# Voice Conversion Based on Locally Linear Embedding

HSIN-TE HWANG[1,3], YI-CHIAO WU[1], YU-HUAI PENG[1,4], 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, Academia Sinica, Taipei, Taiwan*
[2]*Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan*
[3]*Dept. of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan*
[4]*Dept. of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan*

*E-mail: {hwanght; tedwu; jeremycchsu; whm}@iis.sinica.edu.tw, roland19930601@gmail.com, yu.tsao@citi.sinica.edu.tw, {yrwang; schen}@mail.nctu.edu.tw*

This paper presents a novel locally linear embedding (LLE)-based framework for exemplar-based spectral conversion (SC). The key feature of the proposed SC framework is that it integrates the LLE algorithm, a manifold learning method, with the conventional exemplar-based SC method. One important advantage of the LLE-based SC framework is that it can be applied to either one-to-one SC or many-to-one SC. For one-to-one SC, a parallel speech corpus consisting of the pre-specified source and target speakers' speeches is used to construct the paired source and target dictionaries in advance. During online conversion, the LLE-based SC method converts the source spectral features to the target like spectral features based on the paired dictionaries. On the other hand, when applied to many-to-one SC, our system is capable of converting the voice of any unseen source speaker to that of a desired target speaker, without the requirement of collecting parallel training speech utterances from them beforehand. To further improve the quality of the converted speech, the maximum likelihood parameter generation (MLPG) and global variance (GV) methods are adopted in the proposed SC systems. Experimental results demonstrate that the proposed one-to-one SC system is comparable with the state-of-the-art Gaussian mixture model (GMM)-based one-to-one SC system in terms of speech quality and speaker similarity, and the many-to-one SC system can approximate the performance of the one-to-one SC system.

***Keywords:*** voice conversion, locally linear embedding, exemplar-based, many-to-one, manifold learning.

## 1. INTRODUCTION

Voice conversion (VC) is a technique that converts one type of speech to another, without changing the linguistic content. Many applications based on this technique have been proposed, such as impaired speech to normal speech conversion [1], narrowband speech to wideband speech conversion [2], singing VC [3], and body-transmitted speech enhancement [4]. A typical one is speaker VC [5], which converts a source speaker's speech to a target speaker's speech. Generally speaking, speaker VC involves spectral, prosodic, and excitation conversions. In this study, we focus on spectral conversion (SC), whereas a simple linear transformation of F0 is applied for prosodic conversion.

Numerous SC methods have been proposed during the last two decades. In general, most methods assume that both source and target speech utterances are available in the offline stage. We refer to the SC methods following this assumption as one-to-one SC. In

order to make VC more practically applicable, a more flexible VC framework without such assumption or limitation in one-to-one SC is desirable. To tackle this issue, several flexible frameworks have been proposed, such as many-to-one SC [6-8], one-to-many SC [6, 9], and many-to-many SC [10-12]. Unlike the assumption in one-to-one SC, many-to-one, one-to-many, and many-to-many SCs respectively assume that no training data from source, target, and both are available in the offline stage. In this study, we focus on one-to-one and many-to-one SCs.

Among the one-to-one SC methods, statistical methods have become the mainstream method due to their abilities to effectively model the relationship between the source and target speakers' spectral features [5, 13-25]. Generally speaking, they can be categorized into linear and nonlinear methods. Many linear methods have been proposed, such as the Gaussian mixture model (GMM)-based methods [5, 13-15], partial least squares (PLS) regression [16], and local linear transformation [17]. Although these linear methods are reasonably effective, using linear conversion functions for SC is insufficient to model the complex relationship between the source and target speakers' spectral features. To overcome this limitation, several nonlinear methods have been proposed, such as dynamic kernel PLS [18] and the neural network-based methods [19-25]. Because of the statistical average nature inherent in the statistical methods, the converted spectra may be overly smoothed, thereby causing the converted speech to sound "muffled", which is known as the over-smoothing problem. To overcome this problem, the global variance (GV) [14] and modulation spectrum (MS) [15] methods have been proposed. Although notable improvements have been achieved, the loss of spectral details in the converted spectra may be observed due to the fact that the statistical methods usually conduct SC on the low-dimensional spectral features, such as mel-ceptral coefficients (MCCs) [26] and line spectral pair [27], due to the advantages of high computational efficiency and no curse of dimensionality [5, 13-22].

Some SC methods have been proposed to tackle the speech quality deterioration problem by directly operating on the high-dimensional spectral features, e.g., the spectral envelopes (SEs). For instance, the frequency warping (FW) methods were proposed to shift the source SEs to match the target SEs using a warping function [28-30]. The deep neural network-based methods were proposed to learn the mapping between the source and target SEs [23-25]. The exemplar-based methods were proposed to generate the converted SEs using the weighted linear combination of the target SE exemplars, where the weights were estimated by nonnegative matrix factorization (NMF) [31-33] or a locally linear embedding (LLE) method [34]. In contrast, in [35], the authors combined GMM-based and exemplar-based methods by selecting the target exemplar closest to the converted features as the SC output.

On the other hand, for many-to-one SC, the aim is to convert the voices of any arbitrary source speakers to that of a desired target speaker. In [6], the authors proposed an eigenvoice GMM (EV-GMM) approach, which realized many-to-one and one-to-many VCs by estimating the eigen-vector of an arbitrary speaker from the pre-stored multiple speakers' speech corpora. In [7], the authors used NMF to realize many-to-one VC. Their approach attempted to decompose the source and target speakers' spectra into two categories of information, namely speaker individuality and speaker independent information (e.g., phoneme information), within an NMF-based SC framework. In [8], the authors adopted a deep bidirectional long short term memory (DBLSTM) based recurrent

neural network (RNN) to model the relationship between Phonetic PosteriorGrams (PPGs) and target speaker's spectral features, where PPGs were obtained by applying a speaker-independent automatic speech recognition (SI-ASR) system to target speaker's spectral features. Once built, the DBLSTM was applied to predict the target speaker's spectral features from PPGs obtained by applying the same SI-ASR system to source speaker's spectral features at run-time conversion. The key idea is to bridge speakers by means of PPGs obtained from the SI-ASR system.

In this paper, we present a novel LLE-based framework for exemplar-based SC (called the LLE-exemplar-based SC framework hereafter). One important advantage of the proposed framework is that it can be applied to either one-to-one SC or many-to-one SC. The key idea of the proposed SC systems is to adopt LLE, a classical manifold learning method, to characterize the local geometry of the source spectral features (precisely speaking, the local geometry of the locally linear patches in the source spectral feature space) by using either source speaker's spectral feature vectors/exemplars (for one-to-one SC) or multiple speakers' exemplars (for many-to-one SC). Then, the reconstruction weights of the source speech are determined (with the aim of minimizing the local reconstruction error) and used to generate the converted spectral features. By doing so, we assume that some characteristics of an input (source) natural speech utterance (characterized by LLE) can be preserved in the converted speech during the online conversion stage. To further improve the quality of the converted speech, the maximum likelihood parameter generation (MLPG) [14, 36] and GV [37] methods are adopted in the proposed SC systems.

Note that LLE-based one-to-one SC has been published in a conference paper [34]. In this paper, we present more details, discussions, and evaluations of our LLE-based one-to-one SC system, e.g., investigating different spectral features. We also extend the one-to-one SC system to a many-to-one SC system, which is designed based on the assumption that some of the multiple speakers available in the offline stage share similar acoustic characteristics to an unknown source speaker, and thus the local geometry of the source spectral features can be characterized by LLE with these speakers' dictionaries (composed by their speech corpora). The underlying idea comes from the intuition that there exist some speakers whose voices sound like the voice of the source speaker. After charactering the local geometry by using multiple speakers' dictionaries, conversion is conducted in the same way as one-to-one SC.

The remainder of this paper is organized as follows. Manifold learning, in particular the LLE algorithm, is briefly reviewed in Section 2. The proposed LLE-exemplar-based SC framework for one-to-one and many-to-one SCs are described in detail in Section 3. Experimental setup and results are presented in Section 4. Finally, Section 5 gives the conclusions.

## 2. MANIFOLD LEARNING

Manifold learning is a method for nonlinear dimensionality reduction (DR) [38]. It plays an essential role in the development of various techniques and applications, such as representation learning [39], data visualization [40] and super-resolution [41]. Numerous manifold learning methods have been proposed, such as isometric feature mapping (Iso-

map) [42], Laplacian eigenmap (LE) [43], and LLE [44]. These manifold learning methods compute low-dimensional embeddings of high-dimensional input data by discovering an underlying low-dimensional manifold (the intrinsic geometry of the data distribution) in a high-dimensional data space and embedding them onto a low-dimensional embedding space. The proposed SC framework is based on the LLE algorithm.

