A LOCALLY LINEAR EMBBEDING BASED POSTFILTERING APPROACH FOR SPEECH ENHANCEMENT

Yi-Chiao Wu[*], *Hsin-Te Hwang*[*], *Syu-Siang Wang*[†], *Chin-Cheng Hsu*[*], *Ying-Hui Lai*[††], *Yu Tsao*[†], *and Hsin-Min Wang*[*]

[*]Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: { tedwu, hwanght, sypdbhee, jeremycchsu, whm }@iis.sinica.edu.tw
[†]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
E-mail: yu.tsao@citi.sinica.edu.tw
Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan[††]
E-mail: yhlai@ee.yzu.edu.tw

ABSTRACT

This paper presents a novel postfiltering approach based on the locally linear embedding (LLE) algorithm for speech enchantment (SE). The aim of the proposed LLE-based postfiltering approach is to further remove the residual noise components from the SE-processed speech signals through a spectral conversion process, thereby increasing the signal-to-noise ratio (SNR) and speech quality. The proposed postfiltering approach consists of two phases. In the offline phase, paired SE-processed and clean speech exemplars are prepared for dictionary construction. In the online phase, the LLE algorithm is adopted to convert the SE-processed speech signals to the clean ones. The present study integrates the LLE-based postfiltering approach with a deep denoising autoencoder (DDAE) SE method, which has been confirmed to provide outstanding capability for noise reduction. Experimental results show that the proposed postfiltering approach can notably enhance the DDAE-based SE processed speech signals in different noise types and SNR levels.

Index Terms— Speech enhancement, deep neural network, locally linear embedding, postfiltering

1. INTRODUCTION

For a wide range of voice-based applications, such as hearing aids, hands-free communication and automatic speech recognition, speech enchantment (SE) plays a crucial role with the aim of improving the speech quality and intelligibility of corrupted speech. In the past, numerous SE approaches have been proposed. These approaches can be roughly divided into *unsupervised ones*, such as spectral subtraction [1], Wiener filter [2], Kalman filtering [3], and minimum mean-square-error (MMSE) spectral estimator [4], and *supervised ones*, such as sparse coding [5], nonnegative matrix factorization (NMF) [6], [7], deep neural network (DNN) [8], [9], and deep denoising auto-encoder (DDAE) [10], [11], [12]. Because a non-linear and complex function is adopted to characterize the mapping from noisy to clean speech, when a sufficient amount of training data is available, DNN- and DDAE-based approaches can yield outstanding performance [13].

Recently, we have employed the maximum likelihood parameter generation algorithm (MLPG) [14], [15] in a DDAE-based SE system, termed DAS, to overcome the discontinuity effect caused by frame-by-frame processing [12]. Experimental results confirm that DAS can provide higher quality and intelligibility than DDAE alone. In this paper, we further adopt a postfiltering stage based on the local linear embedding (LLE) algorithm [16] to improve the DAS performance.

LLE is a manifold learning algorithm, which has been successfully applied to speaker voice conversion in our previous work [17]. In this study, we investigate its ability in SE. We first employed LLE to directly convert noisy speech to clean speech. Due to its natural limitation, however, LLE could not achieve satisfactory performance when working alone, especially under low signal-to-noise ratio (SNR) noisy conditions. Nevertheless, we noted that LLE-based postfiltering could be suitably combined with DAS to further remove the residual noise components, and thus improve the SNR and speech quality in different noisy conditions.

The paper is organized as follows. The proposed LLE-based postfiltering approach for SE is introduced in Section II. The experimental evaluations are presented in Section III. Finally, Section IV gives the conclusions.

2. LLE-BASED POSFILTERING APPROACH

Figure 1 shows the system architecture of the proposed LLE-based postfiltering approach for SE. The main concept is to perform voice conversion (from DAS-processed speech to clean speech) based on the LLE algorithm. The overall process can be divided into offline and online stages, which will be detailed in this section.

