# A Post-filtering Approach Based on Locally Linear Embedding Difference Compensation for Speech Enhancement

*Yi-Chiao Wu*[1], *Hsin-Te Hwang*[1], *Syu-Siang Wang*[2], *Chin-Cheng Hsu*[1], *Yu Tsao*[2], *Hsin-Min Wang*[1]

[1] Institute of Information Science, Academia Sinica, Taipei, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

yu.tsao@citi.sinica.edu.tw, {tedwu, hwanght, whm}@iis.sinica.edu.tw

## Abstract

This paper presents a novel difference compensation post-filtering approach based on the locally linear embedding (LLE) algorithm for speech enhancement (SE). The main goal of the proposed post-filtering approach is to further suppress residual noises in SE-processed signals to attain improved speech quality and intelligibility. The proposed system can be divided into offline and online stages. In the offline stage, we prepare paired differences: the estimated difference of {SE-processed speech; noisy speech} and the ground-truth difference of {clean speech; noisy speech}. In the online stage, on the basis of estimated difference of a test utterance, we first predict the corresponding ground-truth difference based on the LLE algorithm, and then compensate the noisy speech with the predicted difference. In this study, we integrate a deep denoising autoencoder (DDAE) SE method with the proposed LLE-based difference compensation post-filtering approach. The experiment results reveal that the proposed post-filtering approach obviously enhanced the speech quality and intelligibility of the DDAE-based SE-processed speech in different noise types and signal-to-noise-ratio levels.

**Index Terms**: Speech enhancement, post-filtering, difference compensation, locally linear embedding.

## 1. Introduction

Speech enhancement (SE) has been used as a fundamental unit in a wide range of voice-based applications, such as assistive hearing devices [1, 2, 3], hands-free communication [4], and automatic speech recognition [5]. Traditionally, SE algorithms were derived based on the statistical characteristics of speech and noise signals. Well-known examples include spectral subtraction [6], Wiener filter [7], Kalman filtering [8], and minimum mean-square-error (MMSE) spectral estimation [9]. More recently, machine-learning based approaches, such as sparse coding [10], nonnegative matrix factorization (NMF) [11, 12], deep neural network (DNN) [13, 14], deep denoising auto-encoder (DDAE) [15, 16, 17], recurrent neural network [18], and convolutional neural network (CNN) [19], have attracted great attention. Although these previously developed SE algorithms already yield good performances in many conditions, two issues are still not perfectly addressed, i.e., residual noise and speech distortions are noticeable in enhanced speech signals. To address these two issues, this study proposes a novel difference compensation post-filtering approach based on the locally linear embedding (LLE) algorithm.

The LLE algorithm [20] is a manifold learning algorithm that characterizes the intrinsic geometric structure of high dimensional data. In our previous work [21], we found that LLE could be successfully applied to speaker voice conversion to model the geometric structure of source speech, and then embed the structure to the target speech space. Meanwhile, we derived an LLE-based post-filtering approach to further remove the residual noise components by directly converting the SE-processed speech signals to clean ones [22]. Experimental results showed that the LLE-based post-filtering approach could notably improve the SE-processed speech signals in terms of perceptual evaluation of speech quality (PESQ) [23] and segmental signal-to-noise ratio (SSNR) [24] in different noise types and signal-to-noise-ratio (SNR) levels. However, the short-time objective intelligibility (STOI) [25] score was slightly degraded, suggesting that the processed speech might still suffer from the distortion issue.

On the basis of the success and to overcome the distortion issue of the previous approach in [22], we derive a new post-filtering approach based on LLE-based difference compensation. Because the approach in [22] directly converted SE-processed speech to clean speech and did not utilize noisy speech information, the performance of the post-filtering approach depended heavily on the capability of the preceding SE process. In the present study, however, we propose a two-stage approach: First, we convert the difference of {SE-processed speech; noisy speech} to the difference of {clean speech; noisy speech}. To be more specifically, the difference {SE-processed speech; noisy speech} is used to facilitate an accurate prediction of the difference of {clean speech; noisy speech}. Finally, the converted difference is used to compensate the noisy speech signals to clean ones. Experimental results show that the newly proposed post-filtering approach outperforms the direct conversion-based post-filtering approach in [22], and notably enhances the speech quality and intelligibility of the SE-processed speech in different noise types and SNR levels.

The remainder of this paper is organized as follows. Section II describes the proposed LLE-based difference compensation post-filtering framework. Section III presents the experimental results. Section IV gives the discussion, and Section V summarizes the entire work.

