this paper, we aim at reducing the artifacts that have adverse effects on speech recognition. With this motivation, we propose a modification scheme including a smoothing adaptation to frame signal-to-noise ratio (SNR) and a reestimation of *a priori* SNR for spectral-domain speech enhancement. The experiment shows that the proposed scheme of enhancement significantly improves the performance of the state-of-the-art speech recognition over the baseline speech enhancement techniques.**

Index Terms: speech enhancement, speech recognition, *a priori* SNR

I. Introduction

State-of-the-art automatic speech recognition (ASR) system works well under clean environmental situation. However, it has been observed that the performance of the speech recognition system degrades rapidly in the presence of noise or other distortions [1]. Over the past few decades much research has been devoted to improving the robustness of speech recognition in noisy environment [2] [3] [4].

The presence of noise at runtime introduces a mismatch between the training condition and test condition.

In practice, one of the solutions is the multi-conditional modeling which trains the acoustic model with various noisy databases to cover different kinds of noise environment. Multi-conditional training have proven the great advanced for noisy situation. It is a straightforward way to achieving noise robustness, but it is also known to suffer from a lack of generalisability to unseen conditions and a reduced performance of recognition on high-SNR speech. In other words, such technique fails in face of unknown noise condition. In particular, the multi-condition cannot improve the recognition accuracy while the trained model encounters unseen noise. An alternative to overcome unknown noise condition is to train the acoustic models on clean speech data and apply speech enhancement techniques to improve the runtime speech quality under noise condition [5] [6]. Moreover, the clean speech model always gives better WER performance to clean speech recognition than multi-conditional model does. With the speech enhancement solution, one can focus on developing a high quality clean acoustic model, a sharper model than a multi-condition acoustic model.

To understand the artifacts introduced by speech enhancement, and their effects on speech recognition system, we are interested in looking into various speech enhancement methods. In practice, it is always difficult to reduce noise without introducing the speech distortion due to the random nature of noise and the inherent complexity of speech signal. It has been a fact that the artifacts will be introduced into the speech signal as the noise is reduced in the speech signal. Thus, it is necessary to consider the tradeoff between noise reduction and speech distortion in speech enhancement [7].

Among the most effective enhancement techniques in the past decades, the popular ones include spectral-domain denoising [8] [9] [10] [11] [12] [13], speech production modeling [14] [15], human auditory perceptual criterion [16] [17] [18], the probability of speech presence uncertainty [19] [20], subspace decomposition [21], and the combinations of the above techniques [22].

ASR speech enhancement aims to improve the quality of noisy speech input at runtime to reduce the mismatch with the trained acoustic model. In 1991, Hanson and Clements introduced a constrained iterative enhancement for speech recognition [23], where an iterative Wiener filtering with vocal tract spectral constraints was formulated using inter-frame and intraframe constraints based on line spectral pair transformation. The enhancement approach with interframe constraints ensures more speech-like formant trajectories than those found in the unconstrained approach while the intraframe constraints ensure overall maximization of the speech quality across all classes of speech. The performance was evaluated using a standard, isolated-word recognition system. In 2006, Gemello et al proposed a modification of Ephraim-Malah log-spectral amplitude method by introducing an overestimation of noise power and an adjustment of spectral floor into *a prior*i SNR and *a posteriori* SNR with respect to frame SNR [24]. Significant improvement was reported for Aurora speech recognition system. In 2008, Breithaupt et al proposed a cepstral-domain smoothing method for estimation of *a priori* SNR [25], and the experiment that was done with Wiener filter shows improvement over conventional decision-directed approach. However, the effectiveness of the *a priori* SNR estimation method was only proven in terms of speech enhancement objective measurement but not proven in terms of speech recognition performance. In the same year, Yu et al applied the Ephraim-Malah minimum mean square error (MMSE) criterion into speech feature domain [26] instead of the discrete Fourier transform (DFT) domain for noisy speech recognition. The performance was investigated on the standard Aurora speech recognition platform [27]. In 2010, Paliwal et al investigated the role of speech enhancement in speech recognition [28] where the experiments were conducted on the TIMIT speech corpus, however, there was no any solution provided for the artificial distortion caused by the investigated speech estimators against the speech recognition; and also the speech recognition decoder was only based on small Gaussian mixture model-hidden Markov model (GMM-HMM) where only eight-Gaussian mixtures per state were applied and a bigram language model was used. All the above studies could not make a very clear impression on the goodness and effectiveness of the various enhancement methods in modern speech recognition system, since we observed a fact that the performance of speech enhancer also depends on the particular speech recognition models. In other words, the enhancer may be helpful for certain speech decoder but not always contribute to another speech decoder. For this reason, it is necessary to investigate the performance with typical state-of-the-art decoding platform.

In this paper, we study the spectral-domain speech enhancement and select three typical methods, including Wiener filtering [9], log-spectral amplitude (LSA) [11], and masking-based β-order (β-masking) MMSE [13] algorithms. In [28], Paliwal et al investigated sixteen speech enhancement methods for speech recognition, and gave a conclusion that the improvements in objective speech quality did not translate to the improvement of speech recognition; and an enhancer (with its default settings) that produced best objective speech quality gave a poor performance in speech recognition. Therefore, a speech enhancement algorithm may significantly improve human listening experience [39] [40], direct application of the enhancement algorithm does not always work well for speech recognition system. Classical objective quality measure based on global SNR or average segmental SNR over an utterance does not, in general, provide useful estimates of the perceived speech quality as well as the quality of machine recognition. In our observation, we also noticed that the speech enhancer that has good PESQ (perceptual evaluation of speech quality) performance for human listening brings poor word-error-ratio (WER) performance of speech recognition. The reason is the improvement of PESQ cannot be directly transferred into the improvement of feature distortion that directly affects the performance of speech recognition. So far, there is no any single statistics of the speech quality measure can completely transfer the similarity between the estimated speech and the reference (or clean) speech into the ultimate performance of machine recognition. However, the distortion measure of the feature sequence which is directly used as the input of modern speech recognition system can still represent a rough estimation of the distinction between the estimated speech quality and the clean speech quality for the speech recognition. We propose to improve the ASR speech enhancement systems by alleviating the feature distortion ratio for the purpose of the speech recognition in some aspects: the noise overestimation control, weak spectral component flooring, oversuppression of unwanted residual noise, and a reestimation of *a priori* SNR [41]. Firstly, by introducing smoothing adaptation with respect to frame SNR, we design a smoothing control of the power of the processing noise, show a way to process the weak spectral signal with a time-varying floor of spectral SNRs. Secondly, we develop an oversuppression of the residual noise with smoothing adaptation. Finally, we propose a reestimation of the *a priori* SNR and extend it to a possible iterative process. Experimental results show each and every of the modifications (i.e. the noise control, the weak spectral processing, the residual noise suppression and *a priori* SNR reestimation) are able to effectively improve the performance of the three typical speech enhancement systems in terms of WER.

In order to build up a meaningful investigation system, we setup a state-of-the-art evaluation platform which is reconstructible by open-source speech recognition tool. In particular, we use Kaldi toolkit [29] to build up a large vocabulary speech recognition system with a series of the training models that start from monophone, coarse triphone GMM-HMM to detailed triphone GMM-HMM, and then DNN-HMM which follows the pre-training of deep belief network (DBN). In this speech recognition system, cepstral mean and variance normalization (CMVN), linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT), feature space maximum likelihood linear regression (fMLLR) for speaker adaptive training and state-level minimum Bayes risk (sMBR) techniques are applied. We train the models in each steps by using the labelled clean speech, and measure the performance of recognition using clean, noisy and enhanced speech.

In the remainder of the paper, we give a brief introduction of the spectral-domain speech enhancement algorithms used in this paper in section II. In section III, we propose a series of modification schemes for the speech estimators against speech recognition. We describe the speech recognition platform for the performance evaluation of speech enhancement in section IV. The evaluation is shown in section V and finally the conclusion is given in section VI.

II. Spectral-Domain Speech Enhancement Algorithms

An observed noisy speech signal $x(t)$ is assumed to be a clean speech signal $s(t)$ degraded by uncorrelated additive noise $n(t)$, i.e.,

$$x(t) = s(t) + n(t), \quad 0 \leq t \leq T. \tag{1}$$

Let $S_k(l)$, $N_k(l)$, and $X_k(l)$ denote the kth spectral component of the clean speech signal $s(t)$, noise $n(t)$, and the observed noisy speech $x(t)$, respectively, where l denotes the time frame corresponding to time t in analysis interval [0, T]. The enhanced speech spectrum is given by $\hat{S}_k(l) = G_k(l)X_k(l)$, where $G_k(l)$ is the gain function of the enhancement.

A. Wiener Filtering

With Gaussian distribution assumption of the respective complex spectra of speech and noise, we seek to minimize Bayes risk with the expectation of cost function $C(\hat{S}_k(l), S_k(l))$ given observed signal X_k [30]

$$\begin{aligned} \mathbf{E}[C(\hat{S}_k(l), S_k(l))|X_k] &\propto \int_{X_k} |\hat{S}_k(l) - S_k(l)|^2 Y_k(l) dS_k(l), \\ Y_k(l) &= \exp\Big(- \frac{|X_k(l) - S_k(l)|^2}{\eta_n(k,l)} - \frac{|S_k(l)|^2}{\eta_s(k,l)} \Big) \end{aligned} \tag{2}$$

where $\eta_n(k,l) = \mathbf{E}[|N_k(l)|^2]$ and $\eta_s(k,l) = \mathbf{E}[|S_k(l)|^2]$ are the variances of the kth spectral components of noise and the speech signal, respectively. Consequently, we have the gain function for Wiener filter as shown below

$$G_k(l) = \frac{\xi_k(l)}{1+\xi_k(l)} \tag{3}$$

where $\xi_k(l)$ is the *a priori* SNR.