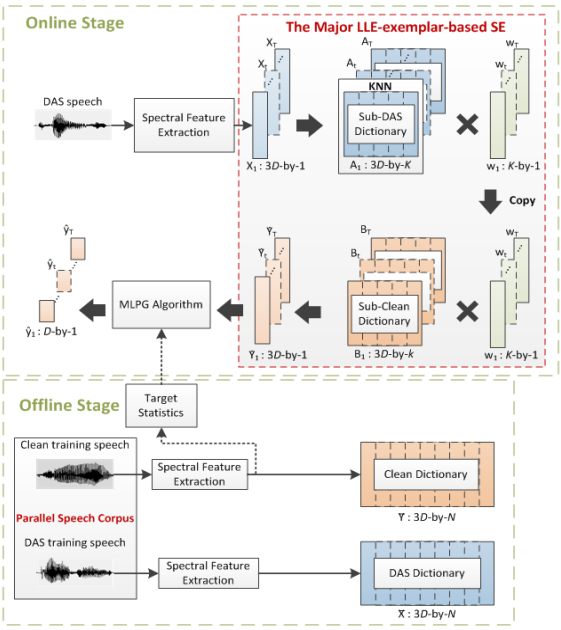

Fig. 1. The proposed LLE-based postfilterning approach.

2.1. The offline stage

From Fig. 1, the paired DAS-processed speech signals and clean ones are prepared in the offline stage. After spectral feature extraction, a pair of dictionaries (DAS and clean dictionaries) is constructed from the joint spectral feature vectors. In the meanwhile, clean speech statistics are estimated, which will be used in the MLPG algorithm.

Let the DAS and clean dictionaries be composed by the source and target spectral feature vectors (called exemplars hereafter) as $\bar{\mathbf{X}}=\left[\bar{\mathbf{X}}_1,\cdots,\bar{\mathbf{X}}_n,\cdots,\bar{\mathbf{X}}_N\right]$ and $\bar{\mathbf{Y}}=\left[\bar{\mathbf{Y}}_1,\cdots,\bar{\mathbf{Y}}_n,\cdots,\bar{\mathbf{Y}}_N\right]$, respectively. $\bar{\mathbf{X}}_n$ and $\bar{\mathbf{Y}}_n$ are the source and target exemplars at frame *n*, respectively. For both source and target signals, the total number of exemplars is *N*. Note that each exemplar (spectral feature vector) in dictionaries $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ is composed by its *D*-dimensional static, delta, and delta-delta features, in order to consider the temporal information.

2.2. The online stage

In the online stage, given an input utterance (DAS-processed speech), spectral feature extraction is performed to obtain the spectral feature vectors $\mathbf{X}$ (source spectral feature vectors). Then, the major LLE-exemplar-based SE module is performed to convert the source spectral features to the converted spectral features, such that the locality structure in the source spectral features is preserved in the converted spectral features. In the following subsections, we will describe the major LLE-exemplar-based SE part and the MLPG algorithm of the proposed postfiltering approach.

2.2.1. The major LLE-exemplar-based SE

The major LLE-exemplar-based SE consists of three steps. The first step identifies the locally linear patch by finding a set of *K* nearest neighbors (measured by the Euclidean distance) from the DAS dictionary for each data point (source spectral feature vector). The second step characterizes the local geometry of each locally linear patch by computing the reconstruction weights that minimize the local reconstruction error as

$$\varepsilon=\sum_{t=1}^{T}\left\|\mathbf{X}_t-\mathbf{A}_t\mathbf{w}_t\right\|^2=\sum_{t=1}^{T}\left\|\mathbf{X}_t-\sum_{k=1}^{K}\mathbf{w}_t(k)\mathbf{a}_{tk}\right\|^2, \qquad (1)$$