## 2. LLE-based difference compensation

Figure 1 illustrates the system architecture of the proposed difference compensation post-filtering approach. The main concept is to reduce the uncertainty between paired enhanced-clean dictionaries. We estimate the difference of the clean-noisy pair from that of the enhanced-noisy pair, rather than directly estimating the clean speech from the enhanced speech. Then, we add the clean-noisy difference to the noisy speech to get the estimated clean speech. The overall system includes offline and online stages.

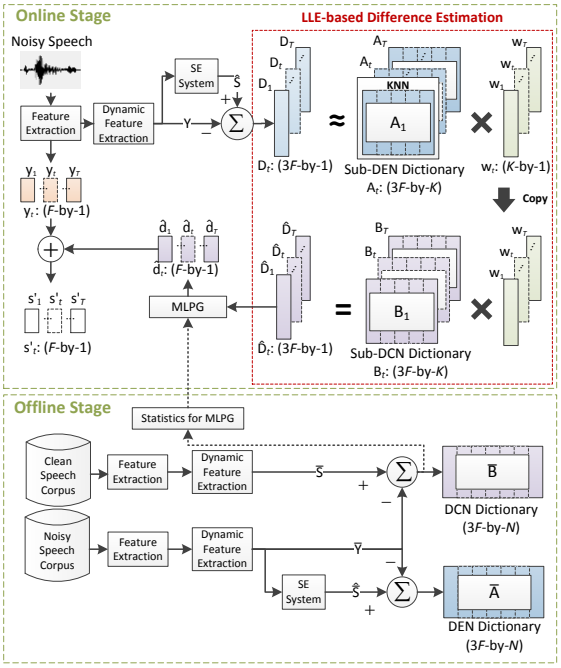


Figure 1: *LLE-based difference compensation system.*

### 2.1. Offline stage

For the offline stage in Figure 1, we construct the paired difference dictionaries: DCN and DEN, where the DCN dictionary is constructed by the differences of {clean speech features; noisy speech features}, and the DEN dictionary is constructed by the differences of {enhanced speech features; noisy speech features}. In addition to the two dictionaries, we also calculate the statistics of the DCN-features to be used by maximum likelihood parameter generation (MLPG) [26, 27] in the online stage.

To construct the dictionaries, we prepare clean speech, $\overline{\mathbf{S}}=\left[\overline{\mathbf{S}}_1,\cdots,\overline{\mathbf{S}}_n,\cdots,\overline{\mathbf{S}}_N\right]$, the corresponding noisy speech, $\overline{\mathbf{Y}}=\left[\overline{\mathbf{Y}}_1,\cdots,\overline{\mathbf{Y}}_n,\cdots,\overline{\mathbf{Y}}_N\right]$, and the corresponding enhanced speech, $\overline{\mathbf{S}}=\left[\overline{\mathbf{S}}_1,\cdots,\overline{\mathbf{S}}_n,\cdots,\overline{\mathbf{S}}_N\right]$, where $\overline{\mathbf{S}}_n$, $\overline{\mathbf{Y}}_n$, and $\overline{\mathbf{S}}_n$ denote the *n*-th clean, noisy, and enhanced speech feature vectors, respectively; the dimension of each feature vector is 3*F*, including static, delta, and delta-delta terms; and *N* denotes the total number of speech frames. Based on the prepared data, we conduct frame-based subtraction between $\overline{\mathbf{S}}$ and $\overline{\mathbf{Y}}$ to obtain the DEN features to form the DEN dictionary $\overline{\mathbf{A}}$ and frame-based subtraction between $\overline{\mathbf{S}}$ and $\overline{\mathbf{Y}}$ to obtain the DCN features to form the DCN dictionary $\overline{\mathbf{B}}$.

### 2.2. Online stage

In the online stage, the input noisy speech is first converted into spectral features, $\mathbf{y}\in\mathcal{R}^{F\times T}$, which then form $\mathbf{Y}\in\mathcal{R}^{3F\times T}$ by appending dynamic features. *T* is the total frame number in the input utterance. Next, SE is applied to $\mathbf{Y}$ to obtain the enhanced spectra features $\mathbf{S}\in\mathcal{R}^{3F\times N}$. We also apply voice activity detection (VAD) to the enhanced speech to determine the time slots of noise and speech, which are then used to predict the SNR level of the given noisy speech. With the predicted SNR, we adjust the volume of the input noisy utterance to make the energy of the clean speech component of the noisy speech match the energy of the enhanced speech. Next, we obtain the input DEN features $\mathbf{D}_t$ (difference of $\mathbf{Y}_t$ and $\mathbf{S}_t$), where *t* is the frame index. Then, the LLE-based difference estimation is performed to convert the input DEN features to the DCN features. Finally, we compensate the noisy speech with the DCN features to obtain the final enhanced speech. In the following subsections, we will detail the LLE-based difference estimation and difference compensation steps of the proposed post-filtering approach.