B. LSA-MMSE

Motivated by a fact that the correlation between the spectral components reduces when the analysis interval length increases, the statistical independence assumption is applied into the estimation of short term speech spectral amplitude. As a result, minimizing the mean square error of log spectral amplitude (LSA) equals $|\hat{S}_k(l)| = \exp\{\mathbf{E}[\ln|S_k(l)| \;/\; X_k]\}$. Consequently, Ephraim and Malah derived the gain function of the LSA-MMSE estimator [11]

$$G_k(l) = \frac{\xi_k(l)}{1+\xi_k(l)} \exp\left\{\frac{1}{2}\int_{\upsilon_k(l)}^{\infty} \frac{e^{-t}}{t} dt\right\} \tag{4}$$

where υ_k is given by

$$\upsilon_k(l) = \frac{\xi_k(l)}{1+\xi_k(l)}\gamma_k(l). \tag{5}$$

The definition of the *a priori* SNR ξ_k and *a posteriori* SNR γ_k is given as follows

$$\xi_k(l) = \frac{\eta_s(l,k)}{\eta_n(l,k)}, \quad \gamma_k(l) = \frac{|X_k(l)|^2}{\eta_n(l,k)}. \tag{6}$$

C. β-masking MMSE

The β-order MMSE speech enhancement method [12] is derived by minimizing the mean square error cost function $J = \mathbf{E}\{(|S_k|^\beta - |\hat{S}_k|^\beta)^2\}$ based on the complex Gaussian distribution model and statistical independence assumption. The gain function of the β-order MMSE expressed by [12]

$$G_k(l) = \frac{\sqrt{\upsilon_k}}{\gamma_k}[\Gamma(\frac{\beta}{2}+1)M(-\frac{\beta}{2};1;-\upsilon_k)]^{1/\beta} \tag{7}$$

where $\Gamma(\frac{\beta}{2}+1)$ is the gamma function and $M(-\frac{\beta}{2};1;-\upsilon_k)$ is the confluent hypergeometric function [10] [12]. Coincidentally, Ephraim-Malah E-M LSA [11] can be seen as a special case of β-order MMSE when $\beta \to 0$ [12].

In β-order MMSE, the value of β can be adapted to proper time-varying properties. In [13], the β-masking MMSE algorithm adapts the β value as follows

$$\begin{aligned}\hat{\beta}(l,k) &= 0.942 + 0.121\Xi(l) + 0.981\Theta_f(l,k)\\ &+ 0.187\max[\Xi(l)+6.7, \quad 0]\Theta_f(l,k).\end{aligned} \tag{8}$$

where $\Xi(l)$ is frame SNR. Θ_f is a normalized version of noise masking threshold at current frame, which represents the perceptual factor of the human auditory system in the frequency domain.

D. About Noise Estimation

Before investigating the speech estimators, we need to have an accurate estimation of the noise spectral variance. Speech enhancement does actually include two main estimation parts: the estimation of noise and the estimation of speech. The quality of estimated speech with the same speech estimator heavily depends on the accuracy of the estimate of the noise statistics. In contrast with the speech estimator that is to reconstruct every instantaneous sample of the speech signal, the noise estimator is not to restore the instantaneous noise spectral power, but only to estimate its expectation, i.e., the noise spectral variance. The main difficulty of noise estimation is due to the nonstationary characteristics of noise and the estimation of the background noise during speech activity. In 2001, Martin proposed to estimate the noise spectral statistics based on tracking the minimum of the noisy speech over a finite window [45]. This is based on a fact that the noisy speech spectral power is frequently reduced to the noise spectral power level during the period of non-speech spectrum or within brief periods in words and syllables. Obviously, using minimum values in a window of considerable length is able to prevent speech spectral power from leaking into the estimate of noise spectral variance. However, it causes a problem that it takes slightly more than the duration of the minimum-search window to update the noise spectrum when the noise floor increases abruptly. As the minimum is usually smaller than the mean, unbiased estimates of noise spectrum were considered with a bias factor based on the statistics of the minimum estimates. In 2006, Rangachari and Loizou introduced the speech presence probability (SPP) into a minimum searching and improved the performance in non-stationary background noise condition [46]. The introduction of SPP further reduces the amount of speech spectral power leaking into the estimates of noise spectral statistics. Recently, an MMSE-based noise estimation method has been reported to be effective on tracking the noise spectral power with short delay [47] [48]. This MMSE-based noise estimation can be interpreted as a voice activity detection (VAD)-based noise tracker when the *a priori* SNR is estimated by means of a limited maximum likelihood estimate. In [49], an improved version of the MMSE-based noise estimation is proposed by introducing SPP with a fixed *a priori* SNR constraint. The usage of the fixed *a priori* SPP leads to an unbiased estimation of the noise statistics, and it is of even lower computation complexity than that in [48]. Compared to minimum statistics noise estimation [45], the MMSE noise estimation improves the SNR and PESQ for non-stationary noise situation. Through many experiments, we observed that, for stationary noise, the minimum statistics [45] and the MMSE-based noise estimation [48] [46] show quite similar conclusion for the WER comparison of different speech enhancement algorithms. For low SNR situation, the WER performance of minimum statistics [45] is obviously better than that of MMSE-based [48]. The WER performance of SPP MMSE-based noise estimation [49] outperforms both minimum statistics [45] and MMSE-based noise estimation [48] in most of noise situation, especially for high SNR situation.

In this paper we only focus on the speech estimation study based on a reliable estimate of noise spectral power density. In the following experiment, we select to report the performance results based on the SPP MMSE noise estimation [49] applied on the reference noise in order to obtain a reliable estimate of noise spectral variance $\eta_n(l,k)$, so that we can have a precise comparison for different speech estimators in terms of speech recognition performance. The idea of selecting the reference noise instead of noisy speech is to avoid the interference from the speech signal leakage. With the progress of the noise estimation techniques which are of less or more drawbacks currently, the noise spectral variance estimation will be approaching to its perfection. We believe that, with the noise estimator applied on the reference noise in place of the noisy speech, the experimental result for the performance comparison among different speech estimators is of meaningful value [1].

III. Proposed Schemes: Artifact Mitigation and Suppression Control in the spectral-domain Enhancers for Speech Recognition

Usually speech enhancement reduces the noise at the expense of introducing artifacts in the form of spectral variation of original speech [7]. The spectral variation includes the changes of the statistical characteristics of the speech and the loss of some discriminating information embedded in the speech signal. The artifacts can cause a new mismatch that is harmful to recognition. Although a speech enhancement algorithm may significantly improve human listening experience [39] [40], direct application of the enhancement algorithm does not always work well for speech recognition.

To speech recognition, the artifacts caused by the enhancement processing is primarily transferred into a certain kind of distortion in feature vector which is used as input vector of speech recognition system. Here, we use a cepstral distortion ratio to represent the feature distortion by the following formula [2]

$$\begin{gathered}\psi_i = \frac{\sum_{l=1}^{J}[c_i^{(e)}(l) - c_i^{(c)}(l)]^2}{\sum_{l=1}^{J}[c_i^{(c)}(l)]^2}, \quad i = 1, ..., M; \\ E = \frac{1}{M}\sum_{i=1}^{M}\psi_i\end{gathered} \tag{9}$$

where $\vec{C}(l) = [c_0(l), c_1(l), ..., c_M(l)]$ is the feature coefficient vector at frame l in speech recognition, J is the number of feature frames, and $c_i^{(e)}(l)$ denotes the i-th feature coefficient of the l-th corrupted or estimated speech frame, and $c_i^{(c)}(l)$ denotes that of the clean speech frame. In this paper, $\vec{C}(l)$ is actually the 13-dimension Mel frequency cepstral coefficient (MFCC) vector with CMVN and VAD processing.

Since the SNRs of spectral component are the most critical parameters for the spectral-domain speech enhancer [25] [42], for the purpose of speech recognition, we attempt to improve the spectral-domain speech enhancement by alleviating the artifacts in some aspects: *a priori* SNR and *a posteriori* SNR estimation, deep suppression of unwanted residual noise, and reestimation of *a priori* SNR.

Considering the definition of *a priori* SNR by (6), together with a fact that the maximum likelihood estimate of $\xi_k(l)$ equals to $\gamma_k(l) - 1$, by introducing a smoothing factor α and replacing the current frame speech component with the estimate in its preceding frame, the conventional decision-directed estimation of the *a priori* SNR is given by [10]

$$\begin{aligned}&\hat{\xi}_k(l) = \\ &\alpha\frac{|G_k(l-1)X_k(l-1)|^2}{\eta_n(k,l)} + (1-\alpha)\max[\gamma_k(l) - 1, \quad 0].\end{aligned} \tag{10}$$

With the decision-directed approach to estimate the *a priori* SNR by (10) for subjective listening, the smoothing factor, α, is conventionally set to 0.98 [10] [12] [13] [38]. It has been reported that the speech estimators with α set to 0.98 for *a priori* SNR estimation results in a great reduction of the noise, and provides enhanced speech with colorless residual noise, which is found to be much less annoying and disturbing for human listening.

