Improving Denoising Auto-encoder Based Speech Enhancement With the Speech Parameter Generation Algorithm

Syu-Siang Wang*‡, Hsin-Te Hwang†, Ying-Hui Lai‡, Yu Tsao‡, Xugang Lu§, Hsin-Min Wang† and Borching Su*
* Graduate Institute of Communication Engineering, National Taiwan University, Taiwan E-mail: d02942007@ntu.edu.tw
† Institute of Information Science, Academia Sinica, Taipei, Taiwan E-mail: hwanght@iis.sinica.edu.tw
‡ Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan E-mail: yu.tsao@citi.sinica.edu.tw
§ National Institute of Information and Communications Technology, Japan

***Abstract*—This paper investigates the use of the speech parameter generation (SPG) algorithm, which has been successfully adopted in deep neural network (DNN)-based voice conversion (VC) and speech synthesis (SS), for incorporating temporal information to improve the deep denoising auto-encoder (DDAE)-based speech enhancement. In our previous studies, we have confirmed that DDAE could effectively suppress noise components from noise corrupted speech. However, because DDAE converts speech in a frame by frame manner, the enhanced speech shows some level of discontinuity even though context features are used as input to the DDAE. To handle this issue, this study proposes using the SPG algorithm as a post-processor to transform the DDAE processed feature sequence to one with a smoothed trajectory. Two types of temporal information with SPG are investigated in this study: static-dynamic and context features. Experimental results show that the SPG with context features outperforms the SPG with static-dynamic features and the baseline system, which considers context features without SPG, in terms of standardized objective tests in different noise types and SNRs.**

I. Introduction

A primary goal of speech enhancement (SE) is to reduce noise components, and thus enhance the signal-to-noise ratio (SNR) of noise-corrupted speech. In a wide range of voice communication applications, SE serves as a key element to increase the quality and intelligibility of speech signals [1], [2], [3]. Generally, SE algorithms can be classified into two categories: unsupervised and supervised ones. The unsupervised algorithms are derived by probabilistic models of speech and noise signals. Notable examples include spectral subtraction [4], Wiener filter [5], Kalman filtering [6], and minimum mean-square-error (MMSE) spectral estimator [7]. These methods assume statistical models for speech and noise signals. The clean speech is estimated from the noisy observation without any prior information on the noise type or speaker identity. One limitation of these approaches is that accurate estimation of noise statistics can be very challenging, especially when the noise is non-stationary. In contrast, the supervised algorithms require a set of training data to learn a transformation structure to facilitate an online SE process. When a sufficient amount of training data is available, the supervised methods can achieve better performance than the unsupervised counterparts [8], [9], [10], [11]. Notable supervised SE algorithms include nonnegative matrix factorization (NMF) [9], [12], sparse coding [13], deep neural network (DNN) [11], [14], and deep denoising auto-encoder (DDAE) [8], [15] algorithms.

The DDAE based SE method includes training and enhancement phases. In the training phase, we need to prepare paired clean and noisy training speech utterances, which are set as the output and input of the DDAE model, respectively. The parameters in the DDAE model are learned based on the minimal mean square error (MMSE) criterion with the goal of transforming the noisy speech to match the clean one. In the enhancement phase, DDAE transforms the noisy speech to the enhanced one using the parameters learned in the training phase. From previous studies, the DDAE approach can effectively remove noise components from noise-corrupted speech and provide better performance in terms of several standardized objective evaluation metrics, compared to conventional SE approaches [8], [15]. However, because DDAE transforms acoustic features in a frame-by-frame manner, the enhanced speech shows some level of discontinuity even though context features are used as input to the DDAE model. In this study, we intend to incorporate the temporal trajectory information of a speech utterance to overcome the discontinuity issue of DDAE.