where $\mathbf{X}_t$ (a *3D*-by-1 vector) denotes the source spectral feature vector (composed by its static, delta and delta-delta features) at frame *t*; *T* is the total number of frames (source spectral feature vectors) of an input utterance for conversion; $\mathbf{A}_t=\left[\mathbf{a}_{t1},\cdots,\mathbf{a}_{tk},\cdots\mathbf{a}_{tK}\right]$ (a 3*D*-by-*K* matrix referred to as the sub-DAS dictionary) is the subset of the DAS dictionary $\bar{\mathbf{X}}$ for $\mathbf{X}_t$; $\mathbf{a}_{tk}$ (a *3D*-by-1 vector) is the *k*-th exemplar (i.e., the *k*-th nearest neighbor of $\mathbf{X}_t$) in the sub-DAS dictionary; and $\mathbf{w}_t$ (a *K*-by-1 vector) is the reconstruction weight vector at frame *t*, subject to $\mathbf{1}^{\mathrm{T}}\mathbf{w}_t=1$, where $\mathbf{1}$ is a *K*-by-1 vector whose elements are all ones, for the purpose of translational invariance. Estimating the reconstruction weights by minimizing ε subject to the constraint is a constrained least squares problem and can be solved separately for each frame. The solution can be obtained by solving the linear system of equations $\mathbf{G}_t\mathbf{w}_t=\mathbf{1}$, and then rescale the weights to satisfy the constraint $\mathbf{1}^{\mathrm{T}}\mathbf{w}_t=1$, where $\mathbf{G}_t$ is the local Gram matrix (*K*-by-*K*) for $\mathbf{X}_t$:

$$\mathbf{G}_t=\left(\mathbf{A}_t-\mathbf{X}_t\mathbf{1}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\mathbf{A}_t-\mathbf{X}_t\mathbf{1}^{\mathrm{T}}\right). \qquad (2)$$

Finally, in the third step, with the assumption that the spectral feature vectors of the SE-processed speech and those of the clean speech form manifolds with similar local geometries in two distinct spectral feature spaces, the converted spectral feature vectors $\hat{\mathbf{Y}}$ is obtained by using the reconstruction weights and the corresponding *K* target exemplars as

$$\hat{\mathbf{Y}}_t=\mathbf{B}_t\mathbf{w}_t=\sum_{k=1}^{K}\mathbf{w}_t(k)\mathbf{b}_{tk}, \qquad (3)$$

where $\mathbf{B}_t=\left[\mathbf{b}_{t1},\cdots,\mathbf{b}_{tk},\cdots,\mathbf{b}_{tK}\right]$ (a 3*D*-by-*K* matrix referred to as the sub-clean dictionary) is the subset of the clean dictionary $\bar{\mathbf{Y}}$ corresponding to the sub-DAS dictionary $\mathbf{A}_t$, in which each $\mathbf{b}_{tk}$ (a *3D*-by-1 vector) is the *k*-th exemplar (corresponding to $\mathbf{a}_{tk}$) in the sub-clean dictionary.

2.2.2. The MLPG algorithm

It has been noted that LLE-exemplar-based voice conversion still suffers from the discontinuity [17]. To overcome this problem, the MLPG algorithm for the proposed method is given as

$$\hat{\mathbf{y}} = (\mathbf{M}^{\mathrm{T}}\mathbf{U}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{U}\hat{\mathbf{Y}}, \quad (4)$$

where $\hat{\mathbf{y}}$ (a DT-by-1 vector) is the converted static spectral feature sequence; $\mathbf{M}$ is a 3DT-by-DT weighting matrix used for appending the dynamic features to the static ones; $\hat{\mathbf{Y}} = [\hat{\mathbf{Y}}_1^{\mathrm{T}}, \cdots, \hat{\mathbf{Y}}_t^{\mathrm{T}}, \cdots, \hat{\mathbf{Y}}_T^{\mathrm{T}}]^{\mathrm{T}}$ (a 3DT-by-1 vector) is the converted spectral feature sequence obtained by (3); $\mathbf{U} = \mathrm{daig}[\mathbf{\Lambda}_1^{(\mathbf{Y})}, \cdots, \mathbf{\Lambda}_t^{(\mathbf{Y})}, \cdots, \mathbf{\Lambda}_T^{(\mathbf{Y})}]$ (a 3DT-by-3DT matrix) is the global precision matrix, where $\mathbf{\Lambda}_1^{(\mathbf{Y})} = \cdots = \mathbf{\Lambda}_t^{(\mathbf{Y})} = \cdots = \mathbf{\Lambda}_T^{(\mathbf{Y})}$ (a 3D-by-3D matrix) is the precision/variance estimated from the clean speech data, which is assumed to be diagonal.