However, it has been observed in our experiment that speech recognition system favors a smaller α value. This could be due to a fact that the α indicates the memory of past frame information, a larger value of α means bringing more past frame information to the current frame. Different from human hearing system, a HMM-based ASR system is believed to favor non-overlapping frame information more than its past information. As a result, a very high value of α brings to the speech signal much harmful artifacts that the HMM system is not trained to accommodate.

By investigating the effect of different α value for speech recognition, we have a conclusion that the best performance of speech recognition is no long with 0.98 of α and the speech recognition reaches the best accuracy where α is in range of 0.7 $\sim$ 0.9.

A. Smoothing Adaptation to Frame SNR for Noise Control and Weak Spectral Floor

In MMSE estimation, the estimate of speech signal spectral amplitude A_k is based on modelling speech and noise spectral components as statistically independent Gaussian random variables [10]. The statistical independence assumption in the Gaussian model is equivalent to the assumption that the Fourier expansion coefficients are uncorrelated. Therefore the k-th spectral gain is only a function of *a priori* and *a posteriori* SNRs of k-th frequency bin rather than those of other frequency bins. In fact, a good estimate of a spectral amplitude is not only contributed from the information of the same frequency parameters but also from other frequency parameters. Great amount of observations has proven that the frame SNR is useful information contributed to the estimation of the speech amplitude [12] [13] [24]. In our previous work [12] [13] [43], the concept of frame SNR was introduced and

[1] Actually, we have also done many sets of experiments using the three different noise estimation methods including minimum statistics [45], MMSE noise estimation [48] and SPP MMSE noise estimation [49] applied on either noisy speech or reference noise, the observations of the WER comparison between different speech enhancement algorithms are almost consistent.

shown to be useful in spectral estimation. The definition of frame SNR is given as follows

$$\Xi(l) = 10\log_{10}\left\{\frac{\sum_k |S_k(l)|^2}{\sum_k |N_k(l)|^2}\right\}. \tag{11}$$

Obviously, the introduction of frame SNR breaks the limitation of the statistical independent assumption on the spectral estimation. In the real-time processing, as we only have the observed noisy speech signal, and we do not have the clean speech signal and noise signal, the frame SNR can be approximated by using the following equation [44]

$$\Xi(l) = 10\log_{10}\max\left\{\frac{\sum_k\left\{\max[(|X_k(l)| - \sqrt{\eta_n(l,k)}\ ),\quad 0]\right\}^2}{\sum_k \eta_n(l,k)},\ \varepsilon\right\} \tag{12}$$

where ε denotes a small positive number set to 2.22×10^{-16}.

Conventional relation to the frame SNR is just a very coarsely concatenation of straight lines [24] [43] [44]. In [12] [13] [43] [44], the frame SNR is introduced to adapt the value of β using broken linear relationship. However, the broken points without smoothing transitions cause harmful error on the estimation of the weak spectrum contaminated with noise. In this paper, we introduce a smoothing relationship by using sigmoid function to solve the broken junction problem.

Controlling the degree of the noise reduction is about making a trade-off between introducing enhancement artifacts and reducing noise so that the potential capability of the speech enhancement can be sufficiently developed for speech recognition. In this paper, we aim to find a suitable degree of noise suppression that achieves a proper effect without introducing much undesired mismatch to feature sequence in speech recognizer. This consideration motivates us to propose to control the estimation of the noise spectral variance for speech estimation.

After the estimation of $\eta_n(k)$, we limit the processing noise variance with a control factor $\rho(l)$ so that the processing noise variance is to be $\breve{\eta}_n(k) = \rho(l)\eta_n(k)$, which is able to mitigate the artificial distortion while speech estimator works on it. Replacing the estimated noise variance with the processing noise variance by using the control factor $\rho(l)$ is a way to control the noise overestimation. In [24], Gemello et al proposed a modified Ephraim-Malah LSA method that herein is marked **GMEM**, where the frame SNR (which was also called global SNR there) is used to control the noise overestimation and the floor of *a priori* and *a posteriori* SNRs with similar broken segmental linear relationship.

Conventional speech enhancement may adversely affect speech recognition accuracy when it is applied into the speech signal that is of very good quality. On the other hand the suppression of some heavy noise is not sufficient for the speech recognition purpose. In this paper, we introduce noise control factor $\rho(l)$ according to the following criteria. Firstly, when the frame SNR is quite high, the degree of noise suppression needs to be mitigated in order to lessen the artifacts. Making the very small value of processing noise retains the sensitivity of the speech decoder to the artificial distortion. It leads to less noise reduction to speech utterance of good quality, thus it keeps the enhancement artifacts to a certain low level. Secondly, when the frame SNR is low, it is believed that the artifacts can be further reduced with more noise suppression as compared to the conventional MMSE enhancer. It means that the processing noise $\breve{\eta}_n(k)$ is overestimated to be greater than the real noise level. Thirdly, the value of control factor higher than 1 results in more suppression of noise for low SNR, however, too high value of control factor may cause harmful distortion in speech quality. Fourthly, because the abrupt broken point causes the damage of speech spectrum, smoothing function is adopted. Based on the above analysis, we propose to make the noise control factor $\rho(l)$ to be adapted by frame SNR $\Xi(l)$ as follows

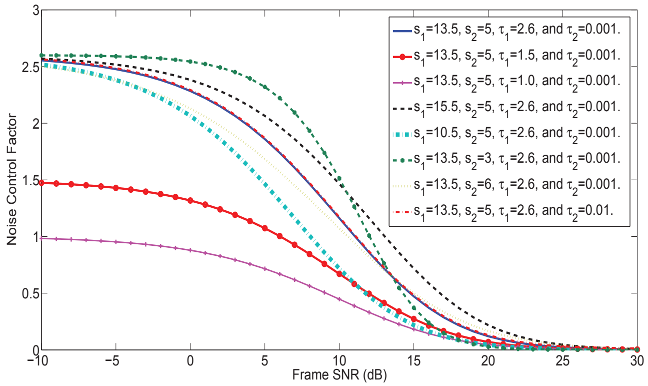

Fig. 1. *Noise control factor depending on frame SNR with different constant sets of τ_1, τ_2, s_1 and s_2.*

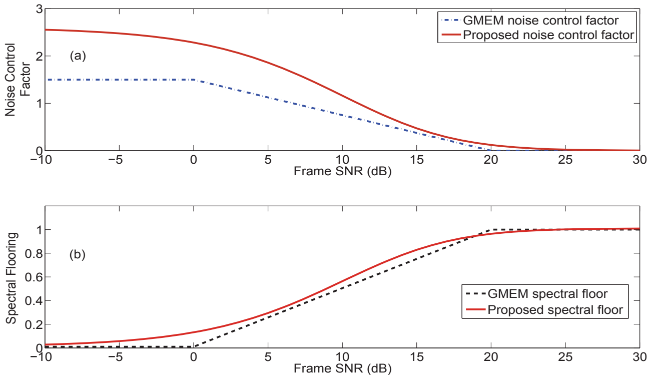

Fig. 2. *Noise control factor and weak spectral floor adapted to frame SNR.*

$$\rho(l) = \tau_1\phi^2(\Xi(l), s_1, s_2) + \tau_2 \tag{13}$$

where τ_1, τ_2, s_1 and s_2 are constants. and ϕ is a general sigmoid function as follows

$$\phi(x, r_1, r_2) = \frac{1}{1 + \exp\{(x - r_1)/r_2\}}. \tag{14}$$

Fig. 1 shows the relationship between the noise control factor $\rho(l)$ and the frame SNR with different configurations of constant parameters τ_1, τ_2, s_1 and s_2. The constant parameters are configured at lowest cepstral distortion ratio using a computer-generated noisy speech database originated from Switchboard database. As a result, we obtain empirically a proper constant parameter set of s_1=13.5, s_2=5, τ_1=2.6, and τ_2=0.001. Fig. 2 (a) shows the noise control factor ρ with respect to the frame SNR ($\Xi(l)$) for the **GMEM** noise control factor and our proposed smoothing noise control factor.

It is difficult to suppress noise without speech distortion, especially for weak speech component which is of very low SNR. In speech recognition, if the loss of weak speech spectrum can be effectively restricted, the information carried by the weak spectrum could be safely transferred to the feature domain and effectively used during matching processing. For weak spectrum component and silence period, a spectral floor is introduced into the *a priori* and *a posteriori* SNRs in [24] to avoid negative spectrum values. However, the adaptation of spectral floor to frame SNR is based on a broken linear relationship. In this paper, our goal is to design the weak spectral floor used to modify both the *a priori* and *a posteriori* SNRs with smoothing adaptation to avoid the broken points that may be harmful to the weak speech signal. Therefore, we propose the weak spectral floor to be adapted by the frame SNR as follows

$$\varsigma(l) = (1+\kappa) - \phi^2(\Xi(l), k_1, k_2) \tag{15}$$

where κ is the lower bound of the flooring factor. Similarly, the values of constant parameters (κ, k_1 and k_2) are selected at the lowest cepstral distortion ratio using the computer-generated noisy speech database. Then, empirically the constant parameters are set to κ=0.01, k_1=13.5, k_2=5. Fig. 2 (b) shows the spectral SNR floor adapted to the frame SNR ($\Xi(l)$) using the **GMEM** spectral floor and our proposed spectral floor respectively.