#### *2.2.1. LLE-based difference estimation*

LLE-based difference estimation consists of three steps, and each step is carried out in a frame-by-frame manner. The first step is to find the local patches in the DEN manifold space for the input data; the second step is to reconstruct the input data by a linear combination of the local patches; the third step is to embed the linear combination weight to the corresponding patches in the DCN manifold space to generate the output data. The main concept of the LLE algorithm is that only the nearest neighbors in the same local patches have a linear geometrical relationship. Thus, the first step only finds a set of *K* nearest neighbors (measured by the Euclidean distance) from the DEN dictionary for each input data to form a sub-DEN dictionary.

In the second step, the cost function of minimizing the linear reconstruction error is defined as

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

where $\mathbf{A}_t=\left[\mathbf{a}_{t1},\cdots,\mathbf{a}_{tk},\cdots,\mathbf{a}_{tK}\right]$ (a 3*F*-by-*K* matrix referred to as the sub-DEN dictionary) is the subset of the DEN dictionary $\overline{\mathbf{A}}$ for $\mathbf{D}_t$, $\mathbf{a}_{tk}\in\mathcal{R}^{3F\times 1}$ is the *k*-th nearest neighbor of $\mathbf{D}_t$ in $\overline{\mathbf{A}}$, and $\mathbf{w}_t\in\mathcal{R}^{K\times 1}$ is the reconstruction weight vector at frame *t*. Estimating the reconstruction weights can be done by solving the linear system of equations $\mathbf{G}_t\mathbf{w}_t=\mathbf{l}$, and then rescaling the weights to satisfy the constraint $\mathbf{l}^{\mathrm{T}}\mathbf{w}_t=1$, where $\mathbf{G}_t$ is the local Gram matrix (*K*-by-*K*) for $\mathbf{D}_t$:

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

Note that, to make each data point and its neighbors invariant to the rotation, rescaling, and translation operations, the LLE algorithm must satisfy the constraint of $\mathbf{l}^{\mathrm{T}}\mathbf{w}_t=1$, where $\mathbf{l}$ is a *K*-by-1 all-ones vector.

In the third step, we assume that the geometrical structures of the input data and the local patch in the DEN manifold space are similar to the DCN counterparts. As a result, the predicted DCN vector $\mathbf{D}$ can be obtained by using the reconstruction weights and the corresponding *K* DCN neighbors as

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

where $\mathbf{B}_t=\left[\mathbf{b}_{t1},\cdots,\mathbf{b}_{tk},\cdots,\mathbf{b}_{tK}\right]$ (a 3*F*-by-*K* matrix referred to as the sub-DCN dictionary) is the subset of the DCN dictionary $\overline{\mathbf{B}}$

corresponding to the sub-DEN dictionary $\mathbf{A}_t$, in which each $\mathbf{b}_{tk}$ (a *3F*-by-1 vector) is the *k*-th exemplar (corresponding to $\mathbf{a}_{tk}$) in the sub-DCN dictionary.

*2.2.2. Difference compensation*

After obtaining the predicted DCN features $\mathbf{D}$, we then apply the MLPG algorithm to compute the smoothed static DCN features $\mathbf{d} \in \mathcal{R}^{F \times T}$. Finally, we can obtain the final predicted clean spectrum $\mathbf{s}' \in \mathcal{R}^{F \times T}$ by

$$\mathbf{s}' = \mathbf{d} + \mathbf{y} . \quad (4)$$

### 2.3. Comparison with the direct conversion method

In this section, we compare the proposed LLE-based difference compensation post-filtering approach (denoted as LDC) with the previous direct LLE-based post-filtering approach (denoted as DL). The main difference between LDC and DL in the offline stage is the construction of the dictionaries. Specifically, since enhanced speech under different noise types, SNR levels and distortions link to the same ground-truth clean speech, the many-to-one issue may occur in DL. Nevertheless, after we introduce the noisy speech information to get the DEN and DCN features, the DEN-DCN paired exemplars become a one-to-one case. Therefore, the paired DEN and DCN dictionaries in LDC can reduce the uncertainty of the paired enhanced and clean dictionaries in DL. In the online stage, DL directly predicts the final enhanced speech while LDC predicts the difference between clean and noisy speech and compensates the input noisy speech with the predicted difference to generate the final enhanced speech. As a result, LDC-processed speech may retain more speech details from noisy speech.