3. EXPERIMENTS

3.1. Experimental setting

Our experiments were conducted on a Mandarin hearing in noise test (MHINT) database, which contained 300 utterances pronounced by a male native Mandarin speaker recorded in a clean condition room with a 16 kHz sampling rate. We compared the baseline DAS system [12] (denoted as DAS) with the DAS system with the proposed LLE-based postfiltering approach (denoted as DAS w/ LLE).

3.1.1. The baseline DAS system

The first 250 utterances of the MHINT dataset were used for training the DAS system. The training utterances were artificially added by car and two-talker noises recorded in a real environment. The SNRs ranged from -10 to 20 dB with a 5 dB interval. As a result, for each noise type, 1750 noisy utterances paired with the corresponding clean utterances were generated as the training set. The neural networks of the DAS system consisted of seven hidden layers with 1200, 300, 300, 514, 300, 300, and 1200 hidden nodes. Two DAS systems, one for the car noise and the other for the two-talker noise, were obtained by the training data.

For signal analysis, the frame length and the frame shift for segmenting a speech waveform with a hamming window were 32 and 16 milliseconds, respectively. Each frame of speech was converted to a static feature vector with 257-dimensional log-power spectral features. The contextual feature vectors were then appended to the static one to form the final spectral feature vector, whose dimension was 771.

3.1.2. The proposed LLE-based postfiltering approach

Five-fold cross validation was performed to evaluate the proposed LLE-based postfiltering approach. In each run, among 50 utterances in the test set, we constructed the DAS and clean dictionaries using 40 utterances while the remaining 10 utterances were used for test. The SNRs for building the dictionaries were -10, 0, and 10 dB. Therefore, for each noise type, 120 clean and the corresponding DAS-processed utterances were used for building the dictionaries. The signal analysis part is the same as that used in developing the DAS system, except that the log-power spectral features among all frames were normalized to the same energy. Moreover, the number of nearest neighbors, namely K in (1), for the LLE algorithm was set to 1024 empirically.

Table 1. PESQ, STOI, and SSNR of DAS w/ LLE and baseline DAS on the test set at different SNRs of the ***two-talker*** noise.

	DAS w/ LLE			DAS		
	PESQ	STOI	SSNR	PESQ	STOI	SSNR
SNR10	**2.22**	0.83	**12.73**	2.21	**0.88**	12.48
SNR6	**2.11**	0.82	**12.08**	2.05	**0.86**	11.76
SNR2	**1.97**	0.80	**10.88**	1.93	**0.84**	10.47
SNR0	**1.86**	0.79	**10.12**	1.83	**0.83**	9.66
SNR-2	**1.78**	0.78	**9.03**	1.75	**0.81**	8.46
SNR-6	1.59	0.75	**6.13**	**1.61**	**0.78**	5.38
SNR-10	1.42	0.69	**2.53**	**1.47**	**0.72**	1.51
Ave	**1.85**	0.78	**9.07**	1.83	**0.82**	8.53

Table 2. PESQ, STOI, and SSNR of DAS w/ LLE and baseline DAS on the test set at different SNRs of the ***car*** noise.

	DAS w/ LLE			DAS		
	PESQ	STOI	SSNR	PESQ	STOI	SSNR
SNR10	**2.03**	0.80	**15.73**	1.96	**0.85**	15.04
SNR6	**1.99**	0.79	**14.91**	1.93	**0.84**	14.17
SNR2	**1.92**	0.78	**13.37**	1.89	**0.83**	12.40
SNR0	**1.86**	0.78	**12.34**	1.85	**0.82**	11.40
SNR-2	**1.82**	0.77	**11.05**	1.81	**0.81**	10.00
SNR-6	1.71	0.75	**7.74**	**1.75**	**0.79**	6.34
SNR-10	1.60	0.72	**3.90**	**1.67**	**0.76**	2.22
Ave	**1.85**	0.77	**11.29**	1.84	**0.81**	10.23