The discontinuity issue is also found in the DNN-based speech synthesis (SS) [16] and DNN-based voice conversion (VC) [17] tasks. Several approaches have been proposed to overcome it, among them an effective approach is the speech parameter generation (SPG) algorithm. The SPG algorithm was first proposed for the hidden Markov model (HMM)-based SS [18], [19], and later was applied in the Gaussian mixture model (GMM)-based VC [20], DNN-based VC [17], [21], and DNN-based SS [16]. Previous studies have confirmed that two types of features are effective in covering temporal information, namely static-dynamic features and context features. The static-dynamic features are obtained by appending dynamic components to the original static ones, while the context features are prepared by attaching adjacent features to the center ones. The SPG algorithm generates speech with smooth temporal trajectories by using the dynamic or contextual features as constraints in the speech generation process. In this study, we use SPG as a post-processor to transform the DDAE enhanced feature sequence to one with a smoothed trajectory. To conduct the static-dynamic-feature-based SPG, we use the static-dynamic features of the noisy speech and clean speech as the input and output of the DDAE model. Similarly, for the context-feature-based SPG, the context features of the noisy speech and clean speech are used as the input and output of the DDAE model. Experimental results show that the SPG smoothed DDAE model with context features achieves better performance than the SPG smoothed DDAE model with static-dynamic features. The results also confirm that DDAE with SPG always outperforms the baseline system (i.e., DDAE without SPG) in various standardized objective tests in different noise types and SNRs.

The remainder of the paper is organized as follows. The DDAE SE system is briefly introduced in Section II. The proposed SPG smoothed DDAE SE framework and experimental evaluations are presented in Sections III and IV, respectively. Finally, the summaries of our findings are given in Section V

II. The Deap Denoising Auto-Encoder

This section reviews the DDAE speech enhancement system. Brief mathematical derivations are also provided.

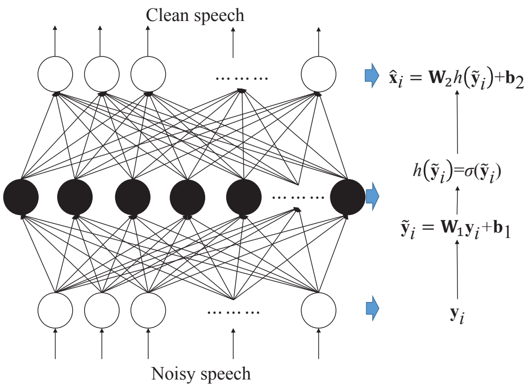

Clean speech
………
$\hat{\mathbf{x}}_i = \mathbf{W}_2 h(\tilde{\mathbf{y}}_i)+\mathbf{b}_2$
$h(\tilde{\mathbf{y}}_i)=\sigma(\tilde{\mathbf{y}}_i)$
………
$\tilde{\mathbf{y}}_i = \mathbf{W}_1\mathbf{y}_i+\mathbf{b}_1$
………
$\mathbf{y}_i$
Noisy speech

Fig. 1. One hidden layer DAE model. $\mathbf{y}_i$ and $\hat{\mathbf{x}}_i$ denote the i-th training sample of the noisy and enhanced speech, respectively.

The DDAE-based speech enhancement method consists of two phases, namely the offline and online phases. The offline phase first prepares paired clean speech and noisy speech utterances, which are used as the output and input of the DDAE model, respectively. The parameters in the DDAE model are estimated by the MMSE criterion with the aim of perfectly transforming noisy speech to the clean one. With the estimated DDAE model parameters, the noisy utterances are reconstructed to the enhanced one in the online phase.

Figure 1 shows the block diagram of a one-layered denoising auto-encoder (DAE). In the figure, the DAE outputs the enhanced feature vector $\hat{\mathbf{x}}_i$ by:

$$\hat{\mathbf{x}}_i = \mathbf{W}_2 h(\tilde{\mathbf{y}}_i) + \mathbf{b}_2, \tag{1}$$

where $\mathbf{W}_2$ and $\mathbf{b}_2$ are the connecting weights and bias vectors for the reconstruction stage, and $h(\tilde{\mathbf{y}}_i)$ is obtained by

$$h(\tilde{\mathbf{y}}_i) = \sigma(\tilde{\mathbf{y}}_i) = \frac{1}{1 + exp(-\tilde{\mathbf{y}}_i)}, \tag{2}$$

with

$$\tilde{\mathbf{y}}_i = \mathbf{W}_1 \mathbf{y}_i + \mathbf{b}_1, \tag{3}$$

where $\mathbf{W}_1$ and $\mathbf{b}_1$ are the connecting weights and bias vectors for the encoding stage.

