# Locally Linear Embedding for Exemplar-Based Spectral Conversion

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

[1] Institute of Information Science, Academia Sinica, Taipei, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
yu.tsao@citi.sinica.edu.tw, {hwanght, whm}@iis.sinica.edu.tw

## Abstract

This paper describes a novel exemplar-based spectral conversion (SC) system developed by the AST (Academia Sinica, Taipei) team for the 2016 voice conversion challenge (vcc2016). The key feature of our system is that it integrates the locally linear embedding (LLE) algorithm, a manifold learning algorithm that has been successfully applied for the super-resolution task in image processing, with the conventional exemplar-based SC method. To further improve the quality of the converted speech, our system also incorporates (1) the maximum likelihood parameter generation (MLPG) algorithm, (2) the postfiltering-based global variance (GV) compensation method, and (3) a high-resolution feature extraction process. The results of subjective evaluation conducted by the vcc2016 organizer show that our LLE-exemplar-based SC system notably outperforms the baseline GMM-based system (implemented by the vcc2016 organizer). Moreover, our own internal evaluation results confirm the effectiveness of the major LLE-exemplar-based SC method and the three additional approaches with improved speech quality.

**Index Terms**: voice conversion, exemplar, locally linear embedding, voice conversion challenge.

## 1. Introduction

Voice conversion (VC) is a technique that transforms a source speaker's voice to that of a specific target speaker [1-13]. A VC system includes two parts, namely spectral and prosody conversions. This study focuses on spectral conversion (SC). Numerous VC approaches have been proposed, such as Gaussian mixture model (GMM)-based [1-4], frequency warping-based [5, 6], neural networks (NN)-based [7-10], and exemplar-based [11-13] approaches. In this study, we investigate the manifold learning algorithm for the exemplar-based SC.

Manifold learning is a popular method playing an essential role in the development of various algorithms, such as nonlinear dimensionality reduction [14], representation learning [15], and data visualization [16]. It attempts to discover underlying manifolds (or the intrinsic geometry of the data distribution) in high-dimensional data spaces and embed them onto low-dimensional embedding spaces. Many manifold learning methods have been proposed, including isometric feature mapping (Isomap) [17], Laplacian eigenmap (LE) [18], and locally linear embedding (LLE) [19]. We integrate LLE with the conventional exemplar-based SC, which is inspired by the success of applying LLE in the super-resolution task in image processing [20].

The proposed SC method is based on the assumption that the source and target feature vectors form manifolds with similar local geometry in two distinct feature spaces. As a result, we characterize the local geometry of the locally linear patches in the source spectral feature space. Given a source feature vector, we estimate the reconstruction weights using the LLE algorithm. Then, the reconstruction weights are applied to the corresponding target exemplars to construct the converted feature vector. To further improve the quality of the converted speech, the maximum likelihood parameter generation (MLPG) algorithm [2, 21] and the postfiltering-based global variance (GV) compensation method [22] are adopted. Experimental results demonstrate the effectiveness of the proposed SC method.

The remainder of this paper is organized as follows. Section 2 describes the proposed LLE-exemplar-based SC system. Section 3 presents the experimental results. Finally, Section 4 gives the conclusions.

## 2. LLE-exemplar-based SC

Figure 1 illustrates the block diagram of the proposed LLE-exemplar-based SC system. From figure 1, it can be seen that there are mainly two stages, namely offline and online stages. In this section, we describe the proposed SC system in detail.

### 2.1. The Offline Stage

As shown in Figure 1, a set of paired dictionaries is composed by the source and target dictionaries in advance, same as the conventional exemplar-based VC systems [11, 12]. Specifically, a parallel speech corpus consisting of the source and target speakers' speeches is needed. After spectral feature extraction, a dynamic time warping (DTW) algorithm is used to time-align the spectral feature sequences of the source and target speech utterances. Accordingly, the paired dictionaries can be constructed from the aligned joint spectral feature vectors. Additionally, some statistics to be used in the MLPG algorithm and the postfiltering-based GV conversion, such as mean and precision/variance of the target spectral features, are estimated according to the maximum likelihood (ML) criterion.