3.2. Objective evaluation

We compared DAS and DAS w/ LLE in terms of three objective evaluation metrics, namely, the perceptual evaluation of speech quality (PESQ) [18], the short-time objective intelligibility measure (STOI) [19], and the segmental signal-to-noise ratio (SSNR, in dB) [20]. The score ranges of PESQ and STOI are {-0.5 to 4.5} and {0 to 1}, respectively. Higher scores of PESQ and STOI denote better speech quality and better intelligibility, respectively. On the other hand, SSNR denotes the degree of noise reduction.

Tables 1 and 2 show the objective evaluation scores obtained by DAS and DAS w/ LLE in the two-talker and car noises at different SNRs, respectively. From Table 1, we first observe that DAS w/ LLE achieves better SSNR scores than DAS at all SNRs. Similar trends can also be found in the car noise condition as shown in Table 2. The result reveals that the residual noises in the DAS-processed speech can be further removed by the LLE-based postfiltering approach, thereby increasing the SSNR scores. We also observe that DAS w/ LLE obtains slightly higher PESQ scores than DAS in higher SNR conditions in both Table 1 and Table 2. The result suggests that the proposed postfiltering approach can slightly improve the sound quality. Finally, it is found that DAS w/ LLE is inferior to DAS under all SNRs and noise types in terms of STOI. The result suggests that the proposed postfiltering approach tends to degrade the

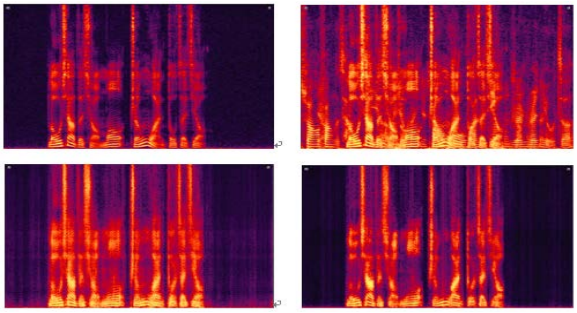

Fig. 2. Spectrograms of an utterance example, original (upper left), noisy speech (upper right), DAS enhanced (bottom left), and DAS w/ LLE (bottom right) with two-talker noise at SNR = 6 dB.

speech intelligibility slightly. Fig. 2 shows the spectrograms of clean speech, noisy speech, speech enhanced by DAS, and speech enhanced by DAS w/ LLE, respectively. From the figure, we observe that both DAS and DAS w/ LLE can remove noise components. We also observe that DAS w/ LLE reveals more detailed sound structures and removes more noise components than DAS.

3.3. Subjective evaluation

We performed subjective tests in terms of noise reduction capability and preference, respectively. In the noise reduction capability test, the subjects were asked to select one from two utterances that had a better noise reduction capability. In the preference test, the subjects were asked to select one from two utterances according to the overall preference. That is, the subjects were hinted to select one from two utterances considering the speech quality, speech intelligibility, and noise reduction capability jointly.

The test utterances were generated under two noise types (i.e., two-talker and car noises) at three different SNRs (i.e., -6, 0, and 6 dB). Note that -6dB and 6dB were not seen in both DAS training and LLE dictionary construction. Fifteen pairs of utterances were tested for each noise type and SNR. We conducted AB tests, i.e., each pair of SE-processed speech utterances by methods **A** and **B** were presented in a random order to the subjects. Twelve subjects were involved in the tests. Figs. 3 and 4 show the results of the noise reduction capability test and the preference test, respectively.