With the noise overestimation and weak speech spectral flooring, the *a posteriori* SNR is modified as follows

$$\breve{\gamma}_k(l) = \begin{cases} \frac{|X_k(l)|^2}{\rho(l)\eta_n(l)}, & \text{if } \frac{|X_k(l)|^2}{\rho(l)\eta_n(l)} \geq \varsigma(l)+1; \\ \varsigma(l)+1, & \text{otherwise} \end{cases} \tag{16}$$

and the *a priori* SNR is modified below

$$\breve{\xi}_k(l) = \begin{cases} \check{\xi}_k(l), & \text{if } \check{\xi}_k(l) \geq \varsigma(l); \\ \varsigma(l), & \text{otherwise} \end{cases} \tag{17}$$

where $\check{\xi}$ is given as follows

$$\check{\xi}_k(l) = \alpha\frac{|\breve{G}_k(l-1)X_k(l-1)|^2}{\rho(l)\eta_n(k,l)} + (1-\alpha)(\breve{\gamma}_k(l)-1) \tag{18}$$

where $\breve{G}_k(l-1) = G_k(\breve{\xi}_k(l-1), \breve{\gamma}_k(l-1))$. It means the MMSE gain function $\breve{G}_k(l-1)$ is actually the function of $\breve{\xi}_k(l-1)$ and $\breve{\gamma}_k(l-1)$ for the previous frame $l-1$.

We aim to design a modification scheme to reduce the feature distortion, so that the speech recognition accuracy may be improved. Using LSA-MMSE speech estimator, Fig. 3 shows the cepstral distortion ratio of the different processing effects including LSA-MMSE with **GMEM** modification (**LSA:GMEM**) [24], LSA-MMSE with our proposed smoothing adaptation (**LSA:P1**) by eqs (13)-(18), and three baselines (noisy speech utterances without enhancement processing (**Noisy**), LSA-MMSE with cepstral domain *a priori* SNR estimation (**LSA:CEP**) [25], and LSA-MMSE with conventional decision-directed approach for *a priori* SNR estimation [11]). Here, **P1** denotes our proposed smoothing adaptation of the noise control factor and weak spectral floor, and **GMEM** denotes conventional broken linear adaptation of the noise control factor and the weak spectral floor [24]. The statistics of the cepstral distortion ratio is computed by using 200 utterances contaminated by 10 dB F16 noise. The 200 utterances are randomly selected from Switchboard database. We can see the improvement of the feature distortion using the enhanced utterances, and our proposed smoothing method is better than the **GMEM** method [2].

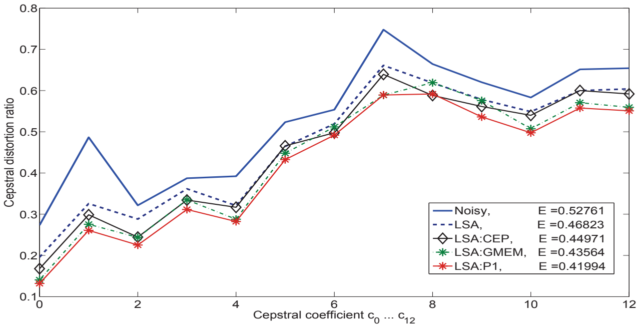

Fig. 3. *Cepstral distortion statistics with 200 utterances.*

B. Oversuppression of Residual Noise

It has been known that musical noise can very apparently appear in spectral suppression using spectral subtraction method [38]. In fact, attenuation of very noisy speech with MMSE algorithm can also cause the musical noise phenomenon. Since residual noise spectrum consists of peaks and valleys with random occurrences, we can seek an oversuppression to attenuate the spectral excursions beyond the MMSE criterion for improving speech quality.

An adaptive oversuppression function in respect to frame SNR can effectively restrict the spectral excursions of noise peaks to a lower bound so that descend the amount of the musical noise. Considering potentially adverse effect of denoising gain applied for high quality speech and insufficient suppression of heavy noise for the purpose of speech recognition, we introduce oversuppression according to the following consideration. When frame SNR Ξ is very high, the oversuppression is not applied. When the frame SNR is low, the oversuppression is applied, and subsequently the oversuppression factor is adjusted to low value depending on the level of frame SNR. Therefore, we propose to further suppress the residual noise by introducing an adaptive smoothing oversuppression factor as follows

$$\omega(l) = 1 + (\varpi - 1)\phi^2(\Xi(l), w_1, w_2) \tag{19}$$

where constant ϖ is the lower bound of the gain control factor, w_1 and w_2 are constant. Similarly, we use the computer-generated speech database to make the configuration of the constant parameters (κ, k_1 and k_2), and finally we get lowest distortion ratio by searching the different constant parameter sets and obtain the proper configuration as $\varpi = 0.1$, w_1=-3 and w_2=2. Subsequently, the gain is modified as follows

$$\breve{\breve{G}}_k(l) = \omega(l)\breve{G}_k(l). \tag{20}$$

Fig. 4 shows the oversuppression factor ω adapted to frame SNR. The adaptive oversuppression brings an obvious improvement for speech recognition. It is believed that the

[2] For fair comparison, the parameter settings of **GMEM** and the proposed method are exactly the same throughout all experiments in this paper, e.g. the α of (10) is set to 0.7 for both.

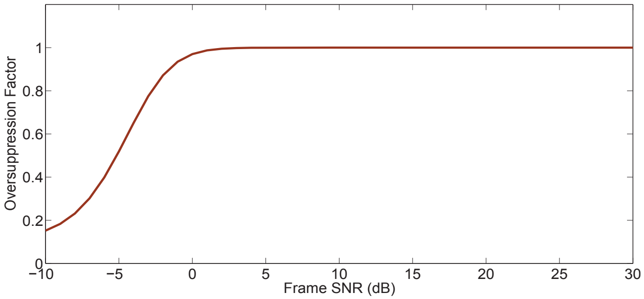

Fig. 4. *Oversuppression factor adapted to frame SNR.*

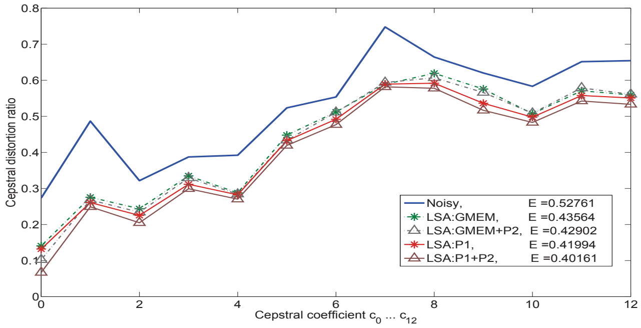

Fig. 5. *Cepstral distortion statistics with 200 utterances.*

adaptive oversuppression make the endpoints more correctly be aligned during speech recognition.

Fig. 5 shows the cepstral distortion ratio of the different processing effects using LSA speech estimator. **P2** denotes the proposed oversuppression method by eqs (19)-(20). Fig. 5 indicates that, in terms of the cepstral distortion ratio, the **LSA:P1+P2** is better than **LSA:P1** in terms of the cepstral distortion ratio, and **LSA:GMEM+P2** is better than **LSA:GMEM**. It means that **P2** brings improvement on the cepstral distortion ratio, and **P1+P2** is better than **GMEM+P2**.

C. Re-estimation of a priori SNR

With the optimization for the speech amplitude estimation, it is believed that the estimation of *a priori* SNR should be improved and closer to the true values if we can use the current frame estimated suppression gain to replace the previous frame estimated gain in the modified decision-directed equation. Since the maximum likelihood estimate of $\mathbf{E}(|S_k(l)|^2)$ is $|\hat{S}_k(l)|^2$, therefore, we can have the re-estimate of *a priori* SNR with the computed gain $\breve{G}_k(l)$ that depends on an initial approximate of the modified *a priori* SNR $\breve{\xi}_k(l)$ using a modified version of the decision-directed approach (17) as follows

$$\tilde{\xi}_k(l) = \max\left[\alpha_c \frac{|\breve{G}_k(l)X_k(l)|^2}{\rho(l)\eta_n(k,l)} + (1-\alpha_c)(\breve{\gamma}_k(l)-1), \quad \varsigma(l)\right]. \tag{21}$$

Experiment shows that the reestimation improves the feature distortion for speech recognition, and the best result falls on $\alpha_c = 1$. Subsequently, the reestimation of the *a priori* SNR is only based its definition in (6).

