JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 34, 1469-1491 (2018)
DOI: 10.6688/JISE.201811_34(6).0007

Locally Linear Embedding Based Post-Filtering for Speech Enhancement*

HSIN-TE HWANG[1,3], YI-CHIAO WU[1], SYU-SIANG WANG[2], CHIN-CHENG HSU[1], YU TSAO[2], HSIN-MIN WANG[1], YIH-RU WANG[3] AND SIN-HORNG CHEN[3]
[1]*Institute of Information Science*
[2]*Research Center for Information Technology Innovation*
Academia Sinica
Taipei, 115 Taiwan
[3]*Department of Electrical and Computer Engineering*
National Chiao Tung University
Hsinchu, 300 Taiwan
E-mail: {hwanght; tedwu; jeremycchsu; whm}@iis.sinica.edu.tw; sypdbhee@gmail.com; yu.tsao@citi.sinica.edu.tw; {yrwang; schen}@mail.nctu.edu.tw

We present a novel speech enhancement method based on locally linear embedding (LLE). The proposed method works as a post-filter to further suppress the residual noises in the enhanced speech signals obtained by a speech enhancement system to attain improved speech quality and intelligibility. We design two types of LLE-based post-filters: the direct LLE-based post-filter (called the DL post-filter) and the LLE-based difference compensation post-filter (called the LDC post-filter). The key technique of the proposed post-filters is to apply the LLE-based feature prediction method, which integrates the LLE algorithm, a classical manifold learning method, with the exemplar-based feature prediction method, to predict either the spectral features of the clean speech from those of the enhanced speech (for DL) or the spectral difference of {clean speech; noisy speech} from that of {enhanced speech; noisy speech} (for LDC). As a result, for DL, the predicted clean speech signals can be directly reconstructed from the predicted clean spectral features. On the other hand, for LDC, the predicted clean spectral features are obtained by compensating the spectral features of the noisy speech with the predicted clean-noisy spectral difference, and then the predicted clean speech signals can be reconstructed accordingly. Experimental results demonstrate the effectiveness of the proposed post-filters for two representative speech enhancement methods, namely the deep denoising autoencoder (DDAE) and the minimum mean-square-error (MMSE) spectral estimation methods.

Keywords: speech enhancement, locally linear embedding, post-filter/postfilter, exemplar-based, manifold learning

1. INTRODUCTION

Speech enhancement has been used as a fundamental unit in a wide range of voice-based applications, such as assistive hearing devices [1, 2], hands-free communication [3, 4], automatic speech recognition [5-7], and speaker recognition [8, 9]. Traditionally, speech enhancement algorithms were derived based on the statistical characteristics of speech and noise signals. A class of approaches design a filter to suppress the noise components in the input noisy speech. Well-known examples include spectral subtraction

Received June 25, 2017; accepted July 2, 2017.
Communicated by Chung-Hsien Wu.
* This work was supported in part by the Ministry of Science and Technology of Taiwan under Grant: MOST 105-2221-E-001-012-MY3.

[10], Wiener filter [11], Kalman filtering [12], and minimum mean-square-error (MMSE) spectral estimation [13]. Another successful class is the subspace-based approaches, which split a noisy speech signal into two subspaces, one for the clean speech signal and the other for the noise comments, and then suppress the noise parts to reconstruct the clean speech signal. Notable subspace techniques include singular value decomposition (SVD) [14] and principal component analysis (PCA) [15, 16]. The class of speech model-based techniques is derived by considering both human speech production models and speech reduction functions to perform noise reduction. Successful examples include the harmonic model [17, 18], the linear prediction (LP) model [19, 20], and the hidden Markov model (HMM) [21, 22].

More recently, machine-learning based speech enhancement approaches, such as sparse coding [23], nonnegative matrix factorization (NMF) [24, 25], and artificial neural networks based approaches, such as deep neural network (DNN) [7, 26], deep denoising auto-encoder (DDAE) [27-29], recurrent neural network [30, 31], and convolutional neural network (CNN) [32], have attracted great attention. Although these previously developed speech enhancement algorithms already yield good performances in many conditions, two issues are still not perfectly addressed, *i.e.*, residual noise and speech distortions are still noticeable in enhanced speech signals. To address these two issues, we propose a novel locally linear embedding (LLE)-based post-filter for speech enhancement.

Our proposed method is inspired by the success of our previous work that integrated the LLE algorithm [33], a manifold learning algorithm that characterizes the intrinsic geometric structure of high dimensional data, with the exemplar-based spectral conversion method (called the LLE-based feature prediction method hereafter) for speaker voice conversion [34]. In this study, we employ the LLE-based feature prediction method in speech enhancement. The intuitive way is to employ LLE-based feature prediction directly to predict the spectral features of clean speech (called the clean spectral features hereafter) from the spectral features of noisy speech (called the noisy spectral features hereafter). Due to its natural limitation, however, LLE-based feature prediction could not achieve satisfactory performance in speech enhancement when working alone in our preliminary experiments, especially under low signal-to-noise ratio (SNR) noisy conditions. Therefore, we adopt it as a post-filter for speech enhancement processed speech to further remove the residual noise components.

Specifically, two types of post-filters based on the LLE-based feature prediction method are presented in this paper: the direct LLE-based (DL) post-filter [35] and the LLE-based difference compensation (LDC) post-filter. The proposed post-filters can be divided into offline and online stages. In the DL post-filter, the offline stage involves preparing the paired enhanced spectral features (obtained by a speech enhancement system) and clean spectral features (also called exemplars) for dictionary construction while the online stage involves performing the LLE-based feature prediction method to predict the clean spectral features from the enhanced spectral features. To overcome the discontinuity problem existing in the predicted clean spectral features, the maximum likelihood parameter generation algorithm [36] is applied after the LLE-based feature prediction method. On the other hand, in the LDC post-filter, the offline stage involves preparing the paired differences: the spectral difference of {enhanced speech; noisy speech} and the spectral difference of {clean speech; noisy speech}, while the online stage involves

performing the LLE-based feature prediction method (followed by the maximum likelihood parameter generation (MLPG) algorithm) to predict the clean-noisy spectral difference from the enhanced-noisy spectral difference and compensating the noisy spectral features with the predicted clean-noisy spectral difference. In this study, we evaluate the effectiveness of the proposed post-filters on a supervised speech enhancement system, *i.e.*, the DDAE-based speech enhancement system [29], and an unsupervised speech enhancement system, *i.e.*, the minimum mean-square-error (MMSE) spectral estimation-based speech enhancement system [13].

The remainder of this paper is organized as follows. The proposed LLE-based post-filters for speech enhancement are described in detail in Section 2. Experimental setup and results are presented in Section 3. Finally, Section 4 gives the conclusions.

2. THE PROPOSED LLE-BASED POST-FILTERS FOR SPEECH ENHANCEMENT

In this section, we first describe the LLE-based feature prediction method in Section 2.1, which is the core technique adopted in the proposed post-filters, and then present the direct LLE-based (DL) post-filter and the LLE-based difference compensation (LDC) post-filter in Sections 2.2 and 2.3, respectively. Finally, a comparison between the DL and LDC post-filters is given in Section 2.4.

2.1 The LLE-Based Feature Prediction Method

As mentioned previously, the LLE-based feature prediction method integrates the LLE algorithm with the exemplar-based feature prediction method. Before we start to describe the LLE-based feature prediction method, we first briefly review the LLE algorithm [33].

The LLE algorithm addresses the problem of nonlinear dimensionality reduction by computing the low-dimensional neighborhood preserving embeddings of high-dimensional data. Let each high-dimensional input data point be sampled from an underlying low-dimensional manifold, and a sufficient number of data be provided, LLE assumes that the manifold is locally linear, and each data point and its neighbors lie on or close to a locally linear patch of the manifold. A manifold can be visualized as a collection of overlapping locally linear patches if the neighborhood size is small and the manifold is sufficiently smooth. Under this condition, the local geometry of a patch (*i.e.*, the local geometry in the neighborhood of each data point) can be characterized by the reconstruction weights that reconstruct each data point from its neighbors. Then, the same reconstruction weights are used for computing the low-dimensional embedding such that the local geometry of the patch is preserved in the low-dimensional embedding space. The LLE algorithm for dimensionality reduction has three steps:

(a) Finding K nearest neighbors for each data point.
(b) Computing the reconstruction weights that best (linearly) reconstruct each data point from its K nearest neighbors found in step (a).
(c) Estimating the low-dimensional embedding for each data point by applying the reconstruction weights obtained in step (b).

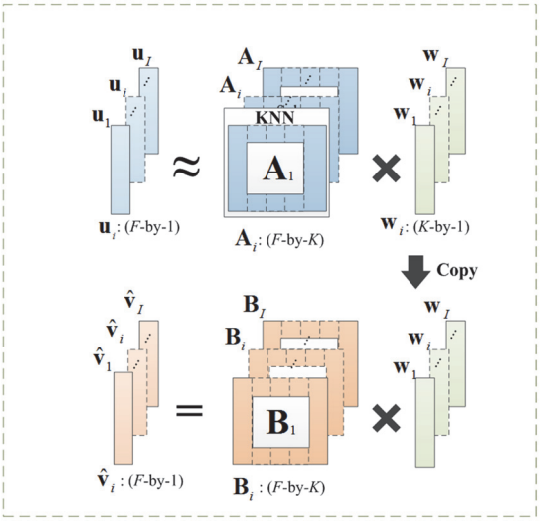


Fig. 1. Overview of the run-time prediction stage of the LLE-based feature prediction method.

