MODIFICATION ON LSA SPEECH ENHANCEMENT FOR SPEECH RECOGNITION

Chang Huai You, Bin Ma, Chongjia Ni

Institute for Infocomm Research, A*STAR, Singapore

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 distortions to the speech signals. In this paper, we aim at reducing the artifacts that has adverse effects on speech recognition. With this motivation, we propose a modification scheme including smoothing adaptation to frame SNR and reestimation of *a priori* SNR for spectral-domain log-spectral-amplitude (LSA) speech enhancement. The experiments show that the proposed scheme of enhancement significantly improves the performance of the state-of-the-art speech recognition over the baseline speech enhancement.

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

1. INTRODUCTION

State-of-the-art automatic speech recognition (ASR) system works well under clean environmental situation [1] [2]. However, 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 models with various noisy databases to cover different kinds of noise environment [3]. Unfortunately, such technique fails in face of unknown noise condition [4]. 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]. 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.

ASR speech enhancement aims to improve the quality of noisy speech input at runtime to reduce the mismatch with the trained acoustic models. In 1991, Hanson and Clements introduced a constrained iterative enhancement for speech recognition [6], where an iterative Wiener filtering with vocal tract spectral constraints was formulated using interframe and intraframe constraints based on line spectral pair transformation. The performance evaluation was based on 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 spectral floor into *a priori* SNR and *a posteriori* SNR with respect to frame SNR [7]. Significant improvement is reported for Aurora speech recognition system. In 2008, Breithaupt et al proposed a cepstral-domain smoothing method for estimation of *a priori* SNR [8], 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 is 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 [9] instead of the discrete Fourier transform (DFT) domain for noisy speech recognition. The performance was investigated on the standard Aurora speech recognition platform [10]. In 2010, Paliwal et al investigated the role of speech enhancement in speech recognition [11] where the experiments were conducted on the TIMIT speech corpus, however, there is no solution provided for the artificial distortion caused by the investigated speech estimators against the speech recognition; and also the speech recognition decoder is only based on small GMM-HMM and a bigram language model. We observed that the enhancer may be helpful for certain speech decoder but not always contribute to another speech decoder, therefore it is meaningful to investigate the performance with a typical state-of-the-art decoding platform.

In [11], 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 gives a poor performance in speech recognition. Therefore, a speech enhancement algorithm may significantly improve human listening experience [12] [13] [14], direct application of the enhancement algorithm does not always work well for speech recognition system. Ephraim-Malah's LSA MMSE [15] is a typical spectral-domain enhancement method which may represent the conventional spectral-domain MMSE speech enhancement techniques [16] [17] [18]. In this paper, we choose the LSA speech estimator for modification, and subsequently propose to improve the ASR speech enhancement system for the purpose of the speech recognition in the following aspects: the noise overestimation control, weak spectral component flooring, oversuppression of unwanted residual noise, and a reestimation of *a priori* SNR.

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 [19]. We build up the 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 the remainder of the paper, we give a brief introduction of the spectral-domain LSA speech enhancement algorithms used in this paper in section 2. In section 3, we propose a series of modification schemes for the speech estimators against speech recognition. The evaluation is shown in section 4 and finally the conclusion is given in section 5.

2. LSA SPEECH ENHANCEMENT ALGORITHM

2.1. LSA Speech Estimator

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.,

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

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.

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 $|\hat{S}_k(l)| = \exp\{\mathbf{E}[\ln|S_k(l)| \mid X_k]\}$. In this regard, Ephraim and Malah derived the gain function of the LSA-MMSE estimator [15]

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

where υ_k is given by

$$\upsilon_k(l) = \frac{\xi_k(l)}{1+\xi_k(l)}\gamma_k(l). \tag{3}$$

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

$$\xi_k(l) = \frac{\eta_s(l,k)}{\eta_n(l,k)}, \quad \gamma_k(l) = \frac{|X_k(l)|^2}{\eta_n(l,k)} \tag{4}$$

where $\eta_n(k,l) = \mathbf{E}[|N_k(l)|^2]$ and $\eta_s(k,l) = \mathbf{E}[|S_k(l)|^2]$ are the respective variances of the kth spectral components of noise and speech within the l-th time frame.