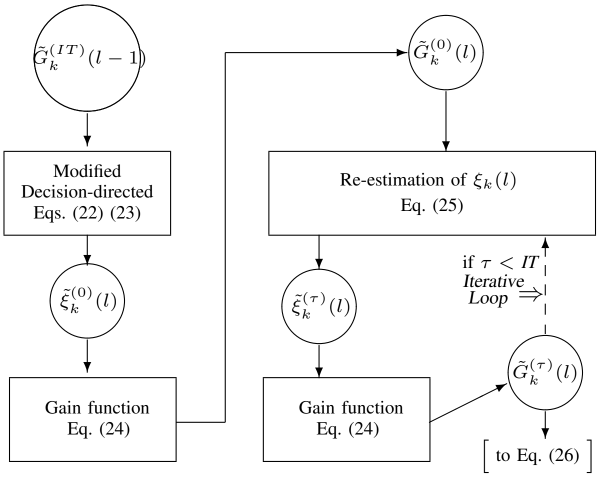

Fig. 6. *The scheme of the proposed estimation of the a priori SNR.*

Eventually, the reestimation and gain function form a pair of iterative algorithm. Theoretically, we can re-estimate the *a priori* SNR iteratively to obtain a proper estimated gain. An investigation shows that increasing iteration may improve some speech objective measurement like SNR and modified Bark spectral distortion (MBSD). Let *IT* denote the number of iteration with *IT*=1 denoting one-time usage of reestimation, we propose a reestimation scheme for the *a priori* SNR as follows

$$\breve{\xi}_k(l) = \breve{\alpha}\frac{|\omega(l)\tilde{G}_k^{(IT)}(l-1)X_k(l-1)|^2}{\rho(l)\eta_n(k,l)} + (1-\breve{\alpha})(\breve{\gamma}_k(l)-1) \tag{22}$$

$$\tilde{\xi}_k^{(0)}(l) = \begin{cases} \breve{\xi}_k(l), & \text{if} \quad \breve{\xi}_k(l) \geq \varsigma(l); \\ \varsigma(l), & \text{otherwise} \end{cases} \tag{23}$$

$$\tilde{G}_k^{(\tau-1)}(l) = G_k(\tilde{\xi}_k^{(\tau-1)}(l), \breve{\gamma}_k(l)), \qquad \tau = 1, ..., IT. \tag{24}$$

$$\tilde{\xi}_k^{(\tau)}(l) = \max\Big[\frac{|\tilde{G}_k^{(\tau-1)}(l)X_k(l)|^2}{\rho(l)\eta_n(k,l)}, \quad \varsigma(l)\Big]. \tag{25}$$

As a result, the estimate of speech spectrum is given by

$$\hat{S}_k(l) = \omega(l)\tilde{G}_k^{(IT)}(l)X_k(l). \tag{26}$$

Here, as mentioned in the beginning of the section, $\breve{\alpha}$ is also empirically set to 0.7. Fig. 6 shows the flow chat of the proposed *a priori* SNR reestimation scheme.

Let **P3** denote the proposed *a priori* SNR reestimation method by eqs (22)-(26). Fig. 7 shows the comparison in terms of the feature statistics of the cepstral distortion ratio computed by using the 200 utterances. It indicates that the **P1+P2+P3**(*IT*=1) is better than **P1+P2**, and **GMEM+P2+P3** is better than **GMEM+P2**, and **P1+P2+P3**(*IT*=1) is better than **GMEM+P2+P3**(*IT*=1). It means that **P3**(*IT*=1) mitigates the artifacts by improving the cepstral distortion ratio. We also can see that the statistics distortion with **P1+P2+P3**(*IT*=1) and **P1+P2+P3**(*IT*=2) is very closer. **P1+P2+P3**(*IT*=3) is not so good as **P1+P2+P3**(*IT*=1).

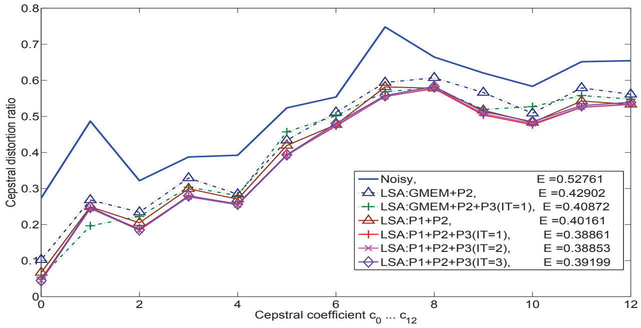

Fig. 7. *Cepstral distortion statistics with 200 utterances.*

IV. Speech Enhancement for Speech Recognition

A. Acoustic Models for Speech Recognition

Speech recognition is based on statistical models trained from transcribed speech and pronunciation dictionaries. In the system, hidden Markov model (HMM) as a temporal topology linking a sequence of states plays a critical role.

A state-of-the-art speech recognizer is setup by elaborating the language modeling and acoustic modeling process in the following.

LM: The language model (LM) is trained with lexicon of 30,858 vocabulary size using SRILM toolkit [31]. We use Part 1 of Fisher transcripts that are equivalent to 700 hours of speech for training.

Acoustic Modeling: by using 260k utterances (313 hr 23 min) from Switchboard-1-LDC97S62 database with Kaldi toolkit [29], we start the acoustic model training from single Gaussian monophone modeling, coarse triphone modeling, and detailed triphone GMM-HMM modeling, and then followed by DNN-HMM modeling etc. The acoustic models with HMM topology [32] [33] are described as follows.

1) **GMM**:
GMM-HMM is sophisticated acoustic model where states are usually delimited as separate GMMs. It is able to cope with the most important sources of speech ambiguity and performs to be enough flexible to allow the realization of recognition systems with large dictionaries. The GMM probability with respect to an observed speech frame represents the possibility of the speech frame generating from the corresponding state, while a transition matrix of the HMM with the probability of moving from one state to another is used to despatch the successive process across time sequence.
GMM-HMM model with 11,500 tied-states and 200,000 total Guassian components are trained with LDA, MLLT and fMLLR-SAT techniques by using 192,000 (286 hours) unique utterances which is aligned by using a previous trained GMM-HMM acoustic model and the **LM** language model [34].

2) **DNN**:
Considering all of the tied-states covering the entire text-dependent phoneme in one deep neural network (DNN) model, DNN is used to generate the posterior probabilities for the HMM states. The parameters of DNN are trained by optimizing the cross-entropy through stochastic gradient descent using error backpropagation procedure [35].
The DNN-HMM model of five hidden layers with 2,048 neurons each hidden layer is trained with cross-entropy using GPU. The 192,000 fMLLR-transformed utterances are aligned by using the **GMM** model and the **LM**, and then are split into two parts: 90% of the features used for DNN training and 10% is used for cross validation of DNN training.

3) **sMBR**:
The state-level minimum Bayes risk (sMBR) is to minimize the expected state errors based on the corresponding state labels by utilizing both HMM topology and language model through the searching lattice [36] [37]. DNN-HMM-sMBR model of five hidden layers with 2048 neurons each hidden layer is trained with sMBR criterion by using the fMLLR-transformed feature of 192,000 utterances which are aligned by using the **DNN** model and the **LM** model.

B. Speech Recognition Platform for Speech Enhancement Evaluation

It is well known that when HMM model is trained in quiet condition and is tested in noisy environment, the recognition accuracy will drop dramatically. Multi-condition training using various patterns of noisy databases can increase the accuracy of noisy recognition accuracy by simulating different kinds of noise environment. However, in practice, we cannot predict the condition of training model to match a dissimilar noise environment. One possible way to solve the problem is to retrain the HMMs in new noisy condition. However, in most of realistic applications, this is either inconvenient or impracticable. The HMM model trained with clean speech database is useful with front-end processing techniques including speech enhancement and feature enhancement applied in noise.

In this paper, we aim to examine the genuine contribution of the speech enhancement algorithms to speech recognition while the state-of-the-art feature enhancement techniques have been applied in the speech recognition system. In particular, we used 13 dimensional MFCC feature for speech recognition system, CMVN, LDA, MLLT and fMLLR feature enhancement techniques. We selected 1,831 English sentences with 21,395 words from the Switchboard corpus, as the test dataset marked as 'SWBD'; and we chose 2,628 English sentences with 21,594 words from the Callhome corpus as test dataset marked as 'CALLHM'. We used the speech utterances selected from the Hub-5-2000 English test corpus which includes about 3.3 hours Switchboard corpus and 20 hours Callhome speech corpus as test dataset. We add different types of noise with specified global SNR to generate various kinds of noisy speech dataset.

As an upper bound reference, Table I shows the WERs of the clean dataset from the 'SWBD' and 'CALLHM' with **GMM**, **DNN** and **sMBR** decoders.