The steps (a), (b), and (c) involve identifying each locally linear patch, characterizing the local geometry of each locally linear patch, and preserving the local geometry in the low-dimensional embedding space, respectively. Next, we describe the LLE-based feature prediction method.

Fig. 1 gives an overview of the run-time prediction stage of the LLE-based feature prediction method. In Fig. 1, given a sequence of input source feature vectors $\{\mathbf{u}_i \in \mathcal{R}^{F\times 1}\}_{i=1}^{I}$, the LLE-based feature prediction method is applied to predict a corresponding target feature vector from each source feature vector independently in a sample-by-sample manner. Accordingly, a sequence of predicted feature vectors $\{\hat{\mathbf{v}}_i \in \mathcal{R}^{F\times 1}\}_{i=1}^{I}$ can be obtained. $\mathbf{u}_i$ and $\hat{\mathbf{v}}_i$ are the ith source and predicted feature vectors, respectively; F is the dimensionality of features; and I is the number of input source/predicted feature vectors.

Specifically, suppose that the paired source and target dictionaries, consisting of the source and target feature vectors (also called exemplars) respectively, have been constructed in the offline stage. Then, the LLE-based feature prediction method for an input source feature vector $\mathbf{u}_i$ conducts the following three steps:

(a) Finding K nearest neighbors (measured by the Euclidean distance) of $\mathbf{u}_i$ from the source dictionary.
(b) Computing the reconstruction weight vector that best (linearly) reconstructs $\mathbf{u}_i$ from its K nearest neighbors found in step (a).
(c) Estimating the corresponding target feature vector by linearly combining K target exemplars (paired with the K nearest neighbors of $\mathbf{u}_i$) in the target dictionary with the reconstruction weight vector obtained in step (b).

The steps (a) and (b) involve identifying the locally linear patch and characterizing the local geometry of the locally linear patch, respectively, as described in steps (a) and (b) of the LLE algorithm for dimensionality reduction. On the other hand, the step (c)

involves estimating the target features by preserving the local geometry of the source features, as opposed to estimating the low-dimensional embedding in step (c) of the LLE algorithm for dimensionality reduction.

We implement steps (b) and (c) of the LLE-based feature prediction method as follows. In step (b), the reconstruction weight vector is computed by minimizing the reconstruction error ε_i subject to the constraint $\mathbf{1}^{\mathrm{T}}\mathbf{w}_i = 1$ (for the purpose of translational invariance) as:

$$\varepsilon_i = \|\mathbf{u}_i - \mathbf{A}_i\mathbf{w}_i\|^2, \text{ s.t. } \mathbf{1}^T\mathbf{w}_i = 1, \tag{1}$$

where $\mathbf{A}_i \in \mathcal{R}^{F\times K}$ is a matrix (a subset of the source dictionary) composed of K nearest neighbors of $\mathbf{u}_i$, *i.e.*, $\mathbf{A}_i = [\mathbf{a}_{i,1}, \ldots, \mathbf{a}_{i,k}, \ldots, \mathbf{a}_{i,K}]$, where $\mathbf{a}_{i,k} \in \mathcal{R}^{F\times 1}$ is the kth nearest neighbor of $\mathbf{u}_i$; $\mathbf{w}_i \in \mathcal{R}^{K\times 1}$ is the reconstruction weight vector for $\mathbf{u}_i$; $\mathbf{1} \in \mathcal{R}^{K\times 1}$ is a vector whose elements are all ones; and the superscript T denotes transposition of the vector. Note that $\mathbf{A}_i$ can be obtained in step (a). Solving $\mathbf{w}_i$ by minimizing ε_i subject to the constraint is a constrained least square problem, and the closed-form solution can be found in [37]. A more efficient way to obtain $\mathbf{w}_i$ is to solve the linear system of equations in advance:

$$\mathbf{G}_i\mathbf{w}_i = 1, \tag{2}$$

where $\mathbf{G}_i \in \mathcal{R}^{K\times K}$ is the local Gram matrix for $\mathbf{u}_i$:

$$\mathbf{G}_i = (\mathbf{A}_i - \mathbf{u}_i\mathbf{1}^T)^T(\mathbf{A}_i - \mathbf{u}_i\mathbf{1}^T). \tag{3}$$

Then, the reconstruction weight vector is rescaled to satisfy the constraint $\mathbf{1}^T\mathbf{w}_i = 1$. The detailed derivations of the solution can be found in [37].

In step (c), with the assumption that the source and target feature vectors share a similar local geometry in their respective feature spaces (manifolds) the predicted feature vector $\hat{\mathbf{v}}_i$ can be obtained by

$$\hat{\mathbf{v}}_i = \mathbf{B}_i\mathbf{w}_i, \tag{4}$$

where the reconstruction weight vector $\mathbf{w}_i$ is obtained in step (b); $\mathbf{B}_i \in \mathcal{R}^{F\times K}$ is a matrix (a subset of the target dictionary) corresponding to $\mathbf{A}_i$, and is composed of K target exemplars, *i.e.*, $\mathbf{B}_i = [\mathbf{b}_{i,1}, \ldots, \mathbf{b}_{i,k}, \ldots, \mathbf{b}_{i,K}]$, where $\mathbf{b}_{i,k} \in \mathcal{R}^{F\times 1}$ is the kth target exemplar in $\mathbf{B}_i$ corresponding to (paired with) $\mathbf{a}_{i,k}$.

Once the sample-by-sample prediction process is finished, a sequence of predicted feature vectors $\{\hat{\mathbf{v}}_i\}_{i=1}^{I}$ can be obtained.

2.2 The Direct LLE-Based Post-Filter

Fig. 2 gives an overview of the DL post-filter. The natural idea is to directly apply the LLE-based feature prediction method to perform post-filtering for speech enhancement, *i.e.*, predicting the clean spectral features from the enhanced spectral features. There are two stages in DL post-filtering: the offline and online stages. The offline stage mainly involves the construction of the paired dictionaries while the online stage performs post-filtering for speech enhancement. In the following, we describe the DL post-filter in detail.

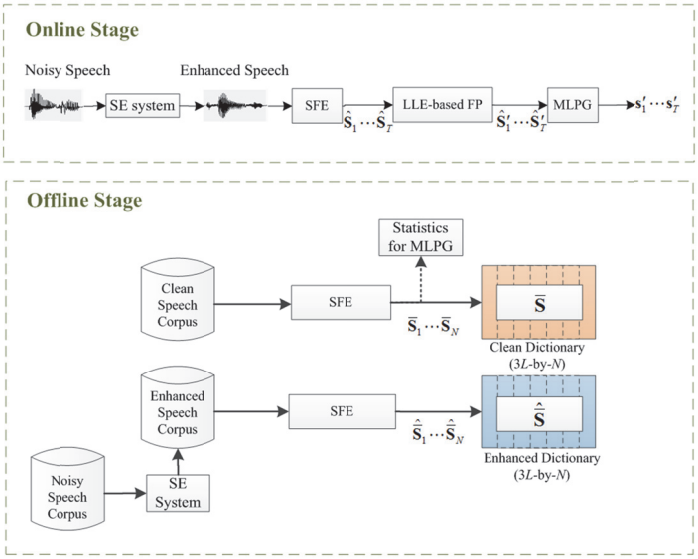


Fig. 2. Overview of the offline and online stages of the proposed direct LLE-based (DL) post-filter. "SE system", "SFE", and "LLE-based FP" modules denote "speech enhancement system", "spectral feature extraction", and "LLE-based feature prediction", respectively.

(A) The Offline Stage

As shown in Fig. 2, the paired enhanced and clean dictionaries are constructed in the following steps:

(a) Preparing three speech corpora: the clean speech, the corresponding noisy speech, and the corresponding enhanced speech corpora, where the enhanced speech corpus is obtained by applying a well-established speech enhancement system/method to the noisy speech corpus while the noisy speech corpus is obtained by artificially adding noises with different SNRs to the clean speech corpus (which will be described in detail in Section 3).
(b) Extracting the enhanced and clean spectral feature vectors from the enhanced and clean speech corpora, respectively. Note that each enhanced or clean spectral feature vector is composed of multi-dimensional static, delta, and delta-delta features.
(c) Constructing the paired dictionaries from the enhanced and clean spectral feature vectors.

Note that after conducting step (b), the statistics (*i.e.*, the precision matrix) to be used by the MLPG algorithm [36] in the online stage is estimated from the clean spectral features. The MLPG algorithm will be described later.