Parameters $\{\theta \mid \theta \in \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2\}$ were determined by optimizing the objective function in (4) through all the training sample vectors.

$$\theta^* = \underset{\theta}{argmin}(L(\theta) + \alpha\psi(\mathbf{W}_1, \mathbf{W}_2) + \beta\phi(h(\tilde{\mathbf{y}}_i), \mathbf{y}_i)\}, \tag{4}$$

where α and β are the weight decay and sparse penalty parameters, respectively; $\psi(\mathbf{W}_1, \mathbf{W}_2) = (\parallel \mathbf{W}_1 \parallel_F^2 + \parallel \mathbf{W}_2 \parallel_F^2)$; $\phi(h(\tilde{\mathbf{y}}_i), \mathbf{y}_i)$ denotes the sparsity constraint, where the KullbackLeibler (KL) divergence [22] between two Bernoulli distributions is used in this study; and $L(\theta)$ is the distance between clean- and reconstructed feature vectors defined as

$$L(\theta) = \sum_{i=1}^{I} \parallel \mathbf{x}_i - \hat{\mathbf{x}}_i \parallel_2^2, \tag{5}$$

where I is the total number of training samples. DDAE is a deap DAE consisting of more layers.

III. The Proposed DDAE with SPG Method

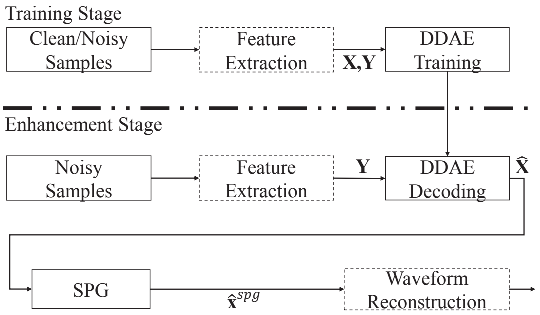


Fig. 2. The proposed SPG smoothed DDAE (DAS) speech enhancement architecture.

Figure 2 shows the block diagram of the proposed DDAE with SPG (denoted as DAS) SE technique.

A. The training stage

At the training stage, after feature extraction, the noisy speech feature vector $\mathbf{Y} = [\mathbf{Y}_1^\intercal, \cdots, \mathbf{Y}_i^\intercal, \cdots, \mathbf{Y}_I^\intercal]^\intercal$ and clean speech feature vector $\mathbf{X} = [\mathbf{X}_1^\intercal, \cdots, \mathbf{X}_i^\intercal, \cdots, \mathbf{X}_I^\intercal]^\intercal$ are used as input and output to the DDAE, respectively, for constructing the model. The superscript $\intercal$ denotes the vector transposition; $\mathbf{Y}_i$ and $\mathbf{X}_i$ are the noisy and clean speech feature vectors at frame i, respectively. Both feature vectors $\mathbf{Y}_i$ and $\mathbf{X}_i$ could be either composed by their static-dynamic or context features. For example, if $\mathbf{Y}_i$ consists of static-dynamic features, then $\mathbf{Y}_i = [\mathbf{y}_i^\intercal, \Delta^{(1)}\mathbf{y}_i^\intercal, \Delta^{(2)}\mathbf{y}_i^\intercal]^\intercal$. The velocity $\Delta^{(1)}\mathbf{y}_i$ and acceleration $\Delta^{(2)}\mathbf{y}_i$ features can be calculated from the static features $\mathbf{y}_{i-1}$, $\mathbf{y}_i$, and $\mathbf{y}_{i+1}$ by

$$\begin{aligned} \Delta^{(1)}\mathbf{y}_i &= \frac{\mathbf{y}_{i+1} - \mathbf{y}_{i-1}}{2}, \\ \Delta^{(2)}\mathbf{y}_i &= \mathbf{y}_{i-1} - 2\mathbf{y}_i + \mathbf{y}_{i+1}. \end{aligned} \tag{6}$$