## 3. Experiments

### 3.1. Experimental setting

We used the DDAE approach [17] as the SE method. We compared the performance of DDAE alone (denoted as DDAE) versus DDAE with two post-filtering approaches, directly LLE-based post-filtering (denoted as DL), and the proposed LLE-based difference compensation post-filtering (denoted as LDC). We evaluated the systems using both objective evaluation and subjective listening test evaluation on the Mandarin hearing in noise test (MHINT) sentences [28], which contained 300 utterances, pronounced by a male native Mandarin speaker and recorded in a clean condition room with a 16-kHz sampling rate. The first 250 utterances of the MHINT dataset were used for training the DDAE model, and the remaining 50 utterances were used for testing. The noisy speech data were obtained by artificially adding noises (car and two-talker noises recorded in a real environment) to the clean speech utterances.

We followed our previous work in [22] to configure the signal analysis settings and system architectures of DDAE and DL. For a fair comparison, the LDC adopted the same setup as DL (log power spectrum, energy normalization, etc.). During evaluation, five-fold cross validation was performed, and the number of nearest neighbors, namely *K* in (1), for both DL and LDC was set to 1024 empirically.

### 3.2. Objective evaluation

Tables 1, 2, and 3 show the PESQ, STOI, and SSNR (in dB) scores of DDAE, DL, and LDC. The PESQ scores indicate speech quality, and the score range is {-0.5 to 4.5}. The STOI scores indicate speech intelligibility, and the score range is {0 to 1}. The SSNR denotes the degree of noise reduction. For all three objective evaluation scores, a higher score means the better performance of the enhanced speech.

Table 1: *PESQ of DDAE, DL, and LDC on the test set at different SNRs of the **car** and **two-talker** noise.*

| PESQ | | | | | | |
|---|---|---|---|---|---|---|
| Noise | Two-talker | | | Car | | |
| Method | DDAE | DL | LDC | DDAE | DL | LDC |
| SNR10 | 2.21 | 2.22 | **2.74** | 1.96 | 2.03 | **3.10** |
| SNR6 | 2.05 | 2.11 | **2.44** | 1.93 | 1.99 | **2.88** |
| SNR2 | 1.93 | 1.97 | **2.22** | 1.89 | 1.92 | **2.59** |
| SNR0 | 1.83 | 1.86 | **2.08** | 1.85 | 1.86 | **2.43** |
| SNR-2 | 1.75 | 1.78 | **1.95** | 1.81 | 1.82 | **2.28** |
| SNR-6 | 1.61 | 1.59 | **1.74** | 1.75 | 1.71 | **2.02** |
| SNR-10 | 1.47 | 1.42 | **1.56** | 1.67 | 1.60 | **1.82** |
| Ave | 1.83 | 1.85 | **2.10** | 1.84 | 1.85 | **2.44** |

Table 2: *STOI of DDAE, DL, and LDC on the test set at different SNRs of the **car** and **two-talker** noise.*

| STOI | | | | | | |
|---|---|---|---|---|---|---|
| Noise | Two-talker | | | Car | | |
| Method | DDAE | DL | LDC | DDAE | DL | LDC |
| SNR10 | 0.88 | 0.83 | **0.90** | 0.85 | 0.80 | **0.90** |
| SNR6 | 0.86 | 0.82 | **0.88** | 0.84 | 0.79 | **0.88** |
| SNR2 | 0.84 | 0.80 | **0.86** | 0.83 | 0.78 | **0.86** |
| SNR0 | 0.83 | 0.79 | **0.84** | 0.82 | 0.78 | **0.85** |
| SNR-2 | 0.81 | 0.78 | **0.82** | 0.81 | 0.77 | **0.83** |
| SNR-6 | **0.78** | 0.75 | **0.78** | 0.79 | 0.75 | **0.80** |
| SNR-10 | 0.72 | 0.69 | **0.73** | **0.76** | 0.72 | 0.75 |
| Ave | 0.82 | 0.78 | **0.83** | 0.81 | 0.77 | **0.84** |