Let $\hat{\overline{\mathbf{S}}} \in \mathcal{R}^{3L\times N}$ and $\overline{\mathbf{S}} \in \mathcal{R}^{3L\times N}$ (as shown in Fig. 2) be the enhanced and clean dictionaries, and be composed of the enhanced and clean spectral feature vectors (or called exemplars) as $\hat{\overline{\mathbf{S}}} = [\hat{\overline{\mathbf{S}}}_1, \ldots, \hat{\overline{\mathbf{S}}}_n, \ldots, \hat{\overline{\mathbf{S}}}_N]$ and $\overline{\mathbf{S}} = [\overline{\mathbf{S}}_1, \ldots, \overline{\mathbf{S}}_n, \ldots, \overline{\mathbf{S}}_N]$, respectively, where the numbers of exemplars in both dictionaries are N; $\hat{\overline{\mathbf{S}}}_n \in \mathcal{R}^{3L\times 1}$ is the nth enhanced exemplar in the enhanced dictionary, and is composed of the L-dimensional static $\hat{\overline{\mathbf{s}}}_n \in \mathcal{R}^{L\times 1}$,

delta $\Delta^{(1)}\hat{\bar{\mathbf{s}}}_n \in \mathcal{R}^{L\times 1}$, and delta-delta $\Delta^{(2)}\hat{\bar{\mathbf{s}}}_n \in \mathcal{R}^{L\times 1}$ features as $\hat{\bar{\mathbf{S}}}_n = \left[\hat{\bar{\mathbf{s}}}_n^T, \Delta^{(1)}\hat{\bar{\mathbf{s}}}_n^T, \Delta^{(2)}\hat{\bar{\mathbf{s}}}_n^T\right]^T$ (for $n = 1\sim N$). Likewise, $\bar{\mathbf{S}}_n \in \mathcal{R}^{3L\times 1}$ is the nth clean exemplar in the clean dictionary, and is composed of the L-dimensional static $\bar{\mathbf{s}}_n \in \mathcal{R}^{L\times 1}$, delta $\Delta^{(1)}\hat{\bar{\mathbf{s}}}_n \in \mathcal{R}^{L\times 1}$, and delta-delta $\Delta^{(2)}\hat{\bar{\mathbf{s}}}_n \in \mathcal{R}^{L\times 1}$ features as $\bar{\mathbf{S}}_n = \left[\bar{\mathbf{s}}_n^T, \Delta^{(1)}\bar{\mathbf{s}}_n^T, \Delta^{(2)}\bar{\mathbf{s}}_n^T\right]^T$ (for n=1~N).

(B) The Online Stage

In Fig. 2, a well-established speech enhancement system is applied to an input noisy speech to obtain the enhanced speech in advance. Then, spectral feature extraction is performed to obtain the sequence of enhanced spectral feature vectors $\{\hat{\mathbf{S}}_t \in \mathcal{R}^{3L\times 1}\}_{t=1}^T$, where T is the number of speech frames of the enhanced speech, and $\hat{\mathbf{S}}_t$ is the enhanced spectral feature vector at frame t, which is composed of the L-dimensional static $\hat{\mathbf{s}}_t \in \mathcal{R}^{L\times 1}$, delta $\Delta^{(1)}\hat{\mathbf{s}}_t \in \mathcal{R}^{L\times 1}$, and delta-delta $\Delta^{(2)}\hat{\mathbf{s}}_t \in \mathcal{R}^{L\times 1}$ features, *i.e.*, $\hat{\mathbf{S}}_t = [\hat{\mathbf{s}}_t^T, \Delta^{(1)}\hat{\mathbf{s}}_t^T, \Delta^{(2)}\hat{\mathbf{s}}_t^T]^T$. Then, the LLE-based feature prediction method is applied to predict the clean spectral feature vectors $\{\hat{\mathbf{S}}'_t \in \mathcal{R}^{3L\times 1}\}_{t=1}^T$ from the enhanced spectral feature vectors $\{\hat{\mathbf{S}}_t\}_{t=1}^T$ independently in a frame-by-frame manner, where $\hat{\mathbf{S}}'_t$ is the predicted clean spectral feature vector at frame t. Note that the paired enhanced and clean dictionaries are used in the LLE-based feature prediction method.

To overcome the discontinuity problem existing in the predicted clean spectral features given by the LLE-based feature prediction method, the MLPG algorithm is applied to the predicted clean spectral feature vectors $\{\hat{\mathbf{S}}'_t\}_{t=1}^T$ to generate a sequence of final predicted static clean spectral feature vectors $\{\mathbf{s}'_t \in \mathcal{R}^{L\times 1}\}_{t=1}^T$, where $\mathbf{s}'_t$ is the final predicted static clean spectral feature vector at frame t. Next, we describe the MLPG algorithm in detail.

(C) The MLPG Algorithm

Since the LLE-based feature prediction method is performed in a frame-by-frame manner, the discontinuity problem exists. As suggested in our previous work [34, 35], the MLPG algorithm can effectively handle the discontinuity problem.

The MLPG algorithm [36] applied to the LLE-based feature prediction method is given as

$$\mathbf{s}' = (\mathbf{M}^T\mathbf{\Lambda}\mathbf{M})^{-1}\mathbf{M}^T\mathbf{\Lambda}\hat{\mathbf{S}}', \tag{5}$$

where $\mathbf{s}' = [(\mathbf{s}'_1)^T, \ldots, (\mathbf{s}'_t)^T, \ldots, (\mathbf{s}'_T)^T]^T \in \mathcal{R}^{LT\times 1}$ is a sequence of final predicted static clean spectral feature vectors; $\mathbf{M} \in \mathcal{R}^{3LT\times LT}$ is a weighting matrix used for appending the dynamic features to the static ones [36]; $\hat{\mathbf{S}}' = [(\hat{\mathbf{S}}'_1)^T, \ldots, (\hat{\mathbf{S}}'_t)^T, \ldots, (\hat{\mathbf{S}}'_T)^T]^T \in \mathcal{R}^{3LT\times 1}$ is a sequence of predicted clean spectral feature vectors obtained by the LLE-based feature prediction method; $\mathbf{\Lambda} = \mathrm{diag}[\mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_t, \ldots, \mathbf{\Lambda}_T] \in \mathcal{R}^{3LT\times 3LT}$ is the global precision matrix, where $\mathbf{\Lambda}_t \in \mathcal{R}^{3L\times 3L}$ is the precision matrix at frame t, which is assumed to be diagonal and is estimated from the clean speech corpus (clean spectral feature vectors). Note that $\mathbf{\Lambda}_1 = \ldots = \mathbf{\Lambda}_t = \ldots = \mathbf{\Lambda}_T$.

2.3 The LLE-Based Difference Compensation Post-Filter

Because the DL post-filter directly predicts the clean spectral features from the enhanced spectral features without utilizing the spectral information of the noisy speech,

the performance of post-filtering may depend heavily on the capability of the preceding speech enhancement system. Alternatively, we propose the LDC post-filter to cope with this problem. Fig. 3 gives an overview of the LDC post-filter, which includes the offline and online stages.

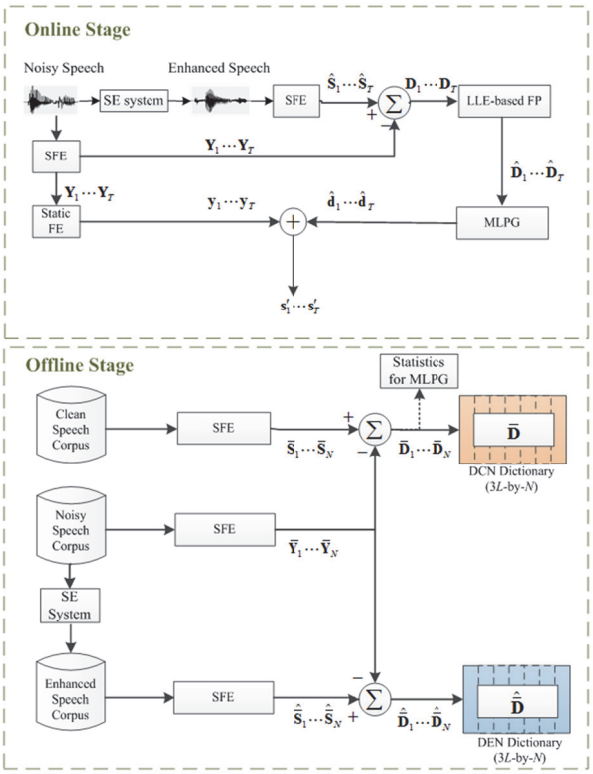


Fig. 3. Overview of the offline and online stages of the proposed LLE-based difference compensation (LDC) post-filter. "SE system", "SFE", and "LLE-based FP" modules denote "speech enhancement system", "spectral feature extraction", and "LLE-based feature prediction", respectively. "Static FE" module in the online stage extracts the static component $\mathbf{y}_t$ from the feature vector $\mathbf{Y}_t$ (for t=1~T).