Speech enhancement is applied at the very first stage before feature extraction. Fig. 8 shows the evaluation platform with the speech enhancement algorithms for different decoders with their corresponding acoustic models. In the recognition, the

TABLE I
Performance of various decoders under clean speech

SWBD	Decoder	**GMM**	**DNN**	**DNN-sMBR**
	WER (%)	23.3	15.5	14.4
CALLHM	Decoder	**GMM**	**DNN**	**DNN-sMBR**
	WER (%)	39.3	28.0	26.8

TABLE II
*Performance evaluation for the LSA-MMSE in terms of WER with different noise types in 10 dB using the **sMBR** decoder for the SWBD databases*

Estimation of spectral SNR	**White**	**F16**	**Factory1**
Noisy (i.e. without denoising)	57.3%	54.1%	53.4%
LSA:GMEM	40.4%	37.4%	39.8%
LSA:P1	37.5%	35.2%	37.6%
LSA:GMEM+P2	38.6%	36.0%	38.4%
LSA:P1+P2	36.3%	34.6%	36.3%
LSA:GMEM+P2+P3(*IT*=1)	34.8%	33.5%	36.1%
LSA:P1+P2+P3(*IT*=1)	33.9%	32.1%	35.2%
LSA:P1+P2+P3(*IT*=2)	33.8%	32.2%	35.5%
LSA:P1+P2+P3(*IT*=3)	34.1%	32.5%	35.7%

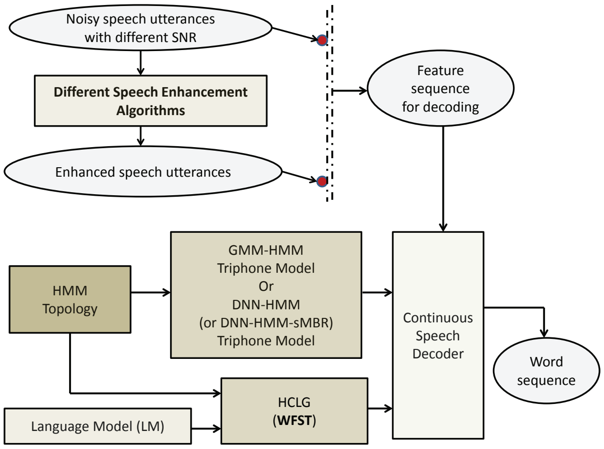
Noisy speech utterances with different SNR
Different Speech Enhancement Algorithms
Enhanced speech utterances
Feature sequence for decoding
HMM Topology
GMM-HMM Triphone Model Or DNN-HMM (or DNN-HMM-sMBR) Triphone Model
Continuous Speech Decoder
Word sequence
Language Model (LM)
HCLG (WFST)

Fig. 8. *The evaluation platform for the speech enhancement algorithms in different decoders.*

decoding-graph WFST (weighted finite-state transducer) is constructed by HCLG = $H \circ C \circ L \circ G$, where G is an acceptor that encodes the grammar (or language model), L represents the lexicon with its output symbols being words and its input symbols being phones, C represents the context-dependency, and H contains the HMM configuration with its output symbols representing context-dependent phones and its input symbols the transitions-ids. The decoders are formed by the acoustic models and the decoding graph WFST.

The purpose of choosing the different decoders is to investigate the effects of the speech enhancement algorithms in different stages of the speech recognition modeling, to gain insights into the usefulness of the speech enhancement algorithms in the machine recognition as opposed to human subjective listening.

V. Performance evaluation

We setup the evaluation platform to measure the performance of different speech enhancement algorithms for the state-of-the-art speech recognition system. The three decoders **GMM**, **DNN** and **sMBR** as introduced in section IV-B are used in the evaluation.

Different type of noise are added into the test speech database to generate different group of noisy speech database with different global SNRs, i.e. 0 dB, 10 dB, 20 dB and 30 dB. And the speech enhancement algorithms are applied into the noisy speech databases. In this paper, we select three types of noises, i.e. white noise, F16 noise and Factory1 noise.

A. Investigated Enhancement Algorithms

We choose four existing algorithms as speech denoising baselines, they are cepstral-domain a priori SNR estimation [25], modified E-M LSA [24], ETSI noise reduction [53] and autoencoder denoising [51] [52]. Their details are described as follows:

CEP: In [25], Breithaupt et al introduced an estimation of *a priori* SNR in cepstral-domain where the pitch can easier detected and special considered with a selective smoothing scheme.

GMEM: In [24], Gemello et at propose to modify E-M LSA by estimating the *a priori* and the *a posteriori* SNRs with the noise overestimation factor and the SNR spectral floor.

ETSI: The standard ETSI noise reduction is based on two stages of applying Mel-warped Wiener filtering. Each stage has linear Wiener filter coefficients smoothed by using Mel filter-Bank, and the impulse response of this Mel-warped Wiener filter obtained by applying a Mel-warped inverse discrete cosine transform (iDCT) [53].

Autoencoder: Autoencoder denoising can be used for noisy speech recognition, its neural network parameters is trained in fMLLR transformed feature domain by using a pair of speech, whose noisy speech is used as input and the clean speech as target output [52]. Noise sources from NOISEX-92 database [50] are selected to add into a clean speech database with different global SNR. In particular, we split 286 hours of speech database from Switchboard into 35 groups; then add 7 types of noises with 5 different SNR (i.e., 0dB, 10dB, 20dB, 30dB and 40dB) separately to form 35 different speech groups. Out of the three selected noises in this paper, only Factory1 noise is involved for autoencoder DNN training, but white and F16 noises are excluded. [3]

To study the contributions of **P1**, **P2** and **P3**, Table II shows comparison between the two adaptation methods in terms of the WER performance of the **sMBR** decoder with different types of noise in 10 dB. It is obvious that the performance of LSA with **P1** is consistently better than the one with **GMEM**.

It can be seen that the proposed reestimation scheme obviously outperforms over the conventional decision-directed approach for speech recognition. However, the WER performance of the reestimation with *IT*=1 and *IT*=2 is very similar, but the one of *IT*=3 is apparently worse than that of *IT*=1. The observation of WER performance is consistent with the cepstral distortion ratio in section III-C where statistic cepstral distortion ratios with **P1+P2+P3**(*IT*=1) and **P1+P2+P3**(*IT*=2) are very closer, while the cepstral distortion ratio with **P1+P2+P3**(*IT*=3) is not so good as that with **P1+P2+P3**(*IT*=1), although our observation with the listening-purpose measurement in terms of SNR and PESQ (perceptual

[3] The purpose of including versus excluding noise is to observe the different effect between including and excluding noises.

evaluation of speech quality) shows the P1+P2+P3(IT=3) is better than P1+P2+P3(IT=2). According to the above observation, we adopt only one-time reestimation (i.e. *IT=1*) in the next experiment.

In the experiment, the performance of the three typical speech enhancement algorithms, i.e. Wiener filter, LSA and β-masking, are investigated with conventional spectral SNR estimation and the proposed estimation. In the following experiment, we are showing the performance of our proposed scheme, **PRO**, which combine the three proposed **P1+P2+P3** methods. Following is a list of the baselines and the investigated algorithms.

- **Noisy**: Noisy speech baseline without denoising;
- **Wiener**: Conventional Wiener filter [9];
- **LSA**: Conventional LSA filter [11];
- **mβ**: Conventional β-masking MMSE [13];
- **LSA:GMEM**: Gemello's modified E-M LSA [24] baseline;
- **ETSI**: ETSI baseline [53];
- **Autoencoder**: Autoencoder denoising baseline [52];
- **Wiener:CEP**: Wiener with cepstral *a priori* SNR [25] baseline;
- **LSA:CEP**: LSA with cepstral *a priori* SNR [25] baseline;
- **Wiener:PRO**: Wiener with the proposed **P1+P2+P3**(*IT*=1);
- **LSA:PRO**: LSA with **P1+P2+P3**(*IT*=1);
- **mβ:PRO**: β-masking with **P1+P2+P3**(*IT*=1).

B. WER performance of different enhancement algorithms

Now we are showing the enhancement performance for speech recognition. The platform system was built using Kaldi toolkit [29], which is the state-of-the-art open source speech recognition software. The performance of various speech enhancement algorithms is measured in terms of WER with different speech recognition decoders. The experimental result is reported in Figs 9-14 [4] for white, F16 and Factory1 noises respectively, where the Switchboard data 'SWBD' represents the same recording channel as training database, and Callhome data 'CALLHM' represents the different recording channel from that of the training database.

From the experimental results, we observed that the speech enhancement may perform different effects on different decoders. This phenomenon is obvious when comparing **GMM** and **DNN**. In high SNR, a speech enhancement can be helpful in some modeling case but possibly bring worse effect in another modeling situation. The instability is due to the unreliable artifacts introduced by the enhancer in different characteristics. Therefore, revealing the enhancement problem hidden inside the enhancement algorithm is meaningful.

In the figures, we highlight the best WER value with bold-font style for easy reference. From the figures, it can be seen that almost all of the speech enhancement algorithms gives positive effect to speech recognition, except a few of them brings negative effect for the case of high SNR (e.g. 30 dB).

Conventional **LSA** helps on **GMM** decoder but does not help on **DNN** and **sMBR** decoders for the case of 30 dB white noise.

It is obvious that the improvement is great with our proposal scheme. For all selected enhancement algorithms, i.e. Wiener filter, LSA and β-masking MMSE, the progress is consistent. In particular, **Wiener:PRO** is generally better than **Wiener:CEP** for all decoders with some exceptions, especially for low SNR situation. We also can see that **LSA:PRO** is almost consistently better than **LSA:CEP** as well as **LSA:GMEM**. The reason is that **PRO** synthesizes the advantage of the smoothing effect, more accurate estimation of *a priori* and *a posteriori* SNRs, and the appropriate smoothing oversuppression of noise in low frame SNR situation.

LSA is consistently better than Wiener filter in the baseline version and in the proposed version. In general, **mβ** is better than LSA for most of cases both in the baseline version and in the proposed version. It examines that the masking information can bring positive contribution to the speech recognition system.

We can see the **mβ:PRO** gives great improvement in low SNR case for all the three speech recognition decoders, and totally helpful in high SNR case.