The LLE algorithm addresses the problem of nonlinear DR 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 DR 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).

The steps (a), (b), and (c) involve identifying each locally linear patch, charactering the local geometry of each locally linear patch, and preserving the local geometry in the low-dimensional embedding space, respectively.

## 3. THE PROPOSED LLE-EXEMPLAR-BASED SPECTRAL CONVERSION FRAMEWORK

The proposed SC framework can operate on either SEs or MCCs. For convenience, in this section, we refer to the SEs and MCCs as the “spectral features”.

### 3.1 One-To-One Spectral Conversion

Fig. 1 gives an overview of the proposed one-to-one SC system. There are two stages: the offline and online stages. The offline stage mainly involves the construction of the paired dictionaries while the online stage performs SC. In the following, we describe the proposed one-to-one SC system in detail.

**(A) The Offline Stage**

As shown in Fig. 1, the paired source speaker and target speaker dictionaries are constructed in the following steps:

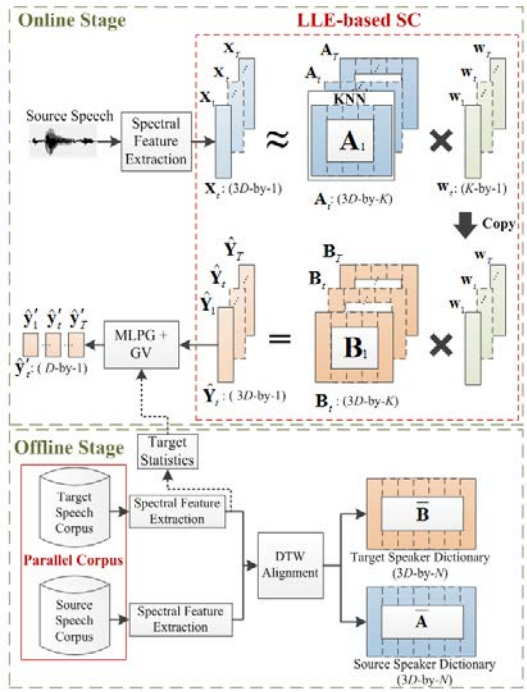



Fig. 1. Overview of the offline and online stages of the proposed one-to-one SC system.

(a) Preparing a parallel speech corpus consisting of the source and target speakers' voices.
(b) Extracting the spectral features (composed of static, delta, and delta-delta features) from the source and target speakers' voices.
(c) Performing dynamic time warping (DTW) to align the spectral feature vector sequence of the source speaker and that of the target speaker to obtain the aligned source and target spectral features.
(d) Constructing the paired dictionaries from the aligned source and target spectral features.

Note that after conducting step (b), some target statistics to be used in the MLPG and GV methods are estimated from the target spectral features, such as the precision matrix and target GV. The MLPG and GV methods will be described later. Besides, after conducting step (c), when multiple source frames are aligned with a certain target frame, or multiple target frames are aligned with a certain source frame, only one source-target frame pair is kept and used to construct the paired dictionaries. This avoids duplicated source frames being selected in the $K$ nearest neighbors selection step during run-time conversion, which might incorrectly characterize the local geometry of the source spectral features. A similar problem exists when the $K$ target exemplars corresponding to the $K$ nearest source neighbors contain duplicated frames. The necessity of this strategy has been confirmed in our preliminary results.

Let $\overline{\mathbf{A}} \in \mathcal{R}^{3D \times N}$ and $\overline{\mathbf{B}} \in \mathcal{R}^{3D \times N}$ (as shown in Fig. 1) be the source speaker and target speaker dictionaries, and be composed of the source and target spectral feature vec-

tors (or called exemplars) as $\overline{\mathbf{A}} = \left[\overline{\mathbf{A}}_1, \cdots, \overline{\mathbf{A}}_n, \cdots, \overline{\mathbf{A}}_N\right]$ and $\overline{\mathbf{B}} = \left[\overline{\mathbf{B}}_1, \cdots, \overline{\mathbf{B}}_n, \cdots, \overline{\mathbf{B}}_N\right]$, respectively, where the numbers of exemplars in both dictionaries are *N*; $\overline{\mathbf{A}}_n \in \mathcal{R}^{3D\times 1}$ is the *n*-th source exemplar in the source speaker dictionary $\overline{\mathbf{A}}$, and is composed of the *D*-dimensional static $\overline{\mathbf{a}}_n \in \mathcal{R}^{D\times 1}$, delta $\Delta^{(1)}\overline{\mathbf{a}}_n \in \mathcal{R}^{D\times 1}$, and delta-delta $\Delta^{(2)}\overline{\mathbf{a}}_n \in \mathcal{R}^{D\times 1}$ features as $\overline{\mathbf{A}}_n = \left[\overline{\mathbf{a}}_n^{\mathrm{T}}, \Delta^{(1)}\overline{\mathbf{a}}_n^{\mathrm{T}}, \Delta^{(2)}\overline{\mathbf{a}}_n^{\mathrm{T}}\right]^{\mathrm{T}}$ (for *n*=1~*N*), where the superscript T denotes transposition of the vector. Likewise, $\overline{\mathbf{B}}_n \in \mathcal{R}^{3D\times 1}$ is the *n*-th target exemplar in the target speaker dictionary $\overline{\mathbf{B}}$, and is composed of the *D*-dimensional static $\overline{\mathbf{b}}_n \in \mathcal{R}^{D\times 1}$, delta $\Delta^{(1)}\overline{\mathbf{b}}_n \in \mathcal{R}^{D\times 1}$, and delta-delta $\Delta^{(2)}\overline{\mathbf{b}}_n \in \mathcal{R}^{D\times 1}$ features as $\overline{\mathbf{B}}_n = \left[\overline{\mathbf{b}}_n^{\mathrm{T}}, \Delta^{(1)}\overline{\mathbf{b}}_n^{\mathrm{T}}, \Delta^{(2)}\overline{\mathbf{b}}_n^{\mathrm{T}}\right]^{\mathrm{T}}$ (for *n*=1~*N*).

**(B) The Online Stage**

From Fig. 1, given a source speech for conversion, spectral feature extraction is performed to extract a sequence of the source spectral feature vectors $\left\{\mathbf{X}_t \in \mathcal{R}^{3D\times 1}\right\}_{t=1}^{T}$, where *T* is the number of speech frames of the source speech; $\mathbf{X}_t$ is the source spectral feature vector at frame *t*, and is composed of the *D*-dimensional static $\mathbf{x}_t \in \mathcal{R}^{D\times 1}$, delta $\Delta^{(1)}\mathbf{x}_t \in \mathcal{R}^{D\times 1}$, and delta-delta $\Delta^{(2)}\mathbf{x}_t \in \mathcal{R}^{D\times 1}$ features as $\mathbf{X}_t = \left[\mathbf{x}_t^{T}, \Delta^{(1)}\mathbf{x}_t^{T}, \Delta^{(2)}\mathbf{x}_t^{T}\right]^{\mathrm{T}}$ (for *t*=1~*T*). Then, the LLE-based SC method is applied to convert the source spectral feature vectors $\left\{\mathbf{X}_t\right\}_{t=1}^{T}$ (independently in a frame-by-frame manner) to obtain a sequence of the converted spectral feature vectors $\left\{\hat{\mathbf{Y}}_t \in \mathcal{R}^{3D\times 1}\right\}_{t=1}^{T}$, where $\hat{\mathbf{Y}}_t$ is the converted spectral feature vector at frame *t*. In order to further improve the quality of the converted speech, the MLPG and GV methods (denoted as "MLPG+GV" in Fig. 1) are applied to the converted spectral feature vectors $\left\{\hat{\mathbf{Y}}_t\right\}_{t=1}^{T}$ to generate a final sequence of converted static spectral feature vectors $\left\{\hat{\mathbf{y}}'_t \in \mathcal{R}^{D\times 1}\right\}_{t=1}^{T}$, where $\hat{\mathbf{y}}'_t$ is the final converted static spectral feature vector at frame *t*. Next, we describe the LLE-based SC, MLPG, and GV methods in detail.

**(C) The LLE-Based SC Method**

The LLE-based SC method for an arbitrary input source spectral feature vector (say $\mathbf{X}_t$ for example) has three steps:

(a) Finding *K* nearest neighbors (measured by the Euclidean distance) of $\mathbf{X}_t$ from the source speaker dictionary.
(b) Computing the reconstruction weight vector that best (linearly) reconstructs $\mathbf{X}_t$ from its *K* nearest neighbors found in step (a).
(c) Estimating the target spectral feature vector (at frame *t*) by linearly combining *K* target exemplars (paired/aligned with the *K* nearest neighbors of $\mathbf{X}_t$) in the target speaker dictionary with the reconstruction weight vector obtained in step (b).