(A) The Offline Stage

As shown in Fig. 3, the paired difference dictionaries (called the DEN and DCN dictionaries, respectively) are constructed in the following steps:

(a) Preparing three speech corpora: the clean speech, the corresponding noisy speech, and the corresponding enhanced speech corpora.
(b) Extracting the clean, noisy, and enhanced spectral features from the clean, noisy, and enhanced speech corpora, respectively. Note that each clean, noisy, or enhanced spectral feature vector is composed of multi-dimensional static, delta, and delta-delta features.

(c) Computing the spectral difference of {enhanced speech; noisy speech} (called the DEN features hereafter) and that of {clean speech; noisy speech} (called the DCN features hereafter).
(d) Constructing the paired DEN and DCN dictionaries from the paired DEN and DCN features.

Note that before conducting step (b), it is essential to normalize the energy of each enhanced speech utterance to match that of the corresponding clean speech utterance beforehand. In other words, we make the energy of each enhanced speech utterance match the energy of the clean speech component of the corresponding noisy speech utterance. Then, we perform steps (b)-(d). In this way, we avoid the mismatch between the DEN and DCN features caused by the energy mismatch between the enhanced speech and the clean speech component of the noisy speech. The necessity of this strategy has been confirmed in our preliminary results.

Additionally, after conducting step (c), we also calculate the statistics (*i.e.*, the precision matrix) of the DCN features to be used by the MLPG algorithm in the online stage.

Let $\overline{\mathbf{D}} \in \mathcal{R}^{3L\times N}$ and $\hat{\overline{\mathbf{D}}} \in \mathcal{R}^{3L\times N}$ be the DCN and DEN dictionaries, and be composed of the DCN and DEN feature vectors (or called exemplars) as $\overline{\mathbf{D}} = [\overline{\mathbf{D}}_1,...,\overline{\mathbf{D}}_n,...,\overline{\mathbf{D}}_N]$ and $\hat{\overline{\mathbf{D}}} = [\hat{\overline{\mathbf{D}}}_1,\cdots,\hat{\overline{\mathbf{D}}}_n,\cdots,\hat{\overline{\mathbf{D}}}_N]$, respectively, where the numbers of exemplars in both dictionaries are N; $\overline{\mathbf{D}}_n \in \mathcal{R}^{3L\times 1}$ is the nth DCN exemplar in the DCN dictionary, and is calculated as $\overline{\mathbf{S}}_n - \overline{\mathbf{Y}}_n$, where $\overline{\mathbf{S}}_n$ is the nth clean spectral feature vector as described previously, and $\overline{\mathbf{Y}}_n$ is the nth noisy spectral feature vector, and is composed of the L-dimensional static $\overline{\mathbf{y}}_n \in \mathcal{R}^{L\times 1}$, delta $\Delta^{(1)}\overline{\mathbf{y}}_n \in \mathcal{R}^{L\times 1}$, and delta-delta $\Delta^{(2)}\overline{\mathbf{y}}_n \in \mathcal{R}^{L\times 1}$ features as $\overline{\mathbf{Y}}_n = \left[\overline{\mathbf{y}}_n^{\mathrm{T}}, \Delta^{(1)}\overline{\mathbf{y}}_n^{\mathrm{T}}, \Delta^{(2)}\overline{\mathbf{y}}_n^{\mathrm{T}}\right]^{\mathrm{T}}$ (for n=1~N). Likewise, $\hat{\overline{\mathbf{D}}}_n \in \mathcal{R}^{3L\times 1}$ is the nth DEN exemplar in the DEN dictionary, and is calculated as $\hat{\overline{\mathbf{S}}}_n - \overline{\mathbf{Y}}_n$, where $\hat{\overline{\mathbf{S}}}_n$ is the nth enhanced spectral feature vector as defined previously (for n=1~N).

(B) The Online Stage

In Fig. 3, a well-established speech enhancement system is applied to the noisy speech to obtain the enhanced speech in advance. Then, spectral feature extraction is performed to obtain the sequence of enhanced spectral feature vectors $\{\hat{\mathbf{S}}_t\}_{t=1}^{T}$. Meanwhile, we also extract the sequence of noisy spectral feature vectors $\{\mathbf{Y}_t \in \mathcal{R}^{L\times 1}\}_{t=1}^{T}$, where $\mathbf{Y}_t$ is the noisy spectral feature vector at frame t, and is composed of the L-dimensional static $\mathbf{y}_t \in \mathcal{R}^{L\times 1}$, delta $\Delta^{(1)}\mathbf{y}_t \in \mathcal{R}^{L\times 1}$, and delta-delta $\Delta^{(2)}\mathbf{y}_t \in \mathcal{R}^{L\times 1}$ features as $\mathbf{Y}_t = [\mathbf{y}_t^T, \Delta^{(1)}\mathbf{y}_t^T, \Delta^{(2)}\mathbf{y}_t^T]^T$ (for t=1~T). Note that before extracting the noisy spectral feature vectors, we should make the energy of the enhanced speech utterance match the energy of the clean speech component of the corresponding noisy speech utterance for the same reason described early in the offline stage. Since the clean speech is not available during the online stage, we cannot apply the same procedure adopted in the offline stage. Alternatively, we first 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 normalize the energy of the input noisy speech such that the energy of the clean speech component of the noisy speech matches the energy of

the enhanced speech. The necessity of this strategy has been confirmed in our preliminary results.

Next, we obtain the DEN feature vectors by calculating the spectral difference of {enhanced speech; noisy speech} as $\{\mathbf{D}_t = \hat{\mathbf{S}}_t - \mathbf{Y}_t\}_{t=1}^{T}$, where $\mathbf{D}_t$ is the DEN feature vector at frame t. Then, the LLE-based feature prediction method is applied to predict the DCN feature vectors $\{\hat{\mathbf{D}}_t \in \mathcal{R}^{3L\times 1}\}_{t=1}^{T}$, from the DEN feature vectors $\{\mathbf{D}_t\}_{t=1}^{T}$ independently in a frame-by-frame manner, where $\hat{\mathbf{D}}_t$ is the predicted DCN feature vector at frame t. Note that the paired DEN and DCN dictionaries are used in the LLE-based feature prediction method.

To overcome the discontinuity problem, the MLPG algorithm is applied to $\{\hat{\mathbf{D}}_t\}_{t=1}^{T}$ in the same way as it is applied in the DL post-filter. As a result, a sequence of predicted static DCN feature vectors $\{\hat{\mathbf{d}}_t \in \mathcal{R}^{L\times 1}\}_{t=1}^{T}$ can be obtained, where $\hat{\mathbf{d}}_t$ is the predicted static DCN feature vector at frame t.

Finally, we obtain a sequence of final predicted static clean spectral feature vectors $\{\mathbf{s}'_t\}_{t=1}^{T}$ by compensating the noisy "static" spectral feature vectors $\{\mathbf{y}_t\}_{t=1}^{T}$ with the predicted static DCN feature vectors $\{\hat{\mathbf{d}}_t\}_{t=1}^{T}$ as $\{\mathbf{s}'_t = \hat{\mathbf{d}}_t + \mathbf{y}_t\}_{t=1}^{T}$.

2.4 Comparison between DL and LDC

The main difference between the DL and LDC post-filters in the offline stage is the construction of the dictionaries. Specifically, since the 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 clean spectral features from the enhanced ones while LDC predicts the clean-noisy spectral difference and compensates the input noisy spectral features with the predicted clean-noisy difference to generate the final predicted clean spectral features. As a result, LDC-processed speech may retain more speech details from the noisy speech.

3. EXPERIMENTS

We conducted two sets of experiments to evaluate the effectiveness of the proposed LLE-based post-filters. In this section, we first describe the experimental setup in Section 3.1. Then, in Section 3.2, we present the evaluation results of applying the LLE-based post-filters to a DDAE-based speech enhancement system [29], which is a representative supervised speech enhancement system. Finally, in Section 3.3, we present the evaluation results of applying the LLE-based post-filters to a MMSE spectral estimation-based speech enhancement system [13], which is a representative unsupervised speech enhancement system.

3.1 Experimental Setup

Our experiments were conducted on the Mandarin hearing in noise test (MHINT)

sentences [38], which contained 300 utterances pronounced by a male native Mandarin speaker recorded in a clean condition room. The maximum, minimum, and average durations of the utterances were around 4.4, 1.9, and 3 seconds, respectively. Speech signals were recorded in a 16 kHz/16 bit format.

In Section 3.2, we compared the following three systems:

- ***DDAE*:** The baseline DDAE-based speech enhancement system integrated with the MLPG algorithm [29].
- ***DDAE-DL*:** The system that applies the DL post-filter to ***DDAE*** for further suppressing residual noises in enhanced speech signals obtained by ***DDAE***.
- ***DDAE-LDC*:** The system that applies the LDC post-filter to ***DDAE*** for further suppressing residual noises in enhanced speech signals obtained by ***DDAE***.