The conventional decision-directed estimation of the *a priori* SNR is given by [16]

$$\hat{\xi}_k(l) = \alpha\frac{|G_k(l-1)X_k(l-1)|^2}{\eta_n(k,l)} + (1-\alpha)\max[\gamma_k(l)-1, \quad 0]. \tag{5}$$

The smoothing factor α is conventionally set to 0.98 [16] [17] [18] [12]. which is found to be much less annoying and disturbing for human listening. However, we observed that the best performance of speech recognition is no long with 0.98, but is in the range of 0.7 $\sim$ 0.9.

2.2. About Noise Estimation

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.

There are many noise estimation methods for speech enhancement purpose. The typical methods include minimum statistics tracking [20] a minimum searching with speech presence probability (SPP) [21], MMSE-based noise estimation method [22] [23], and the SPP MMSE noise estimation method [21]. Through many experiments, we observed that the WER performance of SPP MMSE-based noise estimation [24] outperforms both minimum statistics in [20] and MMSE-based noise estimation in [23] 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 [24] 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 precious 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 results for the performance comparison among different speech estimators is of meaningful value.

3. PROPOSED MODIFICATION SCHEME FOR SPEECH RECOGNITION

For the purpose of speech recognition, we attempt to improve the spectral-domain LSA 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.

3.1. Smoothing Adaptation for Noise Control and Weak Spectral Floor

In MMSE estimation, statistical independence assumption leads to the following subsequence: the spectral gain of a frequency bin is only a function of *a priori* and *a posteriori* SNRs of the 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 [17] [18] [7]. The frame SNR can be approximated by using the following equation [25]

$$\Xi(l) = 10\log_{10}\max\left\{\frac{\sum_k\left\{\max[(|X_k(l)| - \sqrt{\eta_n(l,k)}\,), \quad 0]\right\}^2}{\sum_k \eta_n(l,k)}, \varepsilon\right\} \tag{6}$$

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

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 [7], 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 broken segmental linear relationship.

In this paper, we propose to replace the broken linear relationship with a smoothing relationship. Let ϕ denote a general sigmoid function below

$$\phi(x, r_1, r_2) = \frac{1}{1+\exp\{(x-r_1)/r_2\}} \tag{7}$$

and we make the noise control factor ρ to be adapted by frame SNR $\Xi(l)$ as follows

$$\rho(l) = \tau_1\phi^2(\Xi(l), s_1, s_2) + \tau_2 \tag{8}$$

where τ_1, τ_2, s_1 and s_2 are constants.

In this paper, we empirically set s_1=13.5, s_2=5, τ_1=2.6, and τ_2=0.001. Fig. 1 (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.

If the information carried by the weak spectrum can be safely transferred, the performance of speech recognition will be significantly improved. In [7], a spectral floor is introduced into the *a priori* and *a posteriori* SNRs for weak spectrum component and silence period. However, the adaptation of spectral floor is also 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 smoothly adapted by the frame SNR as follows

$$\varsigma(l) = (1+\kappa) - \phi^2(\Xi(l), k_1, k_2) \tag{9}$$

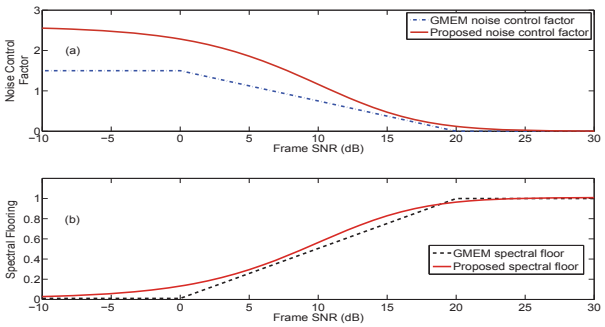

Fig. 1. *Noise control factors and weak spectral floor adapted to frame SNR.*

where κ is the lower bound of the flooring factor. Empirically we set κ=0.01, k_1=13.5, k_2=5. Fig. 1 (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