The steps (a) and (b) involve identifying the locally linear patch and charactering the local geometry of the locally linear patch, respectively, as described in steps (a) and

(b) of the LLE algorithm for DR. On the other hand, the step (c) involves estimating the target spectral feature vector by preserving the local geometry of the source spectral features, as opposed to estimating the low-dimensional embedding in step (c) of the LLE algorithm for DR.

Specifically, we implement steps (b) and (c) as follows. In step (b), the reconstruction weight vector is computed by minimizing the reconstruction error $\varepsilon_t$ subject to the constraint $\mathbf{1}^{\mathrm{T}}\mathbf{w}_t = 1$ (for the purpose of translational invariance) at frame *t*:

$$\varepsilon_t = \left\| \mathbf{X}_t - \mathbf{A}_t \mathbf{w}_t \right\|^2, \text{ s.t. } \mathbf{1}^{\mathrm{T}}\mathbf{w}_t = 1, \tag{1}$$

where $\mathbf{A}_t \in \mathcal{R}^{3D\times K}$ is a matrix (a subset of the source speaker dictionary) composed of *K* nearest neighbors of $\mathbf{X}_t$, i.e., $\mathbf{A}_t = \left[\mathbf{a}_{t,1},\cdots,\mathbf{a}_{t,k},\cdots,\mathbf{a}_{t,K}\right]$, where $\mathbf{a}_{t,k} \in \mathcal{R}^{3D\times 1}$ is the *k*-th nearest neighbor of $\mathbf{X}_t$; $\mathbf{w}_t \in \mathcal{R}^{K\times 1}$ is the reconstruction weight vector at frame *t*; and $\mathbf{1} \in \mathcal{R}^{K\times 1}$ is a vector whose elements are all ones. Note that $\mathbf{A}_t$ can be obtained in step (a). Solving $\mathbf{w}_t$ by minimizing $\varepsilon_t$ subject to the constraint is a constrained least square problem, and the closed-form solution can be found in [34, 45]. A more efficient way to obtain $\mathbf{w}_t$ is to solve the linear system of equations in advance:

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

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

$$\mathbf{G}_t = \left(\mathbf{A}_t - \mathbf{X}_t\mathbf{1}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\mathbf{A}_t - \mathbf{X}_t\mathbf{1}^{\mathrm{T}}\right). \tag{3}$$

Then, the reconstruction weight vector is rescaled to satisfy the constraint $\mathbf{1}^{\mathrm{T}}\mathbf{w}_t = 1$. The detailed derivations of the solution can be found in [45].

In step (c), with the assumption that the source and target spectral feature vectors share a similar local geometry in their respective spectral feature spaces (manifolds), the converted spectral feature vector $\hat{\mathbf{Y}}_t$ at frame *t* can be obtained by

$$\hat{\mathbf{Y}}_t = \mathbf{B}_t \mathbf{w}_t, \tag{4}$$

where the reconstruction weight vector $\mathbf{w}_t$ is obtained in step (b); $\mathbf{B}_t \in \mathcal{R}^{3D\times K}$ is a matrix (a subset of the target speaker dictionary) corresponding to $\mathbf{A}_t$, and is composed of *K* target exemplars, i.e., $\mathbf{B}_t = \left[\mathbf{b}_{t,1},\cdots,\mathbf{b}_{t,k},\cdots,\mathbf{b}_{t,K}\right]$, where $\mathbf{b}_{t,k} \in \mathcal{R}^{3D\times 1}$ is the *k*-th target exemplar in $\mathbf{B}_t$ corresponding to (paired/aligned with) $\mathbf{a}_{t,k}$.

Once the frame-by-frame conversion process is finished, a sequence of converted spectral feature vectors $\left\{\hat{\mathbf{Y}}_t\right\}_{t=1}^{T}$ can be obtained.

**(D) The MLPG and GV Methods**

Since the LLE-based SC method is performed in a frame-by-frame manner, the dis-

continuity problem, which is often encountered in frame-based SC systems, exists. Besides, it also has the over-smoothing effect, which is often observed in the statistical SC methods. In this study, we adopt the MLPG and GV methods in the LLE-based SC framework to handle the discontinuity and over-smoothing problems, respectively.

*1) The MLPG Method:* The MLPG method [14, 36] applied to the proposed method is given as

$$\hat{\mathbf{y}} = (\mathbf{M}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\boldsymbol{\Lambda}\hat{\mathbf{Y}}, \tag{5}$$

where $\hat{\mathbf{y}} = \left[\hat{\mathbf{y}}_1^{\mathrm{T}}, \cdots, \hat{\mathbf{y}}_t^{\mathrm{T}}, \cdots, \hat{\mathbf{y}}_T^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathcal{R}^{DT\times 1}$ is the sequence of the converted static spectral feature vectors; $\hat{\mathbf{y}}_t \in \mathcal{R}^{D\times 1}$ is the converted static feature vector at frame *t*; $\mathbf{M} \in \mathcal{R}^{3DT\times DT}$ is a weighting matrix (given by [14, 36]) used for appending the dynamic features to the static ones; $\hat{\mathbf{Y}} = [\hat{\mathbf{Y}}_1^{\mathrm{T}}, \cdots, \hat{\mathbf{Y}}_t^{\mathrm{T}}, \cdots, \hat{\mathbf{Y}}_T^{\mathrm{T}}]^{\mathrm{T}} \in \mathcal{R}^{3DT\times 1}$ is the converted spectral feature vector sequence obtained by Eq. (4); $\boldsymbol{\Lambda} = \mathrm{diag}[\boldsymbol{\Lambda}_1^{(\mathbf{Y})}, \cdots, \boldsymbol{\Lambda}_t^{(\mathbf{Y})}, \cdots, \boldsymbol{\Lambda}_T^{(\mathbf{Y})}] \in \mathcal{R}^{3DT\times 3DT}$ is the global precision matrix, where $\boldsymbol{\Lambda}_t^{(\mathbf{Y})} \in \mathcal{R}^{3D\times 3D}$ is the precision matrix (at frame *t*) estimated from the target speaker's training data (target spectral feature vectors), which is assumed to be diagonal. Note that $\boldsymbol{\Lambda}_1^{(\mathbf{Y})} = \cdots = \boldsymbol{\Lambda}_t^{(\mathbf{Y})} = \cdots = \boldsymbol{\Lambda}_T^{(\mathbf{Y})}$.

*2) The GV Method:* The converted static spectral feature vector sequence $\hat{\mathbf{y}}$ obtained by Eq. (5) is further processed by the postfiltering-based GV compensation method [37] as

$$\hat{y}_t'(d) = \sqrt{\frac{\mu_v(d)}{\mathrm{var}(d)}}\left(\hat{y}_t(d) - \left\langle \hat{y}(d)\right\rangle\right) + \left\langle \hat{y}(d)\right\rangle, \tag{6}$$

where the index *d*=1~*D*; $\hat{y}_t'(d)$ is the *d*-th element of the final converted feature vector $\hat{\mathbf{y}}_t'$; $\hat{y}_t(d)$ is the *d*-th element of $\hat{\mathbf{y}}_t$ obtained by Eq. (5); $\mu_v(d)$ is the *d*-th element of the mean vector of the target GV, which is obtained using the GVs of the target feature vector sequences calculated from individual utterances in the training data as described in [14]; $\left\langle \hat{y}(d)\right\rangle$ and $\mathrm{var}(d)$ are the mean and variance of the *d*-th component of the converted static spectral feature vector, and can be calculated as

$$\left\langle \hat{y}(d)\right\rangle = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t(d), \tag{7}$$

$$\mathrm{var}(d) = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}_t(d) - \left\langle \hat{y}(d)\right\rangle\right)^2. \tag{8}$$

### 3.2 Many-To-One Spectral Conversion

One important advantage of the proposed LLE-exemplar-based SC framework is that it can be readily extended from one-to-one SC to many-to-one SC simply by replacing the paired source speaker and target speaker dictionaries in the one-to-one SC system to the paired global source speaker and target speaker dictionaries without modifying the

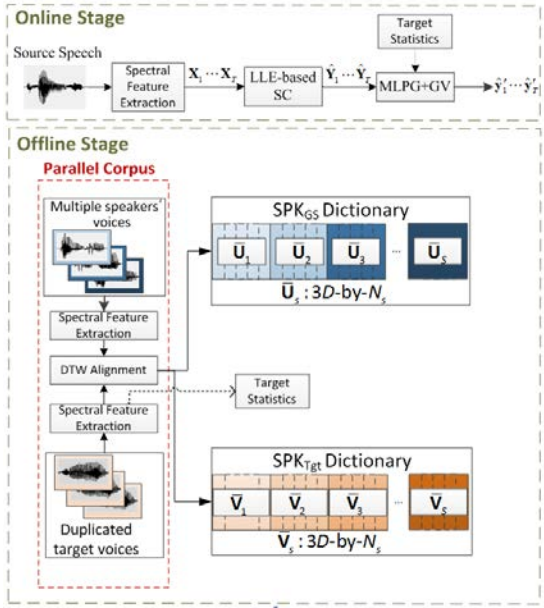



Fig. 2. Overview of the offline and online stages of the proposed many-to-one SC system.

kernel methods in the online stage of the one-to-one SC system. Specifically, the global source speaker dictionary is constructed by using multiple speakers' speech corpora. As shown in Fig. 2, like its one-to-one counterpart, the many-to-one SC system also contains two stages: the offline and online stages.