Specifically, for the ***DDAE*** system, 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. The SNRs ranged from −10 to 20 dB with a 5 dB interval. As a result, for each noise type, 1750 noisy speech utterances paired with the corresponding clean speech utterances were generated as the training set. The DDAE model consisted of seven hidden layers with 1200, 300, 300, 514, 300, 300, and 1200 hidden nodes, respectively. Two DDAEs, 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 were 32 and 16 milliseconds, respectively. The Hamming window was used in the framing process. Each frame of speech was converted to a "static" feature vector with 257-dimensional log-power spectral features, based on a 512-point discrete short-time Fourier transform analysis. The immediately preceding and following contextual feature vectors were then appended to the current one to form the final spectral feature vector, whose dimension was 771 (257 × 3). During spectral feature generation, the MLPG algorithm was adopted to overcome the discontinuity issue as described in [29].

A five-fold cross validation was performed to evaluate ***DDAE-DL*** based on the 50 utterances in the test set. In each run, we constructed the paired enhanced and clean dictionaries using 40 utterances and tested performance using the remaining 10 utterances. The rationale behind this setup is that the proposed post-filters are supposed to support all existing speech enhancement systems. In other words, the training data for these speech enhancement systems should not be assumed available, and the post-filters should be developed independently. The dictionaries were built using the data at -10, 0, and 10 dB SNRs. Thus, for each noise type, there were 120 clean and the corresponding enhanced utterances (obtained by ***DDAE***) to build the paired dictionaries. The dictionaries contained about 24,000 exemplars for each run. The signal analysis part was the same as that used in ***DDAE***, except that the power spectra of each frame were normalized to unit-sum, and the normalizing factor was saved to be used in the reconstruction step. Then, logarithms were applied to the normalized power spectra. The static, delta, and delta-delta features were used, and thus the dimensionality of a final vector was 771 (257 × 3). After performing ***DDAE-DL***, the predicted log normalized power spectra were reverted back to the (linear) normalized power spectra, which were then compensated back

to the power spectra by the normalizing factor. The number of nearest neighbors, namely K in (1)-(4), for the LLE-based feature prediction method applied in DL was set to 1024 empirically. For a fair comparison, the ***DDAE-LDC*** system adopted the same setup as ***DDAE-DL***.

In Section 3.3, we compared the following three systems:

- ***MMSE*****:** The conventional MMSE spectral estimation-based speech enhancement system [13].
- ***MMSE-DL*****:** The system that applies the DL post-filter to ***MMSE*** for further suppressing residual noises in enhanced speech signals obtained by ***MMSE***.
- ***MMSE-LDC*****:** The system that applies the LDC post-filter to ***MMSE*** for further suppressing residual noises in enhanced speech signals obtained by ***MMSE***.

Specifically, for the ***MMSE*** system, the "decision-directed" method is used for tracking of a priori SNR tracking. The signal analysis process was the same as that used in ***DDAE***. For the ***MMSE-DL*** and ***MMSE-LDC*** systems, the setup is the same as that in ***DDAE-DL*** and ***DDAE-LDC***, except that the enhanced dictionary and the enhanced-noisy dictionary were constructed by the enhanced speech utterances obtained by ***MMSE*** rather than ***DDAE***.

For all the systems mentioned above, we used an overlap-add method to synthesize the waveform from the final estimated/enhanced spectral features with the phase information of the original noisy speech.

3.2 Evaluation of the LLE-Based Post-Filters for DDAE

(A) Objective Evaluations

We used the following three metrics for objective evaluation: the perceptual evaluation of speech quality (PESQ) [39], the short-time objective intelligibility measure (STOI) [40], and the segmental signal-to-noise ratio improvement (SSNRI, in dB) [41]. The ranges of PESQ and STOI scores are {–0.5 to 4.5} and {0 to 1}, where higher scores indicate better speech quality and better intelligibility, respectively. On the other hand, SSNRI represents the difference in the segmental SNR between the enhanced speech and the noisy speech for measuring the degree of noise reduction. Therefore, the SSNRI score of the noisy speech is 0 dB, and a higher SSNRI score of the enhanced speech indicates that the noise in the noisy speech has been removed effectively. Tables 1 and 2, respectively, show the objective evaluation scores obtained by ***DDAE***, ***DDAE-DL***, and ***DDAE-LDC*** in the two-talker and car noises at different SNRs. The scores of the unprocessed noisy speech are provided for reference.

We first compared the scores achieved by ***DDAE***, ***DDAE-DL***, and ***DDAE-LDC*** with the scores of the noisy speech. From Table 1, we observe that generally all speech enhancement systems could effectively handle the non-stationary noise (*i.e.*, the two-talker noise) with yielding higher PESQ and SSNRI scores over the noisy speech. With a further analysis, we note that ***DDAE-LDC*** improved the STOI score over the noisy speech at all SNRs except 10dB, while ***DDAE*** and ***DDAE-DL*** only improved the STOI score at low SNRs. From Table 2, it is noted that all speech enhancement systems could improve the SSNRI score when dealing with the stationary noise (*i.e.*, the car noise).

Meanwhile, ***DDAE-LDC*** could yield higher PESQ scores over the noisy speech at all SNRs. However, it degraded the STOI score at high SNRs. On the other hand, ***DDAE*** and ***DDAE-DL*** degraded PESQ and STOI scores at high SNRs. In summary, the results from Tables 1 and 2 reveal that all speech enhancement systems can effectively handle the two-talker and car noises except that ***DDAE*** and ***DDAE-DL*** tends to degrade the speech quality and intelligibility in the car noise, particularly at high SNRs. However, the improvements at low SNRs could be more valuable in many applications.

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

	Noisy Speech		***DDAE***			***DDAE-DL***			***DDAE-LDC***		
	PESQ	STOI	PESQ	STOI	SSNRI	PESQ	STOI	SSNRI	PESQ	STOI	SSNRI
SNR10	2.11	**0.91**	2.21	0.88	2.48	2.22	0.83	2.73	**2.74**	0.90	**3.99**
SNR6	1.81	0.86	2.05	0.86	5.76	2.11	0.82	6.08	**2.44**	**0.88**	**7.04**
SNR2	1.60	0.79	1.93	0.84	8.47	1.97	0.80	8.88	**2.22**	**0.86**	**9.51**
SNR0	1.55	0.75	1.83	0.83	9.66	1.86	0.79	10.12	**2.08**	**0.84**	**10.57**
SNR-2	1.43	0.70	1.75	0.81	10.46	1.78	0.78	11.03	**1.95**	**0.82**	**11.29**
SNR-6	1.32	0.60	1.61	**0.78**	11.38	1.59	0.75	**12.13**	**1.74**	**0.78**	11.39
SNR-10	1.28	0.51	1.47	0.72	11.51	1.42	0.69	**12.53**	**1.56**	**0.73**	11.12
Ave.	1.59	0.73	1.83	0.82	8.53	1.85	0.78	9.07	**2.10**	**0.83**	**9.27**

Table 2. PESQ, STOI, and SSNRI of *DDAE*, *DDAE-DL*, and *DDAE-LDC* evaluated on the test set at different SNRs of the *car* noise.

	Noisy Speech		***DDAE***			***DDAE-DL***			***DDAE-LDC***		
	PESQ	STOI	PESQ	STOI	SSNRI	PESQ	STOI	SSNRI	PESQ	STOI	SSNRI
SNR10	2.61	**0.95**	1.96	0.85	5.04	2.03	0.80	5.73	**3.10**	0.90	**7.59**
SNR6	2.27	**0.92**	1.93	0.84	8.17	1.99	0.79	8.91	**2.88**	0.88	**10.59**
SNR2	1.96	**0.87**	1.89	0.83	10.40	1.92	0.78	11.37	**2.59**	0.86	**12.63**
SNR0	1.84	0.85	1.85	0.82	11.40	1.86	0.78	12.34	**2.43**	**0.85**	**13.40**
SNR-2	1.71	0.82	1.81	0.81	12.00	1.82	0.77	13.05	**2.28**	**0.83**	**13.85**
SNR-6	1.53	0.76	1.75	0.79	12.34	1.71	0.75	13.74	**2.02**	**0.80**	**13.97**
SNR-10	1.43	0.71	1.67	**0.76**	12.22	1.60	0.72	**13.90**	**1.82**	0.75	13.45
Ave.	1.91	0.84	1.84	0.81	10.23	1.85	0.77	11.29	**2.44**	**0.84**	**12.21**

Next, we evaluated the effectiveness of the proposed DL and LDC post-filters by comparing ***DDAE-DL*** and ***DDAE-LDC*** with ***DDAE***. From the results in Tables 1 and 2, we first observe that both ***DDAE-DL*** and ***DDAE-LDC*** achieved better SSNRI scores than ***DDAE*** in both noise types at all SNRs, showing that the residual noises in the ***DDAE*** enhanced speech can be further removed by the DL and LDC post-filters. Comparing ***DDAE-DL*** with ***DDAE***, we observe that ***DDAE-DL*** obtained slightly higher PESQ scores than ***DDAE*** in both noise types at high SNRs (*i.e.*, 10 ~ −2 dB SNRs), suggesting that the DL post-filter can provide additional speech quality improvements in higher SNR conditions. However, we also observe that ***DDAE-DL*** was inferior to ***DDAE*** in terms of STOI under all SNRs and noise types, suggesting that although the DL post-filter can notably improve SSNRI and speech quality, it tends to deteriorate speech intelligibility. On the contrary, comparing ***DDAE-LDC*** with ***DDAE***, we note that ***DDAE-LDC*** improved the PESQ and STOI scores over ***DDAE*** across different SNR levels and noisy types. Consider that ***DDAE-DL*** cannot effectively improve the ***DDAE*** enhanced speech under low SNR conditions, the results suggest that ***DDAE-LDC*** possesses a better ability to avoid distortions than ***DDAE-DL***. With a further comparison,

DDAE-LDC achieved better speech intelligibility, speech quality, and SSNRI scores in most conditions than ***DDAE-DL***. The results confirm that by reducing the uncertainty of the paired dictionaries and maintaining speech details from the noisy speech, the LDC post-filter can attain better enhancement performance than the DC post-filter. In summary, ***DDAE-LDC*** outperformed ***DDAE*** and ***DDAE-DL***.