Similarly, $\mathbf{X}_i = [\mathbf{x}_i^\intercal, \Delta^{(1)}\mathbf{x}_i^\intercal, \Delta^{(2)}\mathbf{x}_i^\intercal]^\intercal$ can be obtained accordingly. If $\mathbf{Y}_i$ and $\mathbf{X}_i$ are composed by the context features, the adjacent n static features are concatenated together with the current static feature frame as $\mathbf{Y}_i = [\mathbf{y}_{i-n}^\intercal, \cdots, \mathbf{y}_i^\intercal, \cdots, \mathbf{y}_{i+n}^\intercal]^\intercal$ and $\mathbf{X}_i = [\mathbf{x}_{i-n}^\intercal, \cdots, \mathbf{x}_i^\intercal, \cdots, \mathbf{x}_{i+n}^\intercal]^\intercal$, respectively. In this paper, n is set to one to make both context and static-dynamic feature vectors have the same dimension and contain the same amount of contextual information. Finally, the DDAE model of the proposed DAS system can be constructed based on the steps presented in Section II. It should be noted that the major difference between building the baseline DDAE model (described in Section II) and the DDAE model of the proposed DAS system is that the baseline system uses the context features as input and only the static features as output to the DDAE for constructing the model, while the proposed system uses the static-dynamic or context features as input and output to the DDAE for constructing the model.

B. The enhancement stage

In the enhancement stage, after feature extraction from an utterance with a total number of $\hat{I}$ frames, the noisy speech feature vector $\mathbf{Y}$ is first transformed by a well trained DDAE model into the enhanced speech feature vector $\hat{\mathbf{X}} = [\hat{\mathbf{X}}_1^\intercal, \cdots, \hat{\mathbf{X}}_{\hat{i}}^\intercal, \cdots, \hat{\mathbf{X}}_{\hat{I}}^\intercal]^\intercal$. Both feature vectors $\mathbf{Y}_{\hat{i}}$ and $\hat{\mathbf{X}}_{\hat{i}}$ could be either composed by their static-dynamic features (e.g., $\hat{\mathbf{X}}_{\hat{i}} = [\hat{\mathbf{x}}_{\hat{i}}^\intercal, \Delta^{(1)}\hat{\mathbf{x}}_{\hat{i}}^\intercal, \Delta^{(2)}\hat{\mathbf{x}}_{\hat{i}}^\intercal]^\intercal$) or context features (e.g., $\hat{\mathbf{X}}_{\hat{i}} = [\hat{\mathbf{x}}_{\hat{i}-1}^\intercal, \hat{\mathbf{x}}_{\hat{i}}^\intercal, \hat{\mathbf{x}}_{\hat{i}+1}^\intercal]^\intercal$). Then, the SPG algorithm with the dynamic or context constraint is employed to generate a smooth static feature vector in the same manner as it is applied to DNN-based VC [17], [21], [23] and SS [16],

$$\hat{\mathbf{x}}^{spg} = f(\hat{\mathbf{X}}; \mathbf{U}^{-1}, \mathbf{M}) = (\mathbf{M}^\intercal\mathbf{U}^{-1}\mathbf{M})^{-1}\mathbf{M}^\intercal\mathbf{U}^{-1}\hat{\mathbf{X}}, \tag{7}$$

where the matrix $\mathbf{U}^{-1}$ is composed by the pre-estimated covariance matrix $\mathbf{\Sigma}^{(\mathbf{XX})}$ from clean training speech features as

$$\mathbf{U}^{-1} = diag[(\mathbf{\Sigma}_1^{(\mathbf{XX})})^{-1}, \cdots, (\mathbf{\Sigma}_{\hat{i}}^{(\mathbf{XX})})^{-1}, \cdots, (\mathbf{\Sigma}_{\hat{I}}^{(\mathbf{XX})})^{-1}]. \tag{8}$$

The covariance matrix $\mathbf{\Sigma}^{(\mathbf{XX})}$ can be assumed diagonal, i.e.,

$$\mathbf{\Sigma}^{(\mathbf{XX})} = diag[\sigma^2(1), \cdots, \sigma^2(d), \cdots, \sigma^2(D)], \tag{9}$$

where $\sigma^2(d)$ is the sample variance of the d-th dimension of the clean speech feature vector. Finally, $\mathbf{M}$ in (7) is a weighting matrix for appending the dynamic features to the static ones when the enhanced speech feature vector $\hat{\mathbf{X}}$ is composed by the static-dynamic features. Similarly, $\mathbf{M}$ can be derived accordingly when the enhanced speech feature vector $\hat{\mathbf{X}}$ is composed by the context features.