**(A) The Offline Stage**

As shown in Fig. 2, the offline stage mainly involves the construction of the paired global source speaker ($SPK_{GS}$ Dictionary) and target speaker dictionaries ($SPK_{Tgt}$ Dictionary) in the following steps:

(a) Preparing a parallel speech corpus consisting of multiple speakers' voices and the desired target speaker's voices in advance. Then, each of the multiple speakers and the target speaker are formed a pair in turn (called a "known source"-target pair).

(b) Constructing the paired dictionaries ( $\bar{\mathbf{U}}_s$ and $\bar{\mathbf{V}}_s$ ) for each "known source"-target pair in the same way as constructing the paired source speaker and target speaker dictionaries (i.e., $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ ) in the one-to-one SC system.

(c) Constructing the paired $SPK_{GS}$ and $SPK_{Tgt}$ dictionaries by combining all the paired dictionaries obtained in step (b).

Like the one-to-one SC system, the target statistics to be used in the MLPG and GV methods are estimated, and the duplicated frames are not included in the paired dictionaries.

Let the $SPK_{GS}$ and $SPK_{Tgt}$ dictionaries be denoted as $\{\bar{\mathbf{U}}_s\}_{s=1}^{S}$ and $\{\bar{\mathbf{V}}_s\}_{s=1}^{S}$ (as shown in Fig. 2), respectively, where both $SPK_{GS}$ and $SPK_{Tgt}$ dictionaries contain $S$ component dictionaries; $\bar{\mathbf{U}}_s \in \mathcal{R}^{3D\times N_s}$ is the $s$-th known source speaker's dictionary in the $SPK_{GS}$

dictionary, and is composed by the *s*-th known source speaker's spectral feature vectors (exemplars), i.e., $\bar{\mathbf{U}}_s = \left[\bar{\mathbf{U}}_{s,1}, \cdots, \bar{\mathbf{U}}_{s,n}, \cdots, \bar{\mathbf{U}}_{s,N_s}\right]$; $\bar{\mathbf{U}}_{s,n} \in \mathcal{R}^{3D\times 1}$ is the *n*-th exemplar in $\bar{\mathbf{U}}_s$, and is composed by the *D*-dimensional static $\bar{\mathbf{u}}_{s,n} \in \mathcal{R}^{D\times 1}$, delta $\Delta^{(1)}\bar{\mathbf{u}}_{s,n} \in \mathcal{R}^{D\times 1}$, and delta-delta $\Delta^{(2)}\bar{\mathbf{u}}_{s,n} \in \mathcal{R}^{D\times 1}$ features as $\bar{\mathbf{U}}_{s,n} = \left[\bar{\mathbf{u}}_{s,n}^{\mathrm{T}}, \Delta^{(1)}\bar{\mathbf{u}}_{s,n}^{\mathrm{T}}, \Delta^{(2)}\bar{\mathbf{u}}_{s,n}^{\mathrm{T}}\right]^{\mathrm{T}}$, for *s*=1~*S* and *n*=1~$N_s$. Likewise, $\bar{\mathbf{V}}_s \in \mathcal{R}^{3D\times N_s}$ is the target speaker dictionary corresponding to $\bar{\mathbf{U}}_s$, and is composed by the target exemplars $\bar{\mathbf{V}}_s = \left[\bar{\mathbf{V}}_{s,1}, \cdots, \bar{\mathbf{V}}_{s,n}, \cdots, \bar{\mathbf{V}}_{s,N_s}\right]$ that are aligned with $\bar{\mathbf{U}}_s = \left[\bar{\mathbf{U}}_{s,1}, \cdots, \bar{\mathbf{U}}_{s,n}, \cdots, \bar{\mathbf{U}}_{s,N_s}\right]$. $\bar{\mathbf{V}}_{s,n} \in \mathcal{R}^{3D\times 1}$ is the *n*-th target exemplar in $\bar{\mathbf{V}}_s$ corresponding to $\bar{\mathbf{U}}_{s,n}$, and is composed by the *D*-dimensional static $\bar{\mathbf{v}}_{s,n} \in \mathcal{R}^{D\times 1}$, delta $\Delta^{(1)}\bar{\mathbf{v}}_{s,n} \in \mathcal{R}^{D\times 1}$, and delta-delta $\Delta^{(2)}\bar{\mathbf{v}}_{s,n} \in \mathcal{R}^{D\times 1}$ features as $\bar{\mathbf{V}}_{s,n} = \left[\bar{\mathbf{v}}_{s,n}^{\mathrm{T}}, \Delta^{(1)}\bar{\mathbf{v}}_{s,n}^{\mathrm{T}}, \Delta^{(2)}\bar{\mathbf{v}}_{s,n}^{\mathrm{T}}\right]^{\mathrm{T}}$, for *s*=1~*S* and *n*=1~$N_s$. The number of exemplars in $\bar{\mathbf{U}}_s$ and $\bar{\mathbf{V}}_s$, $s = 1 \sim S$, is $N_s$.

**(B) The Online Stage**

The many-to-one SC system performs SC in the same way as the one-to-one SC system. Given a source speech for conversion, spectral feature extraction is performed to extract a sequence of the source spectral feature vectors $\{\mathbf{X}_t\}_{t=1}^{T}$. Then, the LLE-based SC method followed by the MLPG and GV methods is applied to convert $\{\mathbf{X}_t\}_{t=1}^{T}$ to obtain the final sequence of the converted static spectral feature vectors $\{\hat{\mathbf{y}}'_t\}_{t=1}^{T}$. Note that the input source speech can be from any arbitrary unseen speaker, and the paired $\mathrm{SPK}_{\mathrm{GS}}$ and $\mathrm{SPK}_{\mathrm{Tgt}}$ dictionaries are used in the LLE-based SC method.

## 4. EXPERIMENTS

We conducted two sets of experiments to evaluate the effectiveness of the proposed LLE-exemplar-based SC framework on one-to-one and many-to-one SCs, respectively. We first describe the experimental setup in Section 4.1, and then present the evaluations of one-to-one and many-to-one SCs in Sections 4.2 and 4.3, respectively.

### 4.1 Experimental Setup

**(A) Speech Corpus:**

Two speech corpora were used in the experiments: the Sinica COSPRO corpus [46] and the Voice Conversion Challenge 2016 (VCC2016) corpus [47].

In the experiments of the one-to-one SC task (Section 4.2), the Sinica COSPRO corpus was adopted. The corpus contained 9 datasets. The intonation-balanced dataset (i.e., COSPRO 03), consisting of Mandarin parallel speech utterances of 3 females and 2 males, was used in the experiments. There were 20 pairs of conversions: 8 intra-gender and 12 inter-gender. For each conversion pair, 10 utterance pairs were randomly selected as the training set, 40 utterance pairs as the development set, and 43 utterance pairs as the test set. Speech signals were recorded in a 16 kHz/16 bit format. Silence segments at the start and end of each utterance in the training set were discarded based on the segmenta-

tion information in the corpus.

In the experiments of the many-to-one SC task (Section 4.3), the VCC2016 corpus was adopted. The corpus consisted of English parallel speech utterances of 5 females and 5 males. Officially, the dataset was divided into training and test sets. The training set comprised 5 source speakers (SF1, SF2, SF3, SM1, and SM2) and 5 target speakers (TF1, TF2, TM1, TM2, and TM3), and each speaker had 162 utterances. The test set comprised the same 5 source and 5 target speakers, and each speaker had 54 utterances. Among these ten speakers, {SF1, SF2, SF3, TF1, TF2} are female, and {SM1, SM2, TM1, TM2, TM3} are male. Speech signals were recorded in a 16 kHz/16 bit format. We conducted objective and subjective evaluations on four pairs of inter-gender conversions, including SF1→TM1, SF2→TM1, SM1→TF1, and SM2→TF1. For each conversion pair, 54 utterance pairs in the official test set were evaluated.