It is noticed that autoencoder performance good only for low SNR case with known noise situation (i.e. Factory1 noise), but its performance drops down rapidly when it is applied for high SNR case. Generally, its performance is not consistently positive. Our proposed methods generally outperform it. Although we believed that if we re-organize the training database for autoencoder DNN, the performance may improve, but the limitation of the method is still obvious as compared to the potential of spectral-domain enhancement algorithm.

We notice that the **ETSI** is a powerful noise reduction system, especially in the low SNR situation in 'SWBD' testbed, it significantly reduce the WER to an amazing level. This advantage is very apparent for F16 noise, WER from 85.7% drops to 58.9%, reaches 31.27% improvement for 0 dB with **sMBR** decoder, although **mβ:PRO** makes a significant improvement and let WER dropped to 62.3%, with 27.3% improvement. However, ETSI does not perform so good in 'CALLHM' testbed.

However, **mβ:PRO** captures up and outperforms **ETSI** in White and Factory1 noises with low SNR. And it wins most of best accuracy totally. For the case of 10 dB Factory1 noise, it drops WER from 53.4% to 35.0% to gain 34.46% improvement with **sMBR** decoder.

The experiments described in the figures made evident that our proposed scheme brings positive and effective progress for the ASR denoising.

VI. Conclusions

We established a state-of-the-art speech recognition platform for speech enhancement evaluation, and investigated typical spectral-domain enhancement algorithms for different speech recognition decoders under various noise conditions. Against the weakness of the conventional speech enhancement algorithm, this paper aims to reveal the potential of the speech enhancement for the purpose of speech recognition. We therefore proposed a series of modification on spectral SNR and the suppression gain to mitigate the feature distortion for speech recognition including smoothing-adaptation scheme for controlling the processing noise power and mitigating the harmful artifacts for weak speech signal, oversuppression of the unwanted residual noise component, and the reestimation

[4] Notice: ** denotes that no test is done, * indicates that the decoder cannot work properly due to strong noise condition.

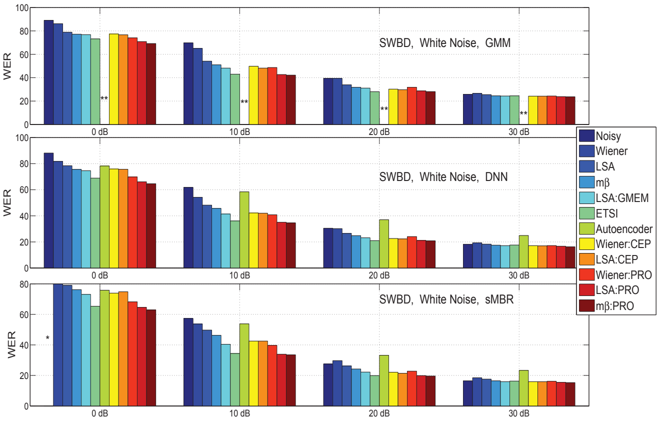

Fig. 9. *Speech Enhancement Performance with SWBD Dataset Under White Noise Condition in terms of WER (%)*

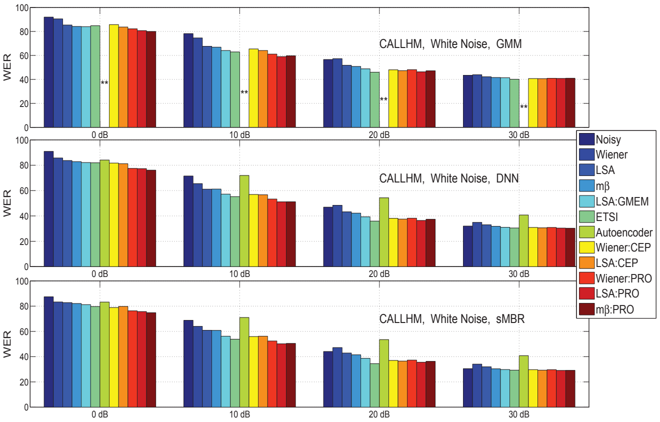

Fig. 10. *Speech Enhancement Performance with CALLHM Dataset Under White Noise Condition in terms of WER (%)*

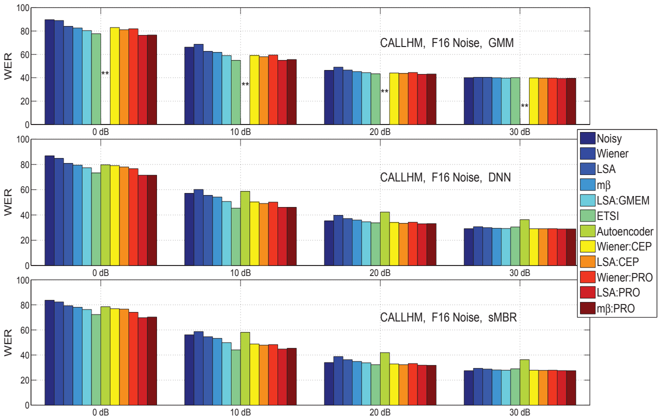

Fig. 12. *Speech Enhancement Performance with CALLHM Dataset Under F16 Noise Condition in terms of WER (%)*

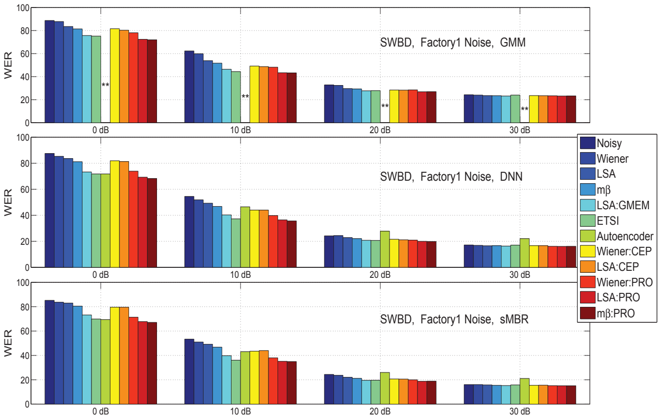

Fig. 13. *Speech Enhancement Performance with SWBD Dataset Under Factory1 Noise Condition in terms of WER (%)*

of *a priori* SNR. With the experimental result, we have the following conclusions: the introduction of frame SNR and the smoothing adaptation methods are effective; an enhancer may be helpful for certain speech decoder but not always contribute to other speech decoder; the proposed scheme is significantly effective for all the three typical speech enhancement algorithms for speech recognition; LSA is evidenced to be almost consistently better than Wiener filter in terms of WER; the general performance of β-masking MMSE is better than the other two, i.e. Wiener filter and LSA.

References

[1] J. Li, L. Deng, Y. Gong, R. Haeb-Umbach, "An Overview of Noise-Robust Automatic Speech Recognition," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 22, issue. 4, pp. 745-777, Feb. 2014.

[2] C. Guan, Y. Cebn And B. Wu, "Direct Modification on LPC Coefficients with Application to Speech Enhancement and Improving the Performance of Speech Recognition in Noise," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-93, 1993

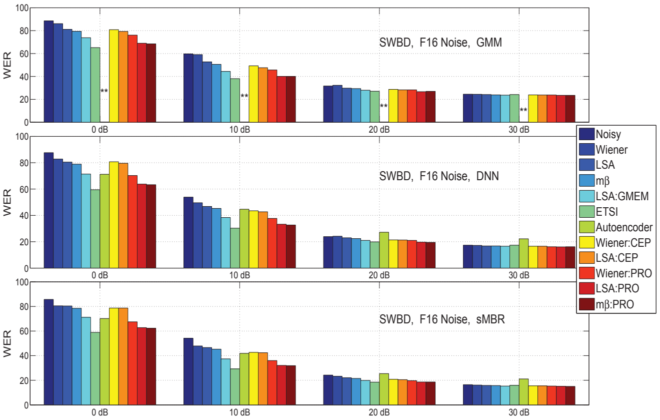

Fig. 11. *Speech Enhancement Performance with SWBD Dataset Under F16 Noise Condition in terms of WER (%)*

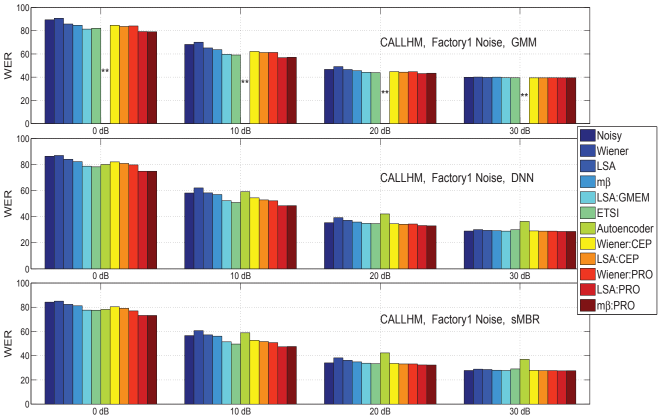

Fig. 14. *Speech Enhancement Performance with CALLHM Dataset Under Factory1 Noise Condition in terms of WER (%)*

[3] C. Breithaupt and R. Martin, “Statistical analysis and performance of DFT domain noise reduction filters for robust speech recognition,” *Int. Conf. on Spoken Lang. Process.* (ICSLP), 2006.

[4] D. O’Shaughnessy, “Invited paper: Automatic speech recognition: History, methods and challenges,” *Pattern Recognition*, vol. 41, no. 10, pp. 29652979, 2008.