In addition to the above quantitative evaluation comparison, we provide the qualitative comparison in Fig. 4, which presents the spectrograms of the clean, noisy and ***DDAE***, ***DDAE-DL*** and ***DDAE-LDC*** enhanced speech. From the figure, we observe that although all speech enhancement systems could effectively remove the noise components in the spectrum domain, ***DDAE*** and ***DDAE-DL*** actually lost some speech details in the high-frequency regions. ***DDAE-LDC***, on the contrary, 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 above objective evaluation results: ***DDAE-LDC*** yields better speech quality, speech intelligibility, and noise reduction than ***DDAE*** and ***DDAE-DL***.

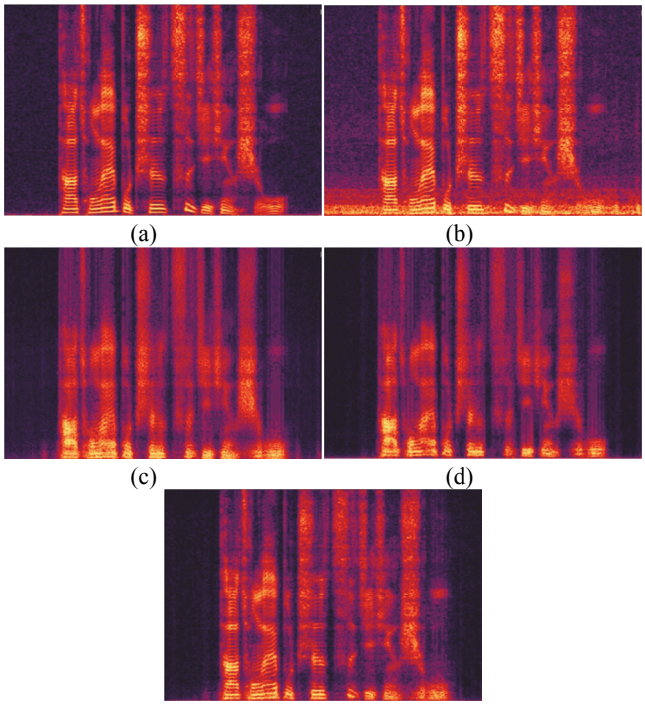


Fig. 4. Spectrograms of an utterance example; (a) original clean speech; (b) unprocessed noisy speech; (c) ***DDAE*** enhanced speech; (d) ***DDAE-DL*** enhanced speech, and (e) ***DDAE-LDC*** enhanced speech, in the ***car*** noise at SNR = 6 dB.

(B) Subjective Evaluations

From the results of objective evaluations, we could not clearly note the effectiveness of the DL post-filter over ***DDAE*** (*i.e.*, ***DDAE-DL*** slightly improved the PESQ and

SSNRI scores over ***DDAE*** but degraded the STOI score). Therefore, we compared ***DDAE-DL*** with ***DDAE*** by conducting the subjective noise reduction capability and preference tests. In the noise reduction capability test, the subjects were asked to select one from two utterances that was with a more notable noise reduction performance. During the preference test, the subjects were asked to select one from two utterances according to the overall preference, including the speech quality, listening effort, and noise reduction capability.

The test utterances were generated under two noise types (*i.e.*, the two-talker and car noises) at three SNRs (*i.e.*, −6, 0, and 6 dB). Note that −6dB and 6dB were not seen in both ***DDAE*** training and dictionary construction of the DL post-filter. Fifteen pairs of utterances were tested for each noise type and SNR combination. We conducted the AB test, *i.e.*, each pair of enhanced speech utterances by methods A and B were presented in a random order to the subjects. Twelve subjects were involved in the tests. Figs. 5 and 6 show the results of the noise reduction capability test and the preference test, respectively. From Fig. 5, we observe that ***DDAE-DL*** outperformed ***DDAE*** in all experimental conditions. The result confirms that the residual noises in the ***DDAE*** enhanced speech signals can be further removed by the DL post-filter through a spectral prediction process. The result is consistent with the objective evaluation result in terms of SSNRI shown in Tables 1 and 2. From Fig. 6, we also observe that ***DDAE-DL*** achieved a significant gain over ***DDAE*** in all experimental conditions. The result again demonstrates the effectiveness of the DL post-filter for speech enhancement. It is worth mentioning that the main factor considered in the preference test is the noise reduction capability according to the

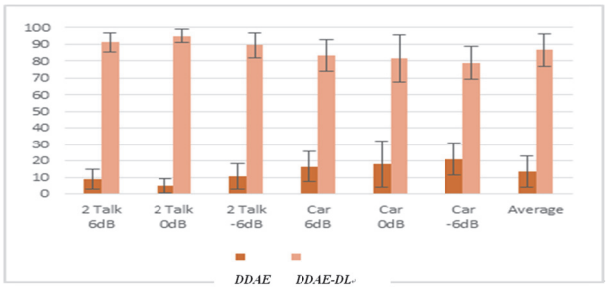


Fig. 5. Noise reduction capability test results for ***DDAE*** and ***DDAE-DL*** in two noise types (2 Talk: the two-talker noise, Car: the car noise) at three SNRs (−6, 0, 6 dB). Error bars indicate the 95% confidence intervals.

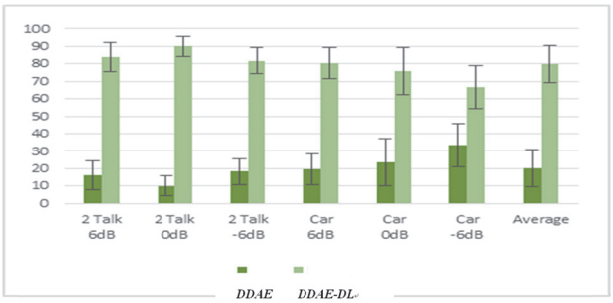


Fig. 6. Preference test results for ***DDAE*** and ***DDAE-DL*** in two noise types (2 Talk: the two-talker noise, Car: the car noise) at three SNRs (−6, 0, 6 dB). Error bars indicate the 95% confidence intervals.

subjects' responses. A possible reason is that the speech quality and listening effort of both systems are similar to each other (cf. the scores of PESQ and STOI in Tables 1 and 2); therefore, the noise reduction capability becomes the most important factor for the listening tests.

We further compared ***DDAE-DL*** with ***DDAE-LDC*** by conducting the subjective preference test. Again, −6 dB and 6 dB SNR levels were not seen in the dictionary constructions of the LDC post-filter. Moreover, 15 pairs of enhanced speech were tested for each condition, and 10 subjects were involved in the tests. The test results are listed in Fig. 7. From the figure, we observe that ***DDAE-LDC*** outperformed ***DDAE-DL*** in all test conditions. Specifically, ***DDAE-LDC*** provided remarkably higher scores under the stationary noise condition (*i.e.*, the car noise) but achieved relatively smaller improvements in the non-stationary noise case (*i.e.*, the two-talker noise). Based on the interviews with the subjects, we found that although the speech quality and listening effort were significantly improved, ***DDAE-LDC*** brought unwanted noise to the enhanced speech, especially under low SNR and non-stationary noise conditions. The listening test result is consistent with the objective evaluation result in terms of SSNRI that the improvements of ***DDAE-LDC*** over ***DDAE-DL*** under the car noise with high SNR condition are more prominent than the improvements under the two-talker noise with low SNR condition.

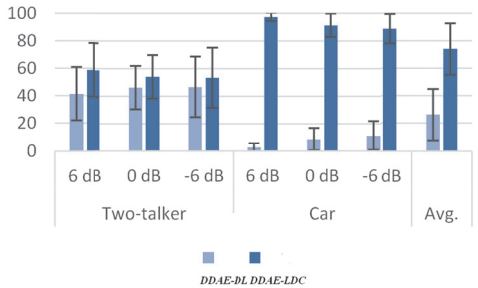


Fig. 7. Preference test results for ***DDAE-DL*** and ***DDAE-LDC*** in two noise types (the car and two-talker noises) at three SNRs (−6, 0, 6 dB). Error bars indicate the 95% confidence intervals.