From Fig. 3, we observe that DAS w/ LLE outperforms baseline DAS in all experimental conditions. The result confirms that the residual noises in the DAS-processed speech signals can be further remove by the LLE-based postfiltering approach through a spectral conversion process. The result is consistent with that of the SSNR-based objective evaluation shown in Tables 1 and 2. From Fig. 4, we also observe that DAS w/ LLE achieves a significant gain over DAS in all experimental conditions. The result again demonstrates the effectiveness of the proposed LLE-based postfiltering approach for SE. It is worth mentioning that the main factor

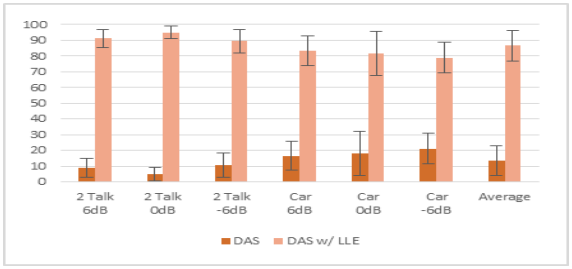

Fig. 3. Noise reduction capability test results for two noise types (2Talk: two-talker noise, Car: car noise) at three different SNRs (-6, 0, 6 dB), respectively. Error bars indicate the 95% confidence intervals.

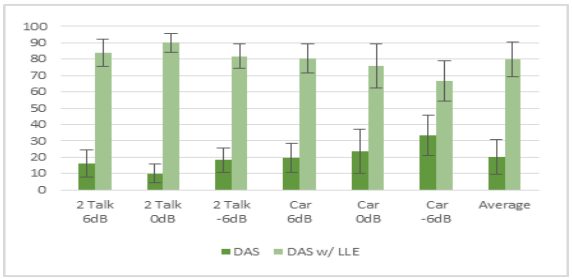

Fig. 4. Preference test results for two noise types (2Talk: two-talker noise, Car: car noise) at three different SNRs (-6, 0, 6 dB), respectively. Error bars indicate the 95% confidence intervals.

considered in the preference test is the noise reduction capability according to the subjects' responses. A possible reason is that the speech quality and speech intelligibility of both approaches are similar to each other (cf. the scores of PESQ and STOI in Tables 1 and 2); therefore, noise reduction capability becomes an important factor while comparing the proposed and baseline approaches.

4. CONCULSIONS

In this paper, we have proposed a novel LLE-based postfiltering approach for SE. Our main contribution is that we investigate the use of the LLE algorithm with the paired SE-processed and clean dictionaries for postfiltering for the SE task. The subjective evaluation results revealed that the DAS system with the proposed LLE-based postfiltering approach (DAS w/ LLE) achieves a significant gain over the baseline DAS system. DAS w/ LLE also shows its potential in the objective evaluation. For future work, we will evaluate our LLE-based postfiltering approach on more SE approaches and noise types. Extending the current approach to a speaker independent one is attractive but challenging.

5. ACKNOWLEDGEMENT

This work was supported in part by the Ministry of Science and Technology of Taiwan under Grant: MOST 105-2221-E-001-012-MY3.

6. REFERENCES

[1] 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.

[2] P. Scalart and J. V. Filho, “Speech enhancement based on a priori signal to noise estimation,” in *Proc. ICASSP*, pp. 629-632, 1996.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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

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

[11] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Ensemble modeling of denoising autoencoder for speech spectrum restoration,” in *Proc. INTERSPEECH*, pp. 885–889, 2014.

[12] S.-S. Wang, H.-T. Hwang, Y.-H. Lai, Y. Tsao, X. Lu, H.-M. Wang, and B. Su, "Improving denoising auto-encoder based speech enhancement with the speech parameter generation algorithm," in *Proc*. *APSIPA ASC*, pp. 365-369, 2015.

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

[14] K. Tokuda, T. Kobayashi, and S. Imai, “Speech parameter generation from hmm using dynamic features,” in *Proc. ICASSP*, pp. 660–663, 1995.

[15] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *Proc. ICASSP*, pp. 1315-1318, 2000.

[16] S. T. Roweis and L.K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323-2326, 2000.

[17] Y.-C. Wu, H.-T. Hwang, C.-C. Hsu, Y. Tsao, and H.-M. Wang, “Locally linear embedding for exemplar-based spectral conversion,” in *Proc. INTERSPEECH*, pp. 1652-1656, 2016.

[18] ITU-T, Rec. P.862, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs International Telecommunication Union-Telecommunication Standardisation Sector, 2001.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[20] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.