IV. Experiments

A. Experimental setup

The experiments were conducted on a Mandarin hearing in noise test (MHINT) database. The database included 320 utterances, which were pronounced by a male native-Mandarin speaker in a clean condition and recorded at 16 kHz sampling rate. These utterances were further down sampled to 8 kHz to form a more challenging task. Among 320 utterances, we selected 250 and 50 utterances as the training and testing data, respectively. Two types of noises, namely the car and pink noises, were artificially added to the clean utterances to generate 20, 10, 5, 0 and -5 dB signal to noise (SNR) noisy speech utterances. The car noise (engine noise) is band-limited and corrupts the low frequency speech content, while the pink noise corrupts larger portions of the speech spectrum. In addition, a DDAE model, consisting of three hidden layers and 300 nodes in each layer, was trained by 250 paired clean and noisy speech utterances for each SNR and noise type. In the testing phase, each trained DDAE model was then used to transform the 50 noisy utterances to the enhanced ones.

The speech waveform was first windowed by 256 samples (32 msec) with 128 samples window shift. The samples in each window were then transformed to log-scale power spectrum coefficients. As mentioned earlier, both DDAE and DAS were implemented with both the static-dynamic (SD) and context (CX) features.

B. Evaluation methods

In this study, all the SE systems were evaluated by (1) the quality test in terms of the perceptual evaluation of speech quality (PESQ) [24] and hearing aids speech quality index (HASQI) [25], (2) the perceptual test in terms of the hearing aids speech perception index (HASPI) [26], and (3) the speech distortion index (SDI) test [27]. The score ranges of PESQ, HASQI, and HASPI are {-0.5 to 4.5}, {0 to 1}, and {0 to 1}, respectively. Higher scores of PESQ and HASQI denote better sound quality and higher scores of HASPI represent better intelligibility. On the other hand, the SDI measures the degree of speech distortion, and a lower SDI value indicates a smaller distortion and thus better performance.

C. Performance evaluation

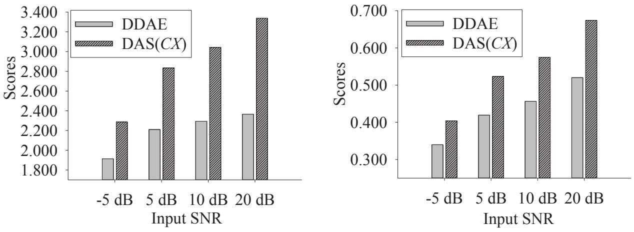
For each test utterance in a specific condition, we could obtain a PESQ, HASQI, HASPI, and SDI score, respectively. We reported the evaluation results by averaging the scores of the 50 test utterances. In the first experiment, we compare the baseline DDAE method [8] and the conventional MMSE method [7]. Table I shows the scores of the unprocessed noisy speech, MMSE enhanced speech, and baseline DDAE enhanced speech in the 0 dB SNR condition averaged over two noise types (i.e., the car and pink noises). Here, the baseline DDAE system uses the context features (the concatenation of

TABLE I
AVERAGE SCORES OF THE UNPROCESSED NOISY SPEECH, MMSE ENHANCED SPEECH, AND DDAE ENHANCED SPEECH IN THE 0 dB SNR CONDITION.

Methods	**Noisy**	**MMSE**	**DDAE**
PESQ	1.690	1.921	**2.270**
HASQI	0.128	0.137	**0.387**
HASPI	0.665	0.761	**0.998**
SDI	0.564	0.349	**0.327**

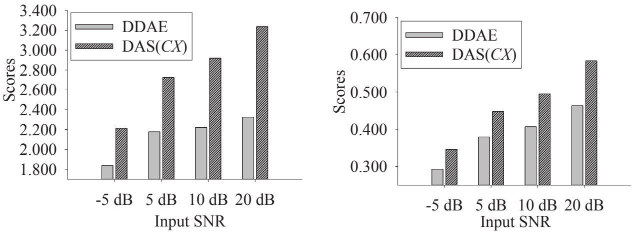
TABLE II
AVERAGE RESULTS OF DDAE, DAS(SD), AND DAS(CX) IN THE 0 dB SNR CONDITION.

Methods	**DAS(SD)**	**DAS(CX)**
PESQ	2.349	**2.550**
HASQI	0.379	**0.432**
HASPI	0.997	**0.999**
SDI	0.319	**0.226**



(a) PESQ (b) HASQI

Fig. 3. Scores of (a) PESQ and (b) HASQI for DDAE and DAS(CX) enhanced speech utterances in -5, 5, 10, and 20dB SNRs on car noise condition.