**(B) Analysis/Conversion/Synthesis:**

We used the STRAIGHT vocoder [48] for feature extraction and waveform generation. During feature extraction, the speech signals were parametrized into the smoothed spectral envelopes (SEs), aperiodicity components (APs), and F0 contours, where the SE, AP, and F0 described the spectral, excitation, and prosodic features, respectively. The FFT length was set to 1024; thus, the AP and SE for each frame consisted of 513 components. The frame shift was 5 milliseconds. The SC systems operated on either MCCs (extracted from SEs) or SEs to obtain the converted SEs (which will be described later in detail).

For all of the compared systems described in the following experiments, we only performed F0 conversion while remaining the other prosodic features (i.e., the source speech's energy and duration) and the excitation features (i.e., APs) unmodified. Specifically, the F0 was converted by the linear mean-variance transformation method as follows:

$$\hat{f}_t^{(y)} = \frac{\sigma^{(y)}}{\sigma^{(x)}}\left( f_t^{(x)} - \mu^{(x)} \right) + \mu^{(y)}, \tag{9}$$

where $f_t^{(x)}$ and $\hat{f}_t^{(y)}$ are a one dimensional log-scaled F0 of the source speech and the converted speech at frame $t$; $\mu^{(x)}$ and $\sigma^{(x)}$ are the mean and standard deviation of log-scaled F0 calculated from the training data of the source speaker; and $\mu^{(y)}$ and $\sigma^{(y)}$ are those of log-scaled F0 of the target speaker calculated from the training data of the target speaker.

Finally, the converted SEs, converted F0, and source speech's APs were passed to the STRAIGHT vocoder for waveform reconstruction.

### 4.2 Evaluation of One-To-One Spectral Conversion

**(A) Reference Systems**

First, we intended to determine suitable spectral features for the proposed LLE-based SC system, and thus tested the performance of the system (in terms of speech quality and speaker similarity) using two types of features, namely MCCs and SEs. Next, we intended to investigate whether the GV method can be compatible with the proposed

SC system, and thus tested the performance of two LLE-based SC systems, one with GV and one without GV. Note that the effectiveness of combining the MLPG method with the LLE-based SC system has been confirmed in our previous work [34]. Thus, the MLPG method was used as the default process in all of the LLE-exemplar-based SC systems in the following experiments. Four SC systems were built for comparison:

- ***GMM**:* The state-of-the-art GMM-based SC system integrated with both MLPG and GV methods [14], where MCCs were used as the spectral features.
- ***LLE**$_{MCC}$:* The LLE-based one-to-one SC system (with MLPG and without GV) as described in Section 3.1, where MCCs were used as the spectral features.
- ***LLE**$_{MCC}$**-GV**:* The LLE-based one-to-one SC system (with both MLPG and GV) as described in Section 3.1, where MCCs were used as the spectral features.
- ***LLE**$_{SE}$**-GV**:* The LLE-based one-to-one SC system (with both MLPG and GV) as described in Section 3.1, where SEs were used as the spectral features.

For the baseline ***GMM*** system, the number of mixture components was 64, according to the objective scores and informal listening tests conducted on the development set. A cross-diagonal covariance matrix was used in the joint density GMM (JDGMM). The spectral features were the first through 24-th MCCs extracted from the STAIGHT SEs. The static, delta, and delta-delta features were used. Accordingly, the dimensionality of a final MCC vector was 72. Note that in the waveform reconstruction step, the converted MCCs obtained by the ***GMM*** system were reverted back to obtain the converted SEs in advance. Then, the waveform can be generated as described in Section 4.1 (B).

Both ***LLE**$_{MCC}$* and ***LLE**$_{MCC}$**-GV*** adopted the same spectral features as the ***GMM*** system. The number of nearest neighbors (i.e., $K$ in Eqs. (1)-(4)) was determined according to the objective scores, computational complexity, and informal listening tests conducted on the development set (which will be described later). The same as the ***GMM*** system, after SC, the converted MCCs obtained by the ***LLE*** system were first reverted back to obtain the converted SEs. Then, the waveform was reconstructed from the SEs.

For ***LLE**$_{SE}$**-GV***, the spectral features were the 513-dimensional log energy-normalized SEs. Specifically, each frame of SEs was normalized to unit-sum, and the energy normalizing factor was taken out as an independent feature and was not modified. Then, a logarithm was applied to each energy-normalized SE value. Moreover, the static, delta, and delta-delta features were used. Accordingly, the dimensionality of a final log energy-normalized SE vector was 1539. After SC, the converted log energy-normalized SEs were reverted back to the (linear) SEs, and the energy was compensated back to the SEs according to the energy normalizing factor. Finally, the waveform was generated as described in Section 4.1 (B).

The dictionaries of the LLE systems (i.e., ***LLE**$_{MCC}$*, ***LLE**$_{MCC}$**-GV***, and ***LLE**$_{SE}$**-GV***) contained about 12,300 to 13,700 exemplars for different conversion pairs.

**(B) Objective Evaluations**

In the objective tests, we evaluated the SC systems on the test and development sets in terms of the spectral distortion and the degree of over-smoothing of the converted MCCs.

*1) Spectral Distortion:* We measured the spectral distortion in terms of mel-cepstral distortion (MCD). The MCD value of a target-converted frame pair at frame $t$ (for all $t$=1~$T$) was computed as the distortion between a pair of reference target and the con-

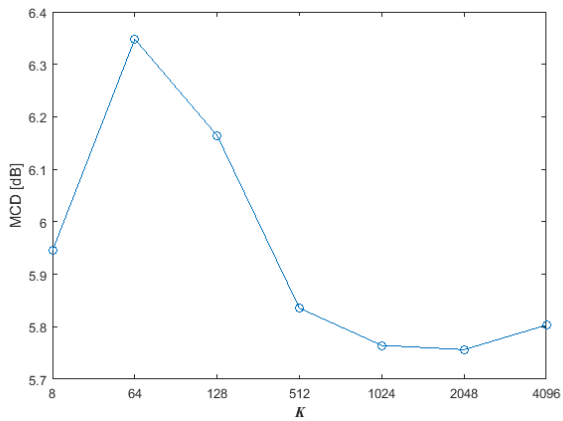



Fig. 3. The average MCD values with different numbers of nearest neighbors on the development set obtained by the ***LLE$_{MCC}$-GV*** system.

**Table 1. The average MCD values [dB] of four SC systems. The MCD value before conversion is 7.19 dB.**

| ***GMM*** | ***LLE$_{SE}$-GV*** | ***LLE$_{MCC}$-GV*** | ***LLE$_{MCC}$*** |
|---|---|---|---|
| 5.58 | 5.31 | 5.59 | **5.09** |

verted mel-cepstra as follows:

$$D_{MCD}(y_t^{(MCC)}, \hat{y}_t^{(MCC)})\ [\mathrm{dB}] = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}(y_t^{(MCC)}(d) - \hat{y}_t^{(MCC)}(d))^2}\ , \tag{10}$$

where $y_t^{(MCC)}(d)$ and $\hat{y}_t^{(MCC)}(d)$ are the $d$-th coefficients of the (reference) target and converted MCCs at frame $t$, respectively. The MCD value of an utterance pair was obtained by averaging over the MCD values of all the frame pairs in the utterance. We reported the average MCD value of all the test and development utterance pairs. A lower MCD value indicates less spectral distortion.

First, we investigated the effect of the number of nearest neighbors (i.e., $K$ in Eqs. (1)-(4)) on the proposed LLE-based SC system. We calculated the (average) MCD values with different numbers of nearest neighbors on the development set. The results of ***LLE$_{MCC}$-GV*** are shown in Fig. 3, where the maximum value of $K$ was set to 4096 because the computational cost of LLE-based SC with $K$=4096 was considerably high for real-world applications. From Fig. 3, we observe that ***LLE$_{MCC}$-GV*** achieved the lowest MCD values when $K$ was around 1024 and 2048. Further analyses demonstrated no significant differences when $K$ was larger than 1024 in terms of listening tests. Similar trends were also found when analyzing the effect of $K$ on the ***LLE$_{MCC}$*** and ***LLE$_{SE}$-GV*** systems. Based on the above analysis, the number of nearest neighbors $K$ for the LLE-based SC systems was set to 1024 in the following experiments. Next, we compared the four SC systems in terms of spectral distortion.