$$\breve{\gamma}_k(l) = \begin{cases} \frac{|X_k(l)|^2}{\rho(l)\eta_n(l)}, & \text{if } \frac{|X_k(l)|^2}{\rho(l)\eta_n(l)} \geq \varsigma(l)+1; \\ \varsigma(l)+1, & \text{otherwise} \end{cases} \tag{10}$$

and the *a priori* SNR is modified below

$$\breve{\xi}_k(l) = \begin{cases} \tilde{\xi}_k(l), & \text{if } \tilde{\xi}_k(l) \geq \varsigma(l); \\ \varsigma(l), & \text{otherwise} \end{cases} \tag{11}$$

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

$$\tilde{\xi}_k(l) = \alpha \frac{|\breve{G}_k(l-1)X_k(l-1)|^2}{\rho(l)\eta_n(k,l)} + (1-\alpha)(\breve{\gamma}_k(l)-1) \tag{12}$$

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 $\breve{\gamma}_k(l-1)$, which totally depends on the parameters in the previous frame.

3.2. Oversuppression of Residual Noise

It has been known that musical noise can very apparently appear in spectral suppression using spectral subtraction method [12]. 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.

We observed that 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. Therefore, we propose to suppress the residual noise by introducing a adaptive smoothing oversuppression factor as follows

$$\omega(l) = 1 + (\varpi - 1)\phi^2(\Xi(l), w_1, w_2) \tag{13}$$

where $\varpi = 0.1$ is the lower bound of the gain control factor, w_1=-3, and w_2=2. Subsequently, the gain is modified as follows

$$\bar{\breve{G}}_k(l) = \omega(l)\breve{G}_k(l) \tag{14}$$

Fig. 2 shows the oversuppression factor ω adapted to frame SNR. When $\Xi \gg 0$ dB, the gain is unchanged since $\omega(l)$ = 1. When the frame SNR is lower than 0 dB, the factor is debated to much low according to level of current frame SNR. The adaptive oversuppression brings an obvious improvement for speech recognition. It is believed that the adaptive oversuppression make the endpoints more correctly be aligned during speech recognition.

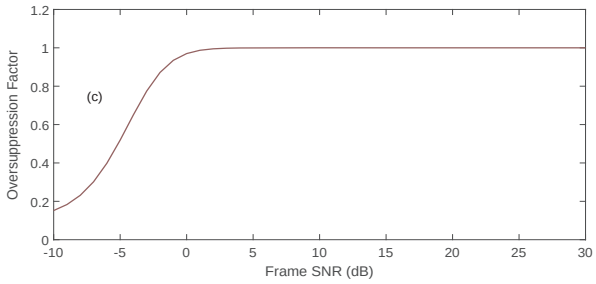

Fig. 2. *Oversuppression factor adapted to frame SNR.*

3.3. 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 (11) as follows

$$\tilde{\xi}_k(l) = \max\left[\alpha_c \frac{|\breve{G}_k(l)X_k(l)|^2}{\rho(l)\eta_n(k,l)} + (1-\alpha_c)(\breve{\gamma}_k(l)-1), \quad \varsigma(l)\right]. \tag{15}$$

Experiment shows that the reestimation reduces the feature distortion for speech recognition, and the best result falls on α_c = 1. Subsequently, the reestimation of the *a priori* SNR is only based its definition in (4).