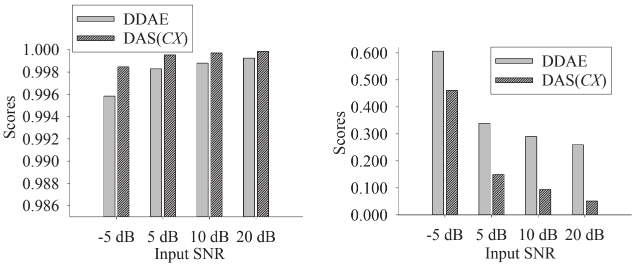
(a) PESQ (b) HASQI

Fig. 4. Scores of (a) PESQ and (b) HASQI for DDAE and DAS(CX) enhanced speech utterances in -5, 5, 10, and 20dB SNRs on pink noise condition.

the static features of the current frame and the immediately preceding and following frames) as the DDAE input, and the static features of the current frame as the DDAE output. From Table I, it can be seen that the objective scores obtained by the MMSE enhanced speech is obviously better than those of the unprocessed noisy speech, i.e., with higher PESQ and HESQI scores (quality test), a higher HASPI score (perception test), and a lower SDI value (distortion measurement). Moreover, the DDAE method outperforms the MMSE method in all metrics, which is consistent to the results in our previous study [8]. The evaluation results for other SNR levels are with similar trends. Therefore, we do not show that set of results.

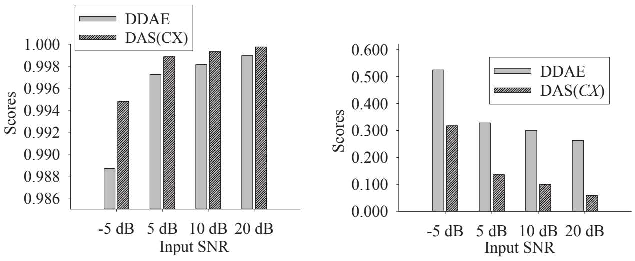
Next, we evaluate the use of the static-dynamic and context features in the proposed DAS SE system. The results are shown in Table II. The testing condition is the same as that in Table I, i.e., we report the average scores over the car and pink noises in the 0 dB SNR condition. The DAS system with the static-dynamic features is denoted as DAS(SD) while the DAS system with the context features is denoted as DAS(CX). From Table II, we can see that DAS(CX) outperforms DAS(SD) in all objective metrics even though the CX features contain the same amount of contextual information with the SD features. A possible explanation for this is given by a previous study on a VC task [23]. In that study, it was shown that the context features (called multiple frames in [23]) have higher inter-dimensional correlation than the static-dynamic features. As a result, the context features might be more suitable for the neural networks (NN) because NN is widely believed to be good at modeling features with strong inter-dimensional correlation. Moreover, from Tables I and II, it is obvious that both DAS(CX) and DAS(SD) outperform the baseline DDAE system, confirming the effectiveness of using the SPG algorithm for further processing the DDAE enhanced speech. This result demonstrates that the proposed DAS system (i.e., the DDAE with SPG system) can produce a smoother static feature sequence than the baseline DDAE system (i.e., the DDAE system without SPG) due to that the dynamic or contextual constraint is considered when the SPG algorithm is applied.

Finally, to further analyze the effectiveness of DAS(CX), which achieves the best performance in Table II, we compare it with the baseline DDAE system in different noise conditions. Figures 3 and 4 show the scores of PESQ and HASQI in different SNRs (including -5, 5, 10 and 20 dB) over car and pink noise conditions, respectively. From both figures, we first observe that both PESQ and HASQI obtained by DAS(CX) and DDAE are consistently increased along with SNRs. The result reveals that the objective measures (PESQ and HASQI) for both systems are positively correlated with the SNR values. In addition, DAS(CX) always outperforms DDAE in all SNRs. Particularly, as shown in figure 3 (a), DAS(CX) achieves a significant improvement over DDAE on the PESQ score. The result demonstrates that the quality (measured by PESQ and HASQI) of the DDAE enhanced speech can be consistently improved by the SPG algorithm. From figures 5 and 6, similar trends can be found that the performances of both DAS(CX) and DDAE are improved along SNR and DAS(CX) outperforms DDAE in all SNRs in terms of the HASPI and SDI measures. The result indicates that both the intelligibility (measured by HASPI) and the distortion (measured by SDI) of the DDAE enhanced speech can be further improved by the SPG algorithm. From figures 3 and 5, we conclude that the proposed DAS(CX) system consistently outperform the baseline DDAE and MMSE speech enhancement systems in different noise conditions.