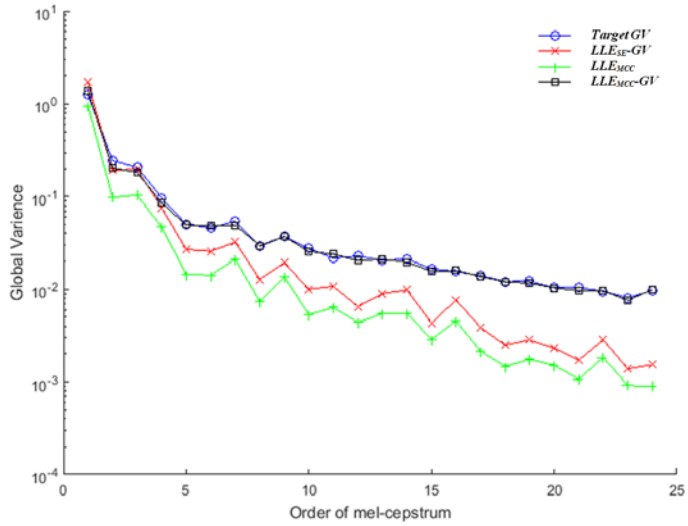



Fig. 4. The average GV measurements of the converted MCCs given by the three proposed SC systems (i.e., ***LLE$_{MCC}$-GV***, ***LLE$_{SE}$-GV***, and ***LLE$_{MCC}$***) and the natural MCCs of the target speech (i.e., ***Target GV***).

Table 1 shows the average MCD values of all the test utterance pairs obtained by the four SC systems. First, comparing the proposed SC systems (operating on MCCs) with and without the GV method, we observe that ***LLE$_{MCC}$*** gave notably lower MCD values than ***LLE$_{MCC}$-GV***, indicating that GV tends to distort the converted speech. Second, comparing ***LLE$_{MCC}$-GV*** with ***LLE$_{SE}$-GV***, we note that ***LLE$_{SE}$-GV*** gave lower MCD values than ***LLE$_{MCC}$-GV***, implying that adopting SE tends to be able to yield better VC performance. Third, comparing the proposed SC systems with the GMM-based SC system, we note that both ***LLE$_{MCC}$*** and ***LLE$_{SE}$-GV*** gave lower MCD values than ***GMM,*** and ***LLE$_{MCC}$-GV*** yielded similar MCD values as ***GMM***. The results indicate that the proposed systems either outperform or perform as well as ***GMM*** in terms of MCD values. However, previous studies have shown that MCD may not perfectly reflect the real subjective evaluation results, particularly when GV is considered in spectral feature generation [14, 22]. Similar results were found in our subjective test (which will be shown later). Thus, we further conducted another objective test to evaluate the SC systems.

*2) Degree of Over-Smoothing:* We adopted the GV measurement to evaluate the degree of over-smoothing of the converted MCCs. The GV measurement of an utterance was computed by Eqs. (7) and (8). We reported the average GV measurement over all the test utterances. Fig. 4 shows the average GV measurements of the converted MCCs given by the three LLE-based SC systems and the natural MCCs of the target speech (referred to as ***Target GV*** hereafter) evaluated on the test set. From Fig. 4, it is clear that the average GV measurement of ***LLE$_{MCC}$-GV*** is larger than those of ***LLE$_{MCC}$*** and ***LLE$_{SE}$-GV***, in particular for the higher order MCCs. Moreover, the average GV measurement of ***LLE$_{MCC}$-GV*** is close to ***Target GV***. These results imply that the converted MCCs given by ***LLE$_{MCC}$*** and ***LLE$_{SE}$-GV*** were overly smoothed, and ***LLE$_{MCC}$-GV*** could effectively

**Table 2. Preference test results (%) of speech quality. *p* is the *p*-value given by the *t*-test for examining the significance of performance difference between the compared systems.**

| ***LLE$_{SE}$-GV*** | ***LLE$_{MCC}$-GV*** | *p* |
|---|---|---|
| 18.12 | **81.88** | 0.000 |

overcome the over-smoothing problem. Besides, the comparison results of ***LLE$_{MCC}$-GV*** and ***LLE$_{MCC}$*** confirm the effectiveness of employing the GV method in the LLE-based SC system operating on MCCs. The comparison results of ***LLE$_{MCC}$-GV*** and ***LLE$_{SE}$-GV*** suggest that for the LLE-based SC system, the GV method can help overcome the over-smoothing issue more effectively when operating on MCCs than SEs. Previous research has shown that the GV measurement of a GMM-based SC system integrated with the GV method can be very close to ***Target GV*** [14]. This was also observed in our baseline GMM-based SC system.

**(C) Subjective Evaluations**

For the subjective evaluation, we randomly selected two conversion pairs from each category (including f-f, m-m, m-f, and f-m; m: male, f: female), resulting in eight conversion pairs. For each conversion pair, eight sentences were randomly selected from the test set, thereby resulting in 64 (8x8) test sentences. Ten listeners were recruited to conduct the speech quality and speaker similarity tests.

*1) Speech Quality:* In the speech quality test, we first compared ***LLE$_{MCC}$-GV*** with ***LLE$_{SE}$-GV*** in order to find out which spectral feature is more suitable for the proposed LLE-based SC system. We conducted a preference test to evaluate the quality of the converted speeches obtained by ***LLE$_{MCC}$-GV*** and ***LLE$_{SE}$-GV***, respectively. Specifically, in the preference test, each pair of converted speeches by systems **A** and **B** were presented in a random order to the listeners. The listeners were asked to judge which sample sounded more natural.

Table 2 shows the average results of the preference test. We can see that ***LLE$_{MCC}$-GV*** yielded remarkable gains over ***LLE$_{SE}$-GV***. According to the responses of the listeners, the converted speech by ***LLE$_{MCC}$-GV*** obviously sounds clearer and brighter than the converted speech by ***LLE$_{SE}$-GV***, and the latter still sounds muffled. A possible reason is that performing the GV method on MCCs is more effective than performing the GV method on SEs, the same as the objective test for comparing ***LLE$_{MCC}$-GV*** with ***LLE$_{SE}$-GV*** in Fig. 4. A similar result has been reported in [23] that the GMM-based system with the GV method worked better on MCCs than SEs. Note that the result is not consistent with that of the objective test in Table 1, where ***LLE$_{MCC}$-GV*** achieved a higher MCD value than ***LLE$_{SE}$-GV*** (5.59dB vs. 5.31dB). This implies that MCD may not perfectly reflect human auditory perception. Similar results have been reported in [14, 22].

Next, we conducted the mean opinion score (MOS) test to evaluate the quality of the converted speeches by ***LLE$_{MCC}$***, ***LLE$_{MCC}$-GV***, and ***GMM***, along with the target sample (denoted as ***Target***) for comparison purpose. During the MOS test, the converted speech samples (obtained by the three SC systems) for each test sentence were presented to a subject in a random order. To evaluate the speech quality, the subjects were request-

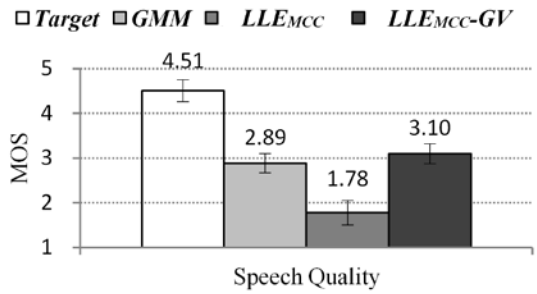



Fig. 5. Subjective test results of the speech converted by ***GMM***, ***LLE$_{MCC}$*** and ***LLE$_{MCC}$-GV***. Error bars indicate 95% confidence intervals. "***Target***" denotes the analysis-synthesized target speech.

**Table 3. Comparison of speaker identification rates (%) of different systems. *p* is the *p*-value given by the one-way ANOVA test for examining the significance of performance difference between the compared systems.**

| ***GMM*** | ***LLE$_{SE}$-GV*** | ***LLE$_{MCC}$-GV*** | *p* |
|---|---|---|---|
| 86.88 | 87.81 | 86.25 | 0.969 |

ed to give a score for each test sentence. The MOS score ranges from 1 (bad) to 5 (excellent), with a higher score denoting better speech quality.

Fig. 5 shows the overall average MOS results. From the figure, we first observe that ***LLE$_{MCC}$-GV*** notably outperforms ***LLE$_{MCC}$***. The result confirms that introducing the GV method to the LLE-based SC system can notably improve the quality of the converted speech. In general, the result is consistent with the result of the objective test in Fig. 4, where ***LLE$_{MCC}$-GV*** achieved GV measurements close to ***Target GV***. However, the result is not consistent with the result of the objective test in Table 1, where ***LLE$_{MCC}$-GV*** achieved a higher MCD value than ***LLE$_{MCC}$*** (5.59dB vs. 5.09dB), again showing that MCD may not perfectly reflect human auditory perception, particularly when the GV method is involved in the spectral feature generation process [14, 22]. Although ***LLE$_{MCC}$-GV*** achieved a higher MOS score (3.10) than ***GMM*** (2.89), the *p*-value of the *t*-test was 0.22, which indicated that the performance difference between ***GMM*** and ***LLE$_{MCC}$-GV*** was not significant. The result is consistent with that of the objective test in Table 1, where ***LLE$_{MCC}$-GV*** achieved a slightly higher MCD value than ***GMM*** (5.59dB vs. 5.58dB).