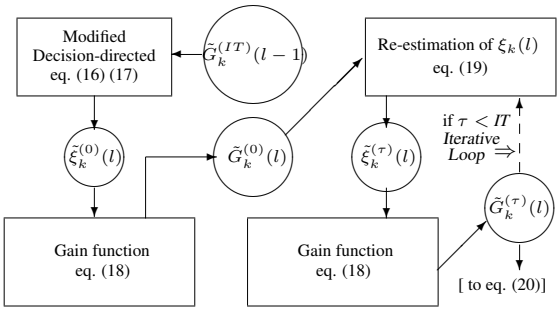

Fig. 3. *The scheme of the proposed estimation of the a priori SNR.*

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

$$\tilde{\xi}_k(l) = \breve{\alpha} \frac{|\omega(l)\tilde{G}_k^{(IT)}(l-1)X_k(l-1)|^2}{\rho(l)\eta_n(k,l)} + (1-\breve{\alpha})(\breve{\gamma}_k(l)-1) \tag{16}$$

$$\tilde{\xi}_k^{(0)}(l) = \begin{cases} \breve{\xi}_k(l), & \text{if } \breve{\xi}_k(l) \geq \varsigma(l); \\ \varsigma(l), & \text{otherwise} \end{cases} \tag{17}$$

$$\tilde{G}_k^{(\tau-1)}(l) = G_k(\tilde{\xi}_k^{(\tau-1)}(l), \breve{\gamma}_k(l)), \qquad \tau = 1, ..., IT. \tag{18}$$

Table 1. *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%

$$\tilde{\xi}_k^{(\tau)}(l) = \max\big[\frac{|\tilde{G}_k^{(\tau-1)}(l)X_k(l)|^2}{\rho(l)\eta_n(k,l)}, \quad \varsigma(l)\big]. \tag{19}$$

Here, the smoothing factor $\breve{\alpha}$ is empirically set to 0.7. As a result, the estimate of speech spectrum is given by

$$\hat{S}_k(l) = \omega(l)\tilde{G}_k^{(IT)}(l)X_k(l). \tag{20}$$

Fig. 3 shows the flow chat of the proposed *a priori* SNR reestimation scheme.

4. PERFORMANCE EVALUATION

4.1. Models for Speech Recognition

A state-of-the-art speech recognizer is setup by elaborating the language modeling and acoustic modeling process as follows: we used 39 dimensional MFCC feature for speech recognition system, and applied RASTA (relative spectra), CMVN, LDA, MLLT and fMLLR feature enhancement techniques. The DNN acoustic model is trained by using 260k utterances (313 hr 23 min) from Switchboard-1-LDC97S62 database with Kaldi toolkit [19]. With sMBR criterion [1] [26], we obtained the DNN-HMM-sMBR model of five hidden layers with 2048 neurons each hidden layer. 1,831 English sentences with 21,395 words from the Switchboard corpus are selected as the test dataset. On the other side, the language model is trained with lexicon of 30,858 vocabulary size by using Part 1 of Fisher transcripts that are equivalent to 700 hours of speech with SRILM toolkit [27].

Different type of noise from NOISEX-92 [28] are added into the test speech database to generate different group of noisy speech database with different global SNRs. In this paper, we select three types of noises, i.e. white noise, F16 noise and Factory1 noise.

Let **P1** denote our proposed smoothing adaptation of the noise control factor and weak spectral floor, **P2** denote the proposed over-suppression method, and **P3** denote the proposed *a priori* SNR reestimation method.

To study the contributions of **P1**, **P2** and **P3**, Table 1 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**. 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. According to the above observation, we adopt only one-time reestimation (i.e. *IT=1*) in the next experiment.

The proposed speech enhancement algorithm is compared with conventional spectral SNR estimation. The particular algorithms are listed below: **Wiener**: Conventional Wiener filter [29], **LSA**: Conventional LSA filter [15], **LSA:CEP**: LSA with cepstral *a priori* SNR [8], **LSA:GMEM**: Gemello's modified E-M LSA [7], **ETSI**: ETSI baseline [30], **LSA:PRO**: LSA with **P1**+**P2**+**P3**(*IT*=1).