(a) HASPI (b) SDI

Fig. 5. Scores of (a) HASPI and (b) SDI for DDAE and DAS(CX) enhanced speech utterances in -5, 5, 10, and 20dB SNRs on car noise condition.



(a) HASPI (b) SDI

Fig. 6. Scores of (a) HASPI and (b) SDI for DDAE and DAS(CX) enhanced speech utterances in -5, 5, 10, and 20dB SNRs on pink noise condition.

V. Conclusion

In this paper, we have proposed incorporating the SPG algorithm with the DDAE speech enhancement system to handle the discontinuity issue and intensively investigated the use of two types of temporal information, namely the static-dynamic and context features, in the SPG algorithm in terms of standardized objective tests in different noise types and SNRs. The experimental results on the MHINT speech corpus have demonstrated that the performance of the DDAE speech enhancement system can be further improved by employing a SPG post-processor and the context features achieve better improvements than the static-dynamic features. In the future work, we will evaluate the proposed DAS system on more noise types and SNRs. We will also apply the sequence error minimization criterion [16] in the DDAE and DAS speech enhancement systems.

VI. Acknowledgements

This work was supported by the Ministry of Science and Technology of Taiwan under contracts MOST103-2221-E-001-003.

References

[1] W. Hartmann, A. Narayanan, E. Fosler-Lussier, and D. Wang, "A direct masking approach to robust ASR," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 1993–2005, 2013.

[2] A. Stark and K. Paliwal, "Use of speech presence uncertainty with mmse spectral energy estimation for robust automatic speech recognition," *Speech Communication*, vol. 53, no. 1, pp. 51–61, 2011.

[3] S. Doclo, M. Moonen, T. Van den Bogaert, and J. Wouters, "Reduced-bandwidth and distributed mwf-based noise reduction algorithms for binaural hearing aids," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 38–51, 2009.

[4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[5] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *ICASSP*, pp. 629–632, 1996.

[6] V. Grancharov, J. Samuelsson, and B. Kleijn, "On causal algorithms for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 764–773, 2006.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[8] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder.," in *INTERSPEECH*, pp. 436–440, 2013.

[9] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[10] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactionson Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[11] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[12] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors.," in *ICASSP*, pp. 4029–4032, 2008.

[13] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech enhancement using generative dictionary learning," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1698–1712, 2012.

[14] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *ICASSP*, pp. 7092–7096, IEEE, 2013.

[15] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Ensemble modeling of denoising autoencoder for speech spectrum restoration," in *INTERSPEECH*, vol. 14, pp. 885–889, 2014.

[16] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (dnn) for parametric tts synthesis," in *ICASSP*, pp. 3829–3833, 2014.

[17] F.-L. Xie, Y. Qian, Y. Fan, F. K. Soong, and H. Li, "Sequence error (se) minimization training of neural network for voice conversion," in *INTERSPEECH*, 2014.

[18] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from hmm using dynamic features," in *ICASSP*, vol. 1, pp. 660–663, 1995.

[19] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *ICASSP*, vol. 3, pp. 1315–1318, 2000.

[20] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2222–2235, 2007.

[21] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[22] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[23] L.-H. Chen, Z.-H. Ling, and L.-R. Dai, "Voice conversion using generative trained deep neural networks with multiple frame spectral envelopes," in *INTERSPEECH*, 2014.

[24] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, vol. 2, pp. 749–752, 2001.

[25] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (hasqi)," *Journal of the Audio Engineering Society*, vol. 58, no. 5, pp. 363–381, 2010.

[26] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (haspi)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[27] J. Chen, J. Benesty, Y. Huang, and E. Diethorn, "Fundamentals of noise reduction in spring handbook of speech processing-chapter 43," 2008.