3.3 Evaluation of the LLE-Based Post-Filters for MMSE

In Section 3.2, we have confirmed the effectiveness of integrating the LLE-based post-filters with the DDAE-based speech enhancement system. In this section, we investigate the compatibility of the LLE-based post-filters with a conventional speech enhancement system, *i.e.*, the MMSE spectral estimation-based speech enhancement system. Tables 3 and 4 report the PESQ, STOI, and SSNRI scores obtained by ***MMSE***, ***MMSE-DL***, and ***MMSE-LDC*** in the two-talker and car noises, respectively, at different SNRs.

We first compared the scores achieved by ***MMSE***, ***MMSE-DL***, and ***MMSE-LDC*** with the scores of the noisy speech. From Table 3, we observe that generally all speech enhancement systems yielded improvements in terms of SSNRI over the noisy speech scores except that ***MMSE-LDC*** gave negative SSNRI scores in the two-talker noise at high SNRs (*i.e.*, 6~10dB). Moreover, we also observe that ***MMSE*** and ***MMSE-LDC*** systems gave slightly higher PSEQ scores than the noisy speech in the two-talker noise at high SNRs (*i.e.*, 2~10 dB), and there were no obviously differences in the PESQ score

among ***MMSE***, ***MMSE-LDC***, and the noisy speech at low SNRs. However, ***MMSE-DL*** yielded lower PESQ scores than the noisy speech across all SNRs. Finally, all the three speech enhancement systems tended to degrade the speech intelligibility with yielding lower STOI scores than the noisy speech. In summary, the results from Table 3 reveal that although all speech enhancement systems show their capability to improve SNR in the non-stationary noise (*i.e.*, the two-talker noise), they cannot effectively improve or even may degrade the speech quality and speech intelligibility over the noisy speech.

Table 3. PESQ, STOI, and SSNRI of *MMSE*, *MMSE-DL*, and *MMSE-LDC* evaluated on the test set at different SNRs of the *two-talker* noise.

	Noise Speech		***MMSE***			***MMSE-DL***			***MMSE-LDC***		
	PESQ	STOI	PESQ	STOI	SSNRI	PESQ	STOI	SSNRI	PESQ	STOI	SSNRI
SNR10	2.11	**0.91**	**2.20**	0.88	0.26	1.85	0.78	**0.29**	2.17	0.88	–1.51
SNR6	1.81	**0.86**	**1.88**	0.83	0.67	1.66	0.74	**2.40**	1.83	0.83	–0.55
SNR2	1.60	**0.79**	**1.61**	0.72	0.91	1.49	0.68	**3.84**	1.60	0.75	0.40
SNR0	1.55	**0.75**	**1.55**	0.69	0.92	1.39	0.65	**4.25**	1.53	0.72	0.67
SNR-2	1.43	**0.70**	**1.44**	0.62	0.96	1.32	0.60	**4.69**	1.42	0.67	0.74
SNR-6	1.32	**0.60**	1.27	0.51	0.97	1.25	0.55	**4.64**	**1.32**	0.58	0.77
SNR-10	**1.28**	**0.51**	1.27	0.43	1.08	1.17	0.46	**4.20**	1.25	0.50	0.91
Ave.	1.59	**0.73**	**1.60**	0.67	0.82	1.45	0.64	**3.47**	1.59	0.71	0.21

Table 4. PESQ, STOI, and SSNRI of *MMSE*, *MMSE-DL*, and *MMSE-LDC* evaluated on the test set at different SNRs of the *car* noise.

	Noise Speech		***MMSE***			***MMSE-DL***			***MMSE-LDC***		
	PESQ	STOI	PESQ	STOI	SSNRI	PESQ	STOI	SSNRI	PESQ	STOI	SSNRI
SNR10	2.61	**0.95**	3.06	0.92	4.83	2.15	0.82	4.85	**3.25**	0.93	**5.23**
SNR6	2.27	**0.92**	2.70	0.89	6.01	2.02	0.80	**7.43**	**2.95**	0.90	6.98
SNR2	1.96	**0.87**	2.36	0.84	5.74	1.89	0.77	**8.53**	**2.53**	0.86	7.47
SNR0	1.84	**0.85**	2.21	0.81	5.55	1.82	0.76	**8.88**	**2.38**	0.83	7.65
SNR-2	1.71	**0.82**	2.05	0.79	5.38	1.75	0.74	**9.21**	**2.15**	0.81	7.46
SNR-6	1.53	0.76	1.79	0.73	3.72	1.55	0.69	**8.33**	**1.84**	**0.76**	6.17
SNR-10	1.43	0.71	1.59	0.66	2.95	1.41	0.64	**7.58**	1.61	0.70	5.05
Ave.	1.91	0.84	2.25	0.81	4.88	1.80	0.74	**7.83**	2.39	0.83	6.57

Next, we evaluated the effectiveness of the proposed DL and LDC post-filters by comparing ***MMSE-DL*** and ***MMSE-LDC*** with ***MMSE*** in the two-talker noise. From Table 3, we first observe that ***MMSE-DL*** outperformed ***MMSE*** in terms of SSNRI across all SNR levels, indicating that the DL post-filter can effectively remove the residual noises in the ***MMSE*** enhanced speech (consistent with the result in Table 1). On the contrary, ***MMSE-LDC*** gave lower SSNRI scores than ***MMSE***, suggesting that the LDC post-filter tends to introduce additional noise components to the ***MMSE*** enhanced speech (different from the result in Table 1). We also observe that ***MMSE-DL*** generally yielded lower STOI scores than ***MMSE***, indicating that the DL post-filter tends to degrade the speech intelligibility, and ***MMSE-LDC*** gave higher STOI scores than ***MMSE*** at most SNR levels, indicating that the LDC post-filter can further improve the intelligibility of the ***MMSE*** enhanced speech (consistent with the result in Table 1). On the other hand, we observe that ***MMSE*** obtained the highest PESQ scores at most SNR levels, showing that both LLE-based post-filters may degrade the speech quality (different from the result

in Table 1). In summary, the results above reveal that the integration of the proposed post-filters with ***MMSE*** is not as effective as the integration with ***DDAE*** in the two talker noise (a non-stationary noise). The reason could be that the results of the post-filters heavily depend on the performance (in terms of PESQ, STOI, and SSNRI) of the preceding speech enhancement system. Since ***MMSE*** could not handle the non-stationary noise well, the post-filters could not provide further improvements. This can be verified by comparing the objective results of ***DDAE*** in Table 1 with those of ***MMSE*** in Table 3. Actually the results are expectable since our preliminary results also revealed that applying the proposed post-filters directly to the noisy speech (without enhancement) could not achieve satisfactory speech enhancement performance.

Next, we evaluated the effectiveness of the proposed DL and LDC post-filters in the car noise, which is relatively more stationary than the two-talker noise. From Table 4, we first observe that both ***MMSE-DL*** and ***MMSE-LDC*** outperformed ***MMSE*** in terms of SSNRI, suggesting that the proposed post-filters can effectively remove the residual noises in the ***MMSE*** enhanced speech (consistent with the result in Table 2). Next, we observe that ***MMSE-DL*** gave lower STOI scores than ***MMSE*** across all SNR levels, indicating that the DL post-filter tends to degrade the speech intelligibility (consistent with the result in Table 2). On the other hand, ***MMSE-LDC*** gave higher STOI scores than ***MMSE***, indicating that the LDC post-filter can further improve the speech intelligibility (consistent with the result in Table 2). We also observe that ***MMSE-DL*** obtained lower PESQ scores than ***MMSE*** across all SNR levels, showing that the DL post-filter tends to degrade the speech quality (different from the result in Table 2). Meanwhile, ***MMSE-LDC*** obtained higher PESQ scores than ***MMSE*** across all SNR levels, indicating that the LDC post-filter can further improve the speech quality (consistent with the result in Table 2). In summary, the results from Tables 3 and 4 reveal that applying the LDC post-filter for ***MMSE*** is effective under the stationary noise condition. The result again confirms that if the preceding speech enhancement system can handle the noisy speech well, the LDC post-filter can further improve the speech enhancement performance. On the other hand, although the DL post-filter can effectively remove the residual noises, it tends to degrade the speech quality and intelligibility under both stationary and non-stationary noise types.

4. CONCLUSIONS

In this paper, we have proposed a novel LLE-based post-filtering approach with the aim to further suppress the residual noises in the enhanced speech signals obtained by a speech enhancement system. Two types of LLE-based post-filters have been presented: the DL post-filter and the LDC post-filter. The DL post-filter improves the enhanced speech (obtained by a speech enhancement system) by directly predicting the clean spectral features from the enhanced spectral features while the LDC post-filter improves the enhanced speech by predicting the spectral difference of {clean speech; noisy speech} from that of {enhanced speech; noisy speech} and then compensating the noisy spectral features with the predicted spectral difference. Our major findings are:

- Both of the proposed post-filters can further improve the supervised DDAE-based speech enhancement system under different noise types and SNR levels. Particularly,

the LDC post-filter achieve notable improvements over the DL post-filter due to the fact that the LDC post-filter introduces the noisy speech information in the difference prediction and compensation stages. As a result, the paired DEN and DCN dictionaries in LDC can reduce the uncertainty of the paired enhanced and clean dictionaries in DL, and the LDC-processed speech may retain more spectral details from noisy speech.