Table 3: *SSNR of DDAE, DL, and LDC on the test set at different SNRs of the **car** and **two-talker** noise.*

| SSNR | | | | | | |
|---|---|---|---|---|---|---|
| Noise | Two-talker | | | Car | | |
| Method | DDAE | DL | LDC | DDAE | DL | LDC |
| SNR10 | 12.48 | 12.73 | **13.99** | 15.04 | 15.73 | **17.59** |
| SNR6 | 11.76 | 12.08 | **13.04** | 14.17 | 14.91 | **16.59** |
| SNR2 | 10.47 | 10.88 | **11.51** | 12.40 | 13.37 | **14.63** |
| SNR0 | 9.66 | 10.12 | **10.57** | 11.40 | 12.34 | **13.40** |
| SNR-2 | 8.46 | 9.03 | **9.29** | 10.00 | 11.05 | **11.85** |
| SNR-6 | 5.38 | **6.13** | 5.39 | 6.34 | 7.74 | **7.97** |
| SNR-10 | 1.51 | **2.53** | 1.12 | 2.22 | **3.90** | 3.45 |
| Ave | 8.53 | 9.07 | **9.27** | 10.23 | 11.29 | **12.21** |

From the results in Table 1, we found that LDC outperformed DDAE across different SNR levels and noise types. Compared with DL, we noted that LDC improved the SE-processed speech under both high and low SNR conditions while DL was not able to improve the SE-processed speech under low SNR conditions, suggesting that the two-stage conversion of LDC had a better ability to avoid distortions than the direct conversion of DL. Next, from the STOI results in Table 2, we observed very similar trends as that of the PESQ results: LDC yielded higher STOI scores than DDAE and DL across different SNR levels and noise types. Finally, from the SSNR results in Table 3, we noted that LDC also achieved better SSNR scores in most conditions except very low SNR conditions. In summary, the results above reveal that, compared with DL, LDC achieves better speech intelligibility, speech quality, and SSNR scores in most conditions, confirming that by reducing the uncertainty of the paired dictionaries and maintaining speech details from the noisy speech, LDC can attain better enhancement performance.

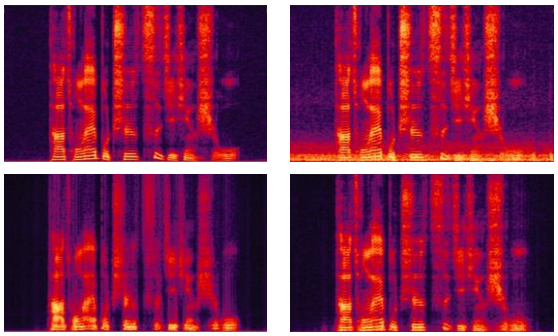

Figure 2: *Spectrograms of an utterance example, original (upper left), noisy speech (upper right), DL enhanced (bottom left), and LDC (bottom right) with* ***car*** *noise at SNR = 6 dB*

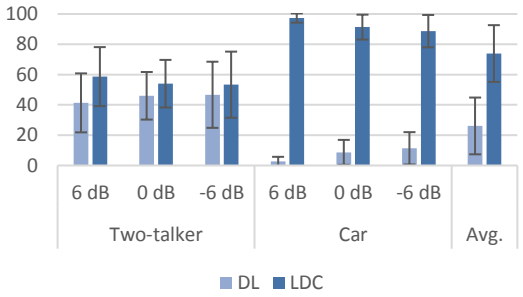


Figure 3: *Preference test results for two noise types (****car*** *and* ***two-talker*** *noise) at three different SNRs (-6, 0, 6 dB), respectively. Error bars indicate the 95% confidence intervals.*

We also compared the spectrograms of clean speech, noisy speech, DL enhanced speech, and LDC enhanced speech in Figure 2. We observed that although both DL and LDC effectively removed noise components in the spectrum domain, DL actually lost some speech details in the high-frequency bands. LDC, however, preserved most high-frequency-band speech structures, and thus the spectrogram was closer to that of the clean speech. The spectrogram analyses were actually consistent with the objective evaluation results in the previous section: LDC yields better speech quality, speech intelligibility, and noise reduction than DL.

### 3.3. Subjective evaluation

We also carried out a subjective listening test to evaluate the proposed approach. The AB test was adopted to evaluate the DL and LDC approaches. The enhanced speech signals were named A or B randomly, and the subjects were asked to pick the preferred speech from A and B according to the speech quality, listening effort, and noise reduction capability. We performed the subjective test under six conditions, including two noise types (i.e., two-talker and car noise) and three SNR levels (i.e., -6, 0 and 6 dB). Note that -6 dB and 6 dB SNR levels were not seen in either the DL or the LDC dictionary constructions. Moreover, 15 pairs of enhanced speech were tested for each condition, and 10 subjects were involved in the tests.