[5] R. Flynn and E. Jones, “Robust Distributed Speech Recognition using Speech Enhancement,” *IEEE Transactions on Consumer Electronics*, vol. 54, no. 3, pp. 1267-1273, March 2008.

[6] J. Tchorz and B. Kollmeier, “A model of auditory perception as front end for automatic speech recognition,” *J. Acoust. Soc. Amer.*, vol. 106, no. 4, pp. 2040-2050, 1999.

[7] X. Lu, M. Unoki, S. Matsuda, C. Hori, and H. Kashioka, “Controlling tradeoff between approximation accuracy and complexity of a smooth function in a reproducing kernel Hilbert space for noise reduction,” *IEEE Trans. on Signal Process.*, vol. 61, no. 3, pp. 601-610, 2013.

[8] R. J. McAulay and M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137145, 1980.

[9] V. Stahl, A. Fisher and R. Bippus, “Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering,” *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP, 2000.

[10] Y. Ephraim and D. Malah, “Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator,” *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-32, No. 6, pp. 1109-1121, Dec. 1984.

[11] Y. Ephraim and D. Malah, “Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator,” *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-33, No. 2, pp. 443-445, Apr.1985.

[12] C.H. You, S.N. Koh, and S. Rahardja, “β-Order MMSE Spectral Amplitude Estimation for Speech Enhancement,” *IEEE Trans. Speech and Audio Processing*, Vol. 13, No. 4, pp. 475-486, Jul. 2005.

[13] C.H. You, S.N. Koh, and S. Rahardja, “Masking-Based β-Order MMSE Speech Enhancement”, *Speech Communication*, Vol. 48, Issue 1, pp. 57-70, Jan. 2006.

[14] S. Gannot, D. Burshtein and E. Weinstein, “Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms,” *IEEE Trans. on Speech and Audio Processing*, Vol. 6 No. 4 pp 373-385, Jul. 1998.

[15] C.H. You, S.N. Koh and S. Rahardja, “Kalman Filtering Speech Enhancement Incorporating Masking Properties for Mobile Communication in a Car Environment,” *Proc. IEEE International Conference on Multimedia and Expo*, ICME’2004, Taiwan, Jun. 2004.

[16] R.P. Hellman, “Asymmetry of Masking between Noise and Tone,” Perception and Psychophysics, Vol. 11, pp.241-246, 1972.

[17] J. D. Johnston, “Transform Coding of Audio Signal Using Perceptual Noise Criteria,” *IEEE J. Select Areas Commun.*, Vol. 6, pp. 314-323, Feb. 1988.

[18] D. Tsoukalas, M. Paraskevas and J. Mourjopoulos, “Speech Enhancement Using Psychoacoustic Criteria,” *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-93., Vol. 2, pp. 359-362, 1993.

[19] I. Cohen, “Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-spectral Amplitude Estimator,” *IEEE Signal Process. Letters*, Vol. 9, I. 4, pp. 113-116, 2002.

[20] B. Fodor and T. Fingscheidt, “MMSE Speech Enhancement Under Speech Presence Uncertainty Assuming (Generalized) Gamma Speech Priors Throughout,” *Int. Conf. Acoust., Speech and Signal Processing*, ICASSP, pp. 4033-4036, 2012.

[21] Y. Ephraim, H. L. Van Trees, “A Signal Subspace Approach for Speech Enhancement,” *IEEE Trans. Speech and Audio Processing*, Vol. 3, No. 4, pp. 251-266, Jul. 1995.

[22] C.H. You, S. Rahardja and S.N. Koh, “Speech Enhancement for Telephony Name Speech Recognition,” *ICME* 2008

[23] J. H. L. Hansen and M. A. Clements “Constrained Iterative Speech Enhancement with Application to Speech Recognition,” *IEEE Trans. on Sign. Process.*, vol. 39, no. 4. Apr. 1991.

[24] R. Gemello, F. Mana, and R. D. Mori, “Automatic Speech Recognition with a Modified EphraimMalah Rule,” *in IEEE Sig. Process. Lett.* Vol. 13, No. 1, pp. 56-59, Jan. 2006

[25] C. Breithaupt, T. Gerkmann, and R. Martin, “A Novel a Priori SNR Estimation Approach Based on Selective Cepstro-Temporal Smoothing,” *IEEE Int. Conf. Acoust., Speech, Signal Processing*, ICASSP, pp. 4897-4900, Apr. 2008.

[26] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, A. Acero, “Robust Speech Recognition Using a Cepstral Minimum-Mean-Square-Error-Motivated Noise Suppressor,” *IEEE Trans. on Audoi, Speech, and Language Process.*, vol. 15, no. 5, July 2008.

[27] H. G. Hirsch and D. Pearce, “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” *in Proc. ISCA ITRW ASR*, 2000.

[28] K.K. Paliwal, J.G. Lyons, S. So, A.P. Stark, K.K Wójcicki, “Comparative Evaluation of Speech Enhancement Methods for Robust Automatic Speech Recognition,” *Int. Conf. Sig. Proce. and Comm. Sys.*, Gold Coast, Australia, ICSPCS, Dec. 2010.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, K. Veselý, N. Goel, M. Hannemann, P.Motlicek, Y. Qian, P. Schwarz, J. Silovsky, and G. Stemmer. The Kaldi speech recognition toolkit. *ASRU*, 2011.

[30] P. J. Wolfe and S. J. Godsill, “Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement,” *EURASIP Journal on Applied Signal Processing*, vol. 10, pp. 10431051, 2003.

[31] http://www.speech.sri.com/projects/srilm/manpages/

[32] Rabiner, L. R., and Juang, B. H., “An introduction to hidden Markov models,” *IEEE ASSP Mag.* 3, 416, 1986.

[33] Rabiner, L. R., “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, 77, 257286, 1989.

[34] D. Povey, “Discriminative training for large vocabulary speech recognition,” *Ph.D. dissertation*, University of Cambridge, Cambridge, UK, 2003.

[35] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” Nature, vol. 323, pp. 533536, October 1986.

[36] M. Gibson and T. Hain, “Hypothesis spaces for minimum Bayes risk training in large vocabulary speech recognition,” *Proc. INTERSPEECH*, pp. 24062409, Sep. 2006.

[37] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, “Sequence-discriminative training of deep neural networks,” *Interspeech*, 2013.

[38] O. Cappé, “Elimination of The Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor,” *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 2, pp. 345-349, 1994.

[39] J.S. Lim and A.V. Oppenheim, “Enhancement and Band-Width Compression of Noisy Speech,” *Proceedings Of The IEEE*, Vol. 67, No. 12, pp. 1586-1604, Dec. 1979.

[40] R. J. McAulay and M. L. Malpass, “Speech Enhancement Using a Soft-Decision Noise Suppression Filter,” *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-28, No. 2, pp.137-145, Apr. 1980.

[41] C.H. You, B. Ma, and C. J. Ni, “Modification on LSA Speech Enhancement for Speech Recognition,” *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, Accepted, New Orlean, Mar. 2017.

[42] Z. Chen and V. Hohmann “Online Monaural Speech Enhancement Based on Periodicity Analysis and *a Priori* SNR Estimation,” *IEEE/ACM Trans. Aud., Speech, and Lang. Process.* Vol. 23, No. 11, Nov. 2015.

[43] C.H. You, S.N. Koh, and S. Rahardja, “Adaptive β-Order MMSE Estimation for Speech Enhancement”, *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-03, Vol. 1, pp. 852-855, 2003.

[44] C.H. You, S.N. Koh, and S. Rahardja, “An MMSE Speech Enhancement Approach Incorporating Masking Properties”, *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-04, Vol. 1, pp. 725-728, May 2004.

[45] R. Martin, “Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics,” *IEEE Trans. Speech and Aud. Process.*, Vol. 9, No. 5, pp. 504 -512, Jul. 2001.

[46] S. Rangachari and P. C. Loizou, “A noise-estimation algorithm for highly non-stationary environments,” *Speech Communi.*, Vol. 48, pp. 220-231, 2006.

[47] R. Yu, “A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction,” *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 44214424, 2009.

[48] R.C. Hendriks, R. Heusdens and J. Jensen, “MMSE based noise PSD tracking with low complexity,” *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 4266-4269, 2010.

[49] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based Noise Power Estimation with Low Complexity and Low Tracking Delay,” *IEEE Trans. Audio, Speech and Language Process.*, vol. 20, no. 4, pp. 13831393, 2012.

[50] A. Varga and H. Steeneken, “Assessment for Automatic Speech Recognition: II. NOISEX-92: a Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems,” *Speech Communication*, Vol.12, No. 3, pp. 247-251, Jul. 1993.

[51] X. Lu, Y. Tsao, S. Matsuda and C. Hori, “Speech enhancement based on deep denoising autoencoder,” *In Proc. Interspeech*, pp. 436440, 2013.

[52] L. Wang, B. Ren, Y. Ueda, A. Kai, S. Teraoka and F. Fukushima, “Denoising autoencoder and environment adaptation for distant-talking speech recognition with asynchronous speech recording,” *Proc. of APSIPA ASC 2014*, Dec. 2014.

[53] ETSI ES 202 212 V1.1.2 (2005-11) “Two stage mel-warped Wiener filter approach,” *ETSI Standard: Extended advanced front-end feature extraction algorithm*, 2005