*2) Speaker Similarity:* In the speaker similarity test, we compared the ***GMM***, ***LLE$_{MCC}$-GV***, and ***LLE$_{SE}$-GV*** systems. The ABX test was adopted. The natural source and target speeches were presented to the listener in a random order as A and B, and the corresponding converted speech was presented as X. To prevent the listener from evaluating only a specific prosodic pattern of each utterance, the same sentence was used for A and B, and a different one was used for X [14]. Listeners were asked to judge whether the utterance X sounded more like utterance A or B. Note that we only reported the results of intra-gender conversion since all the inter-gender conversion pairs were identified correctly in our preliminary result. Similar results have also been reported in [18, 33].

Table 3 shows the overall average results of the ABX test. There are no significant differences among three systems according to the one-way ANOVA test (i.e., *p*-value is greater than 0.05).

From the results of the objective and subjective tests, we can confirm that our best LLE-based SC system (i.e., the ***LLE$_{MCC}$-GV*** system) performs as well as the state-of-the-art GMM-based SC system (i.e., the ***GMM*** system).

### 4.3 Evaluation of Many-To-One Spectral Conversion

#### (A) Reference Systems

In order to show the effectiveness of applying the proposed LLE-based framework for many-to-one SC, we compared the LLE-based many-to-one SC system (denoted as ***MTO-LLE***) with the best LLE-based one-to-one SC system (i.e., ***LLE$_{MCC}$-GV***) in the following experiments. For both ***LLE$_{MCC}$-GV*** and ***MTO-LLE*** systems, the number of nearest neighbors $K$ was set to 1024.

For the ***LLE$_{MCC}$-GV*** system, we randomly selected 10 utterance pairs from the 162 utterance pairs in the official training set to construct the paired dictionaries for each conversion pair. Other settings were the same as described in Sections 4.1 and 4.2. The dictionaries of the ***LLE$_{MCC}$-GV*** system contained about 4,700 to 6,500 exemplars for different conversion pairs.

For the ***MTO-LLE*** system, the $\text{SPK}_{\text{GS}}$ dictionary was constructed using the voices of speakers of the same gender. Thus, the built dictionary $\text{SPK}_{\text{GS}}$ is a gender dependent dictionary. In the online stage, the gender dependent dictionary $\text{SPK}_{\text{GS}}$ is used to convert the voice of an unseen source speaker of the same gender. For example, for the conversion pair SF1→TM1, the source speaker was a female; therefore, the female version of $\text{SPK}_{\text{GS}}$ would be used. To build the gender dependent dictionary $\text{SPK}_{\text{GS}}$, voices from another four female speakers (i.e., SF2, SF3, TF1, and TF2) in the VCC2016 corpus were used. Note that the voices of the source speaker SF1 were not involved in building the dictionary $\text{SPK}_{\text{GS}}$. Our preliminary results showed that a gender dependent dictionary $\text{SPK}_{\text{GS}}$ achieved better VC performance than a gender independent dictionary $\text{SPK}_{\text{GS}}$ (i.e., constructed without considering the gender information). For each conversion pair, 648 (162x4) utterance pairs extracted from the official training set were used to build the $\text{SPK}_{\text{GS}}$ dictionary and the corresponding $\text{SPK}_{\text{Tgt}}$ dictionary. The ***MTO-LLE*** system operated on the same spectral features as the ***LLE$_{MCC}$-GV*** system.

#### (B) Objective Evaluations

The objective evaluation was conducted on the test set in terms of MCD using Eq. (10). We reported the average MCD value of all the test utterance pairs. Fig. 6 shows the average MCD values of ***MTO-LLE*** with different sizes of dictionaries and ***LLE$_{MCC}$-GV***. First, we analyzed the effect of the size of the dictionary (i.e., the number of speakers used for constructing the $\text{SPK}_{\text{GS}}$ dictionary) on the performance of the proposed ***MTO-LLE*** system, where ***D1***, ***D2***, ***D3***, and ***D4*** in Fig. 6 denote the gender dependent $\text{SPK}_{\text{GS}}$ dictionaries constructed by using one, two, three, and four speakers' voices, respectively. Note that ***D1*** was a subset of ***D2***, ***D2*** was a subset of ***D3***, and ***D3*** was a subset of ***D4***, and the dictionaries ***D1***, ***D2***, ***D3***, and ***D4*** contained about 87,100 to 113,200, 186,600 to 223,000, 286,000 to 324,700, and 385,600 to 424,300 exemplars, respectively, for different conversion pairs. From Fig. 6, it is observed that the MCD value decreases

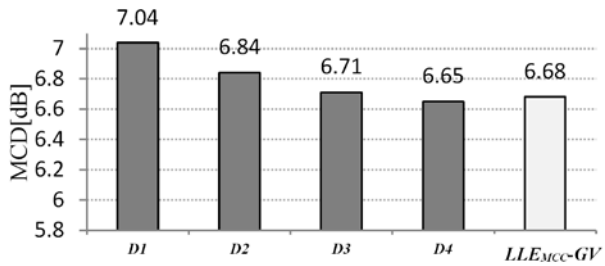



Fig. 6. The average MCD values of ***MTO-LLE*** with different sizes of dictionaries (***D1***~***D4***) and ***LLE$_{MCC}$-GV***. The MCD value before conversion is 9.45dB.

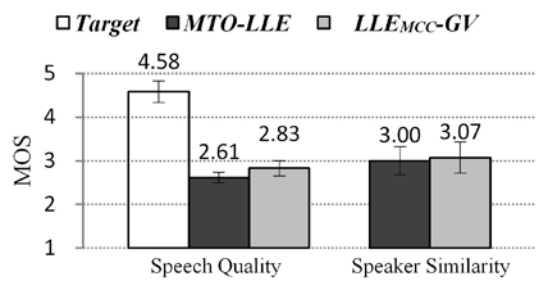



Fig. 7. Subjective test results of the speech converted by ***MTO-LLE*** and ***LLE$_{MCC}$-GV***, where ***D4*** was adopted for ***MTO-LLE***. Error bars indicate 95% confidence intervals. "***Target***" denotes the analysis-synthesized target speech.

as the number of speakers used for constructing the $SPK_{GS}$ dictionary increases. The result implies that the VC performance can be improved by increasing the size of the $SPK_{GS}$ dictionary (as well as the corresponding $SPK_{Tgt}$ dictionary). The result is expectable, as when more exemplars are available in the $SPK_{GS}$ dictionary, more exemplars that are similar/close to the source spectral features can be found, thereby increasing the capability of the $SPK_{GS}$ dictionary for characterizing the local geometry of the source spectral features.

Next, we compared ***MTO-LLE*** (using ***D4*** as the $SPK_{GS}$ dictionary) with ***LLE$_{MCC}$-GV***. From Fig. 6, we observe that ***MTO-LLE*** and ***LLE$_{MCC}$-GV*** obtained similar MCD values. The result indicates that the proposed many-to-one SC system performs as well as the proposed one-to-one SC system, when both systems are applied to the same source-target pair. In fact, ***MTO-LLE*** even gives a slightly lower MCD value than ***LLE$_{MCC}$-GV*** (6.65dB vs. 6.68dB). The reason could also be that the dictionary in ***MTO-LLE*** contained much more exemplars than the dictionary in ***LLE$_{MCC}$-GV***.

**(C) Subjective Evaluations**

We conducted the MOS test to evaluate the speech quality and speaker similarity of the converted speech. Ten listeners participated in the test. For each conversion pair, five sentences were randomly selected from the test set, thereby resulting in 20 (4x5) test sentences. During the MOS test, the converted speech samples (obtained by the two SC

systems) for each test sentence were presented to a subject in a random order. To evaluate the speech quality, for each test sentence, the listeners were requested to give a score from 1 (bad) to 5 (excellent) for two converted speech samples and an analysis-synthesized target speech sample. On the other hand, to evaluate the speaker similarity, for each test sentence, the listeners were requested to give a score from 1 (very dissimilar) to 5 (very similar) for each converted speech sample by comparing it with the corresponding analysis-synthesized target speech sample. Fig. 7 shows the overall average results of the MOS test.

From Fig. 7, we can see that ***MTO-LLE*** performs almost as well as ***LLE$_{MCC}$-GV*** in the speaker similarity test, but slightly worse in the speech quality test. With a further *t*-test on the scores of the compared systems, the *p*-values in the speech quality and speaker similarity tests were *0.058* and *0.779*, respectively. Both *p*-values are larger than *0.05*, which implies that the differences between ***MTO-LLE*** and ***LLE$_{MCC}$-GV*** are not significant. In general, the result of the subjective test is consistent with that of the objective test in Fig. 6.