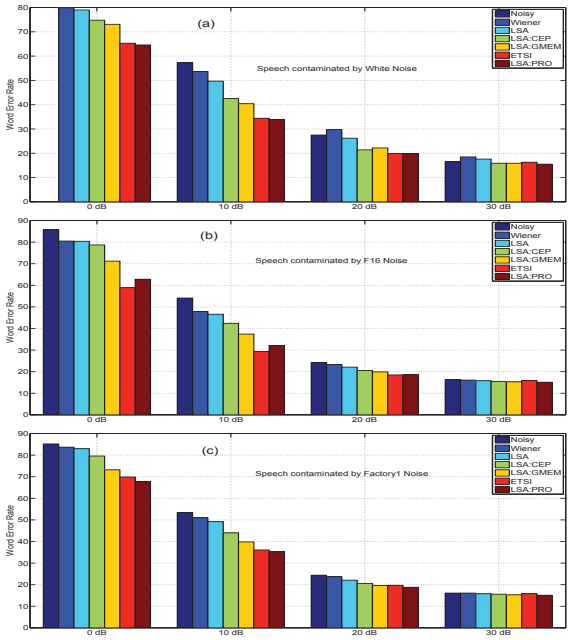

Fig. 4. *WER with DNN-sMBR decoder by applying different denoising algorithms on the databases with different noise types (a) White noise; (b) F16 noise; (c) Factory1 noise in different SNRs, i.e. 0 dB, 10 dB, 20 dB and 30 dB.*

Fig. 4 shows the WER of the different speech enhancement systems with different noise conditions using the DNN-sMBR decoder.

We can see that our proposed **LSA:PRO** is almost consistently better than **LSA:CEP** and **LSA:GMEM**. LSA is consistently better than Wiener filter.

It is shown that the **ETSI** is a powerful noise reduction system, especially in the low SNR situation. This advantage is very apparent for F16 noise, WER from 85.7% drops to 58.9%, reaches 31.27% improvement for 0 dB, although **LSA:PRO** makes a significant improvement and let WER dropped to 62.8%, with 26.7% improvement. However, **LSA:PRO** outperforms **ETSI** in White and Factory1 noises with both low SNR and high SNR. And it wins most of best accuracy totally. For the case of 10 dB Factory1 noise, it drops WER from 53.4% to 35.2% to gain 34.1% improvement.

The experiment makes evident that our proposed scheme brings positive and effective progress for the ASR denoising. Actually, we have applied the same modification scheme on other spectral-domain speech enhancement methods such as Wiener filtering [29], MMSE [16], β-order MMSE [17] and masking-based β-order MMSE [18]. The improvement with the proposed modification on the above-mentioned speech estimators for ASR is also very significant.

5. CONCLUSION

We have introduced a series of modification on LSA-MMSE speech enhancement to mitigate the artifacts for speech recognition. In particular, we have proposed smoothing-adaptation scheme for controlling the processing noise power and mitigating the harmful artifacts for weak speech signal, over-suppressing the residual noise, and re-estimating *a priori* SNR. We analyzed the effectiveness of the proposed modification scheme, and the experimental result shows that the proposed scheme is significantly effective for the state-of-the-art speech recognition.

6. REFERENCES

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

[2] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition," Feb. 2014. [Online]. Available: http://arxiv.org/abs/1402.1128

[3] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multi-style training for robust isolated word speech recognition," *Proc. of ICASSP*, pp. 705-708, Dallas, TX, USA, 1987.

[4] J. Ming, B. Hou, "Speech Recognition in Unknown Noisy Conditions." in book *Robust Speech Recognition and Understanding*, Chapter 11, M. GrimmandK. Kroschel (eds.), I-TECH Education and Publishing, pp. 175186, 2007.

[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. 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.

[7] 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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

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

[28] 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.

[29] 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.

[30] 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