- The LDC post-filter can also improve the unsupervised MMSE spectral estimation-based speech enhancement system, under the stationary noise type, *e.g.*, the car noise, with different SNR levels. However, it fails to improve the MMSE-based speech enhancement system under the non-stationary noise type, *e.g.*, the two talker noise. On the other hand, the DL post-filter can effectively remove the residual noises in the enhanced speech obtained by the MMSE-based speech enhancement system in different noise types and SNR levels. However, it may notably degrade the speech quality and speech intelligibility.
- Whether the proposed post-filters can further suppress the residual noises in the enhanced speech signals seems to depend on the capability/performance of the preceding speech enhancement system.

For future work, we will evaluate the proposed LLE-based post-filters on more speech enhancement systems and noise types. In the meanwhile, we will derive algorithms to speed up the online nearest-neighbor searching for the LLE algorithm. Finally, we plan to extend the current scenario (with specified target speaker) to a speaker independent one.

REFERENCES

1. H. Levitt, "Noise reduction in hearing aids: an overview," *Journal of Rehabilitation Research and Development*, Vol. 38, 2001, pp. 111-121.
2. 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*, Vol. 64, 2016, pp. 1568-1578.
3. B. H. Juang and F. K. Soong, "Hands-free telecommunications," in *Proceedings of International Workshop on Hands-Free Speech Communication*, 2001, pp. 5-10.
4. S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Processing*, Vol. 64, 1998, pp. 21-32.
5. J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust Automatic Speech Recognition: A Bridge to Practical Applications*, 1st ed., Academic Press, NY, 2015.
6. B. Li, Y. Tsao, and K. C. Sim, "An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition," in *Proceedings of Interspeech*, 2013, pp. 3002-3006.
7. A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 7092-7096.
8. A. El-Solh, A. Cuhadar, and R. A. Goubran, "Evaluation of speech enhancement techniques for speaker identification in noisy environments," in *Proceedings of the 9th IEEE International Symposium on Multimedia Workshops*, 2007, pp. 235-239.

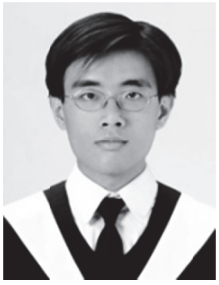
9. J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Proceedings of International Conference on Spoken Language Processing*, 1996, pp. 929-932.
10. S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics*, *Speech and Signal Processing*, Vol. 27, 1979, pp. 113-120.
11. P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proceedings of IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, 1996, pp. 629-632.
12. V. Grancharov, J. Samuelsson, and B. Kleijn, "On causal algorithms for speech enhancement," *IEEE Transactions on Audio*, *Speech*, *and Language Processing*, Vol. 14, 2006, pp. 764-773.
13. 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, 1984, pp. 1109-1121.
14. M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication*, Vol. 10, 1991, pp. 45-57.
15. Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, Vol. 11, 2003, pp. 334-341.
16. J. Karhunen and J. Joutsensalo, "Representation and separation of signals using nonlinear PCA type learning," *Neural Networks*, Vol. 7, 1994, pp. 113-127.
17. R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics*, *Speech and Signal Processing*, Vol. 34, 1986, pp. 744-754.
18. T. F. Quatieri and R. J. McAulay. "Shape-invariant time-scale and pitch modifications of speech," *IEEE Transactions on Signal Processing*, Vol. 40, 1992, pp. 497-510.
19. J. Makhoul, "Linear prediction: A tutorial review," in *Proceedings of IEEE*, Vol. 63, 1975, pp. 561-580.
20. B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Transactions on Acoustics*, *Speech and Signal Processing*, Vol. ASSP-27, 1979, pp. 247-254.
21. L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, 1986, pp. 4-16.
22. L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proceedings of IEEE*, Vol. 77, 1989, pp. 257-286.
23. 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, 2012, pp. 1698-1712.
24. 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, 2013, pp. 2140-2151.
25. K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proceedings of IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2008, pp. 4029-4032.

26. 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, 2014 pp. 65-68.
27. X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proceedings of Interspeech*, 2013, pp. 436-440.
28. X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Ensemble modeling of denoising autoencoder for speech spectrum restoration," in *Proceedings of Interspeech*, 2014, pp. 885-889.
29. 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 *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2015, pp. 365-369.
30. F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proceedings of the 12th International Conference on Latent Variable Analysis and Signal Separation*, 2015, pp. 91-99.
31. L. Sun, J. Du, L. R. Dai, and C. H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proceedings of Hands-free Speech Communications and Microphone Arrays*, 2017, pp. 136-140.
32. S. W. Fu, Y. Tsao, and X. Lu, "SNR-Aware convolutional neural network modeling for speech enhancement," in *Proceedings of Interspeech*, 2016, pp. 8-12.
33. S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, Vol. 290, 2000, pp. 2323-2326.
34. Y. C. Wu, H. T. Hwang, C. C. Hsu, Y. Tsao, and H. M. Wang, "Locally linear embedding for exemplar-based spectral conversion," in *Proceedings of Interspeech*, 2016, pp. 1652-1656.
35. 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 *Proceedings of IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2017, pp. 5555-5559.
36. K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proceedings of IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2000, pp. 1315-1318.
37. L. K. Saul and S. T. Roweis, "An introduction to locally linear embedding," http://www.cs.nyu.edu/~roweis/lle/papers/lleintro.pdf, 2001.
38. 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, 2007, pp. 70S-74S.
39. 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.
40. 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, 2011, pp. 2125-2136.

41. Y. Tsao, and Y. H. Lai, "Generalized maximum a posteriori spectral amplitude estimation for speech enhancement," *Speech Communication*, Vol. 76, 2016, pp. 112-126.

Hsin-Te Hwang (黃信德) received the M.S. degree in Department of Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan. He is currently pursuing the Ph.D. degree in Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan. He is also a Research Assistant in Institute of Information Science, Academia Sinica, Taipei, Taiwan. His research interests include speech signal processing, particularly, voice conversion, speech enhancement, and speech synthesis.

Yi-Chiao Wu (吳宜樵) received the M.S. degree in Institute of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan. He is a Research Assistant in Institute of Information Science, Academia Sinica, Taipei, Taiwan. His research interests include speech signal processing, particularly, voice conversion, speech enhancement, and speaker identification.



Syu-Siang Wang (王緒翔) received the M.S. degree in Department of Electrical Engineering, National Chi Nan University, Nantou, Taiwan. He is currently working toward the Ph.D. degree in the Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan. He is a Research Assistant in the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His research interests include signal processing, speech recognition, and deep learning.

Chin-Cheng Hsu (許晉誠) received the M.S. degree in Institute of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan. He is a Research Assistant in Institute of Information Science, Academia Sinica, Taipei, Taiwan. His research interests include machine learning and speech signal processing, particularly, voice conversion, speech synthesis, and speech enhancement.

Yu Tsao (曹昱) received the B.S. and M.S. degrees in Electrical Engineering from National Taiwan University in 1999 and 2001, respectively, and the Ph.D. degree in Electrical and Computer Engineering from Georgia Institute of Technology in 2008. From 2009 to 2011, Dr. Tsao was a researcher at National Institute of Information and Communications Technology, Japan, where he engaged in research and product development in automatic speech recognition for multilingual speech-to-speech translation. Currently, he is an Associate Research Fellow at the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. He received the Academia Sinica Career Development Award in 2017. Dr. Tsao's research interests include speech and speaker recognition, acoustic and language modeling, audio-coding, and bio-signal processing.

Hsin-Min Wang (王新民) received the B.S. and Ph.D. degrees in Electrical Engineering from National Taiwan University, Taipei, Taiwan, in 1989 and 1995, respectively. In October 1995, he joined the Institute of Information Science, Academia Sinica, Taipei, Taiwan, where he is currently a Research Fellow and the Deputy Director. He also holds a joint appointment as a Professor in the Department of Computer Science and Information Engineering, National Cheng Kung University. His major research interests include spoken language processing, natural language processing, multimedia information retrieval, machine learning, and pattern recognition.

Yih-Ru Wang (王逸如) received the B.S. and M.S. degrees from the Department of Communication Engineering, National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 1982 and 1987, respectively, and the Ph.D. degree from the Institute of Electronic Engineering, NCTU, in 1995. He was an Instructor in the Department of Communication Engineering, NCTU, from 1987 to 1995. In 1995, he became an Associate Professor. His research interests include automatic speech recognition and computational linguistics.

Sin-Horng Chen (陳信宏) received the B.S. degree in Communication Engineering and the M.S. degree in Electronics Engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1976 and 1978, respectively, and the Ph.D. degree in Electrical Engineering from Texas Tech University, Lubbock, in 1983. He became an Associate Professor and a Professor in the Department of Communications Engineering, NCTU, in 1983 and 1990, respectively. He is currently a Professor of ECE Department and Senior Vice President of NCTU. His major research interest is in speech signal processing, especially in Mandarin speech recognition and text-to-speech.