From the results of the objective and subjective tests, we can confirm that the ***MTO-LLE*** many-to-one system can yield comparable performance to the ***LLE$_{MCC}$-GV*** one-to-one system. Consider that ***MTO-LLE*** does not require the source speaker's training speech, which is required in ***LLE$_{MCC}$-GV***, we believe that ***MTO-LLE*** can be suitably applied in real-world VC scenarios. In that case, converting speech in real time is a critical issue at the online conversion stage. We have found that the architecture of the proposed LLE-based SC framework is suitable for implementing a real-time many-to-one SC system. We are currently working on the new implementation.

## 5. CONCLUSIONS

In this paper, we have proposed a novel LLE-based SC framework. The proposed SC framework can be carried out in either the one-to-one or many-to-one manner. Experimental results confirm the effectiveness of the proposed SC framework. Our major findings include:

- The GV method can effectively overcome the over-smoothing problems existing in the proposed LLE-based SC systems, thereby notably improving the quality of the converted speech.
- Better sound quality can be achieved by using low-dimensional MCCs instead of high-dimensional SEs as spectral features in the proposed SC systems, mainly due to the fact that the GV method works more effectively with MCCs than SEs.
- The proposed one-to-one SC system (operating on MCCs) is comparable with the state-of-the-art GMM-based one-to-one SC system in terms of speech quality and speaker similarity.
- The proposed many-to-one SC system performs as well as the proposed one-to-one SC system in terms of speech quality and speaker similarity. However, the many-to-one SC system is more flexible than the one-to-one SC system since the former can convert the voice of any arbitrary unseen source speaker to that of a desired target speaker, without the requirement of the source speaker's training data.

Another advantage of the LLE-based SC framework is that its architecture is suitable for real-time SC. That is, the online conversion stage can be very efficient. Currently, we are working on the implementation of a real-time many-to-one SC system. In the future, we plan to investigate other manifold learning methods for SC and extend the proposed SC framework to other flexible SC scenarios, such as one-to-many and many-to-many SC tasks.

## ACKNOWLEDGEMENT

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

## REFERENCES

1. S. W. Fu, P. C. Li, Y. H. Lai, C. C. Yang, L. C. Hsieh, and Y. Tsao, "Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery," to appear in *IEEE Transactions on Biomedical Engineering*.
2. K. Park and H. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Proc. ICASSP*, 2000, pp. 1843–1846.
3. K. Kobayashi, T. Toda, G. Neubig, S. Sakti, S. Nakamura. "Statistical singing voice conversion based on direct waveform modification with global variance," in *Proc. INTERSPEECH*, 2015, pp. 2754–2758.
4. T. Toda, K. Nakamura, H. Sekimoto, and K. Shikano, "Voice conversion for various types of body transmitted speech," in *Proc. ICASSP*, 2009, pp. 3601–3604.
5. Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process*., vol. 6, no. 2, pp.131–142, Mar. 1998.
6. T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," in *Proc. ICASSP*, 2007, pp. 1249–1252.
7. R. Aihara, T. Takiguchi, and Y. Ariki, "Many-to-one voice conversion using exemplar-based sparse representation," in *Proc. WASPAA*, 2015, pp. 1–5.
8. L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. ICME*, 2016, pp.1–6.
9. D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in *Proc. INTERSPEECH*, 2011, pp. 653–656.
10. Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Many-to-many eigenvoice conversion with reference voice," in *Proc. INTERSPEECH*, 2009, pp. 1623–1626.
11. T. Nakashika, T. Takiguchi, and Y. Ariki, "Parallel-data-free, many-to-many voice conversion using an adaptive restricted Boltzmann machine, in *Proc. MLSLP*, 2015.
12. R. Aihara, T. Takiguchi, and Y. Ariki., "Multiple non-negative matrix factorization for many-to-many voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 24, no. 7, pp.1175–1184, 2016.

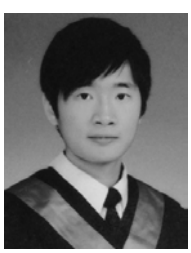
13.A. Kain, and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, 1998, pp. 285–288.
14.T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process*., vol. 15, no. 8, pp. 2222–2235, Nov. 2007.
15.S. Takamichi, T. Toda, A. Black, G. Neubig, S. Sakti, and S. Nakamura, " Post-filters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 757–767, Apr. 2016.
16.E. Helander, T.Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. Audio, Speech, Lang. Process*., vol. 18, no. 5, pp. 912–921, Jul. 2010.
17.V. Popa, H. Silen, J.Nurminen, and M. Gabbouj, "Local linear transformation for voice conversion," in *Proc. ICASSP*, 2012, pp. 4517–4520.
18.E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp.806–817, Mar. 2012.
19.S. Desai, A. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 954–964, Jul. 2010.
20.T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets," in *Proc. INTERSPEEH*, 2013, pp. 369–372.
21.Z. Wu, E. S. Chng, and H. Li, "Conditional restricted Boltzmann machine for voice conversion," in *Proc. ChinaSIP*, 2013, pp. 104–108.
22.H. T. Hwang, Y. Tsao, H. M. Wang, Y. R. Wang and S. H. Chen, "A Probabilistic interpretation for Artificial Neural Network-based Voice Conversion," in *Proc. APSIPA ASC*, 2015, pp.552–558.
23.L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L -R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp.1859–1872, Dec. 2014.
24.C. C. Hsu, H. T. Hwang, Y. C. Wu, Y. Tsao, and H. M. Wang, "Voice conversion from non-parallel corpora using variational Auto-encoder," in *Proc. APSIPA ASC*, 2016, pp.1–6.
25.C. C. Hsu, H. T. Hwang, Y. C. Wu, Y. Tsao, and H. M. Wang, "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," to appear in *Proc. INTERSPEECH*, 2017.
26.T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, 1992, pp. 137–140.
27.F. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. ICASSP*, 1984, pp. 37–40.
28.D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 922–931, Jul. 2010.

29. E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1313–1323, May 2012.
30. D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 3, pp. 556–566, Mar. 2013.
31. R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Proc. SLT*, 2012, pp. 313–317.
32. Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in *Proc. SSW*, 2013, pp. 201–206.
33. Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang, Process.*, vol. 22, no. 10, pp.1506–1521, Oct. 2014.
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 *Proc. INTERSPEECH*, 2016, pp. 1652–1656.
35. H. Y. Gu and S. F. Tsai, "A voice conversion method combining segmental GMM mapping with target frame selection", *Journal of Information Science and Engineering*, vol. 31, no. 2, pp. 609–626, 2015.
36. K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp.1315–1318.
37. H. Silén, E. Helander, J. Nurminen, M. Gabbouj, "Ways to implement global variance in statistical speech synthesis," in *Proc. INTERSPEECH*, 2012, pp. 1436–1439.
38. L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik. "Dimensionality reduction: A comparative review." *Journal of Machine Learning Research* 10.1–41 (2009): 66–71.
39. Y. Bengio, A. Courville, and P. Vincent, P. "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8), pp. 1798–1828, 2013.
40. G. Hinton and S. Roweis, "Stochastic neighbor embedding," *Advances in neural information processing systems*, vol. 15, pp. 833–840, 2002.
41. H. Chang, D.Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. CVPR*, 2004.
42. J.B. Tenenbaum, V. De Silva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
43. M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in neural information processing systems*, vol. 14, pp. 585–591, 2001.
44. S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
45. L.K. Saul and S.T. Roweis, "An introduction to locally linear embedding," (2001) Available from https://www.cs.nyu.edu/~roweis/lle/papers/lleintro.pdf.

46.C.-y. Tseng, Y. C. Cheng and C. H. Chang, “Sinica COSPRO and Toolkit - corpora and platform of Mandarin Chinese fluent speech,” in *Proc. Oriental COCOSDA*, 2005, pp. 23–28.
47.T. Toda, L. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, “The voice conversion challenge 2016,” in *Proc. INTERSPEECH*, 2016, pp. 1632–1636.
48.H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” *Speech Commun*., vol. 27, no. 3–4, pp.187–207, 1999.

**Hsin-Te Hwang (黃信德)** received the M.S. degree in Dept. of Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan. He is currently pursuing the Ph.D. degree in Dept. 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.



**Yu-Huai Peng (彭玉淮)** received the B.S. degree in Dept. of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan. He is currently pursuing the M.S. degree in Dept. of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan. He is an intern in Institute of Information Science, Academia Sinica, Taipei, Taiwan. His research interests include speech signal processing, particularly, voice conversion.

**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 (NICT), 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 (CITI), 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 (NCTU), 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.