From Figure 3, we observed that LDC outperformed DL in all test conditions. Specifically, LDC provided remarkably better subjective performance under a more stationary noise condition (car noise), but achieved slightly better performance in the non-stationary noise case (two-talker noise). Based on the interviews with the subjects, we found that although speech intelligibility and quality were significantly improved, LDC brought unwanted noise to the enhanced speech, especially under low SNR and non-stationary noise conditions. The listening test results were consistent with the objective SSNR scores that the improvements of LDC under the car noise / high SNR case were more prominent than that under the two-talker noise / low SNR case.

## 4. Discussions

From the objective and subjective evaluations, we confirm that the proposed difference compensation post-filtering approach can improve the SE performance by introducing the noisy speech information in the difference conversion phase and utilizing the noisy speech in the compensation phase. Furthermore, because we conduct subtraction between the clean/enhanced speech and the noisy speech in the logarithmic domain, the difference actually corresponds to the clean/enhanced speech divided by the noisy speech. Therefore, the difference used in this study is related to the MASK or SNR gain. In [29], Wang et al. proposed using MASK as the DNN training target and achieved better performance than directly predicting the clean speech spectrum. Our result is in line with their conclusion. Additionally, Kinoshita et al. [30] mentioned the ill-posed problem of many-to-one mapping between the observed noisy speech and the target clean speech in DNN. In this study, we fix the many-to-one issue and the results indicate the effectiveness of our approach.

## 5. Conclusions

In this paper, we have proposed a novel LLE-based difference compensation post-filtering approach for SE. Our main contribution is that we investigated an implicit way to improve the SE-processed speech by converting the difference between the SE-processed and noisy speech to the difference between the clean and noisy speech and compensating the noisy speech with the predicted difference, rather than directly converting the SE-processed speech to the clean speech. The objective evaluation results reveal that the DDAE system with the proposed LLE-based difference compensation post-filtering approach (LDC) achieves a significant improvement in terms of speech quality and intelligibility scores, compared to the baseline DDAE system and the DDAE system with the direct LLE-based post-filtering approach (DL), across different SNR levels and noise types. The subjective evaluation results also show that LDC gives a notable gain over DL for the stationary noise type and a comparable performance under non-stationary noise conditions. In the future, we plan to investigate the connection between the difference compensation approach and the Wiener filter. Moreover, because our framework is a post-filtering approach, we will investigate the compatibility of the proposed approach with other SE techniques.

## 6. Acknowledgements

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

# 7. References


[1] H. Levitt, "Noise reduction in hearing aids: A review," Journal of rehabilitation research and development, vol. 38, p. 111, 2001.

[2] D.-L. Wang, "Deep learning reinvents the hearing aid," *IEEE Spectrum*, vol. 54, no. 3, pp. 32-37, 2017.

[3] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Transactions on Biomedical Engineering*, 2016.

[4] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication," *Speech Communication*, vol. 53, pp. 677-689, 2011.

[5] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 745-777, 2014.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

[16] 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.

[17] 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.

[18] Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., & Schuller, B. (2015, August). Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In International Conference on Latent Variable Analysis and Signal Separation (pp. 91-99). Springer International Publishing.

[19] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement," in INTERSPEECH, 2016, pp. 3768-3772.

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

[21] 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.

[22] Y.-C. Wu, H.-T. Hwang, S.-S. Wang, C.-C. Hsu, Y.-H. Lai, Y. Tsao, and H.-M. Wang, "A locally linear embedding based postfiltering approach for speech enhancement," in *Proc. ICASSP*, 2017.

[23] 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.

[24] 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.

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

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

[27] 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.

[28] L. L. Wong, S. D. Soli, S. Liu, N. Han, and M.-W. Huang, "Development of the mandarin hearing in noise test (MHINT)," Ear and hearing, vol. 28, no. 2, pp. 70S–74S, 2007.

[29] Y. Wang, N. Arun, and D.-L. Wang, "On training targets for supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[30] K. Kinoshita, M. Delcroix, A. Ogawa, T. Higuchi and T. Nakatani, "Deep mixture density network for statistical model-based feature enhancement," in *Proc. ICASSP*, 2017.