Let the source and target dictionaries be composed by the source and target spectral feature vectors (called exemplars hereafter) as $\bar{\mathbf{X}} = \left[\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_n, \cdots, \bar{\mathbf{X}}_N\right]$ and $\bar{\mathbf{Y}} = \left[\bar{\mathbf{Y}}_1, \cdots, \bar{\mathbf{Y}}_n, \cdots, \bar{\mathbf{Y}}_N\right]$, respectively. The total number of both exemplars is *N*. $\bar{\mathbf{X}}_n$ and $\bar{\mathbf{Y}}_n$ are the source and target exemplars at frame *n*, respectively. In order to take into account temporal information for achieving better conversion performance, both source and target exemplars are composed by their static, and dynamic features as $\bar{\mathbf{X}}_n = \left[\bar{\mathbf{x}}_n^{\mathrm{T}}, \Delta^{(1)}\bar{\mathbf{x}}_n^{\mathrm{T}}, \Delta^{(2)}\bar{\mathbf{x}}_n^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\bar{\mathbf{Y}}_n = \left[\bar{\mathbf{y}}_n^{\mathrm{T}}, \Delta^{(1)}\bar{\mathbf{y}}_n^{\mathrm{T}}, \Delta^{(2)}\bar{\mathbf{y}}_n^{\mathrm{T}}\right]^{\mathrm{T}}$, for *n*=1~*N*. The superscript $\mathrm{T}$ denotes transposition of the vector. $\bar{\mathbf{x}}_n$, $\Delta^{(1)}\bar{\mathbf{x}}_n$, and $\Delta^{(2)}\bar{\mathbf{x}}_n$ are *D*-dimensional source static, delta and delta-delta features, respectively. Likewise, $\bar{\mathbf{y}}_n$,

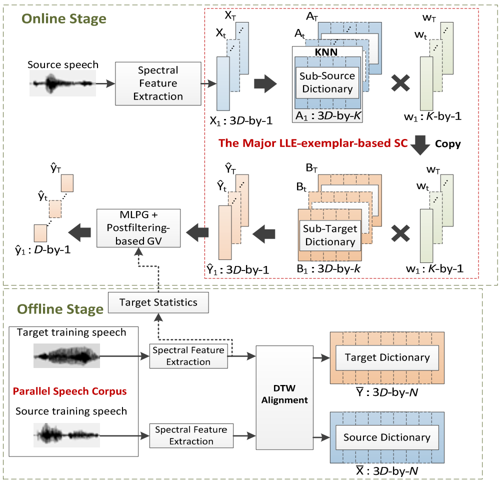


Figure 1: *The Proposed LLE-exemplar-based SC system.*

$\Delta^{(1)}\bar{\mathbf{y}}_n$, and $\Delta^{(2)}\bar{\mathbf{y}}_n$ are $D$-dimensional target static, delta and delta-delta features, respectively.

## 2.2. The Online Stage

From Figure 1, the online stage first applies feature extraction to convert the input utterance (source speech) into a sequence of spectral feature vectors. Then, the LLE-exemplar-based SC algorithm is performed to convert the source spectral features (composed by static and dynamic features). Finally, the MLPG and GV compensation methods are adopted to improve the quality of the converted speech. In the following subsections, we will review the LLE algorithm and describe each component of the proposed SC system.

### *2.2.1. The LLE algorithm*

The LLE algorithm is a manifold learning method that computes the low-dimensional embeddings that best preserve the local geometry of each locally linear patch in the high-dimensional space [19]. 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. In other words, each data point and its neighbors are expected to lie on or close to a locally linear patch of the manifold. Thus, the local geometry of these patches can be characterized by the reconstruction weights for reconstructing each data point from its neighbors. Moreover, the chart from the manifold to the low-dimensional feature space will be roughly linear on these small patches. Based on this idea, the LLE algorithm has three steps: 1) identifying the locally linear patch by finding a set of $K$ nearest neighbors for each data point; 2) characterizing the local geometry of each locally linear patch by estimating the reconstruction weights of the corresponding neighbors that minimize the local reconstruction error; 3) computing the low-dimensional embeddings by finding a mapping that best preserves the local geometry and is nearly linear.

### *2.2.2. The major LLE-exemplar-based SC*

As in steps 1 and 2 of the LLE algorithm, we characterize the local geometry of each locally linear patch in the source spectral feature space first. Then, the converted feature vectors are estimated from the paired dictionaries by preserving the local geometry (as opposed to estimating the low-dimensional embeddings in step 3 of the LLE algorithm). Specifically, we identify $K$ nearest neighbors (measured by the Euclidean distance) from the source dictionary for each source spectral feature vector. Then, we estimate the reconstruction weights by minimizing the local reconstruction error:

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

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

$$\hat{\mathbf{w}}_t = \frac{\mathbf{G}_t^{-1}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\mathbf{G}_t^{-1}\mathbf{1}} , \qquad (2)$$

where $\mathbf{G}_t$ is the local Gram matrix ($K$-by-$K$) for $\mathbf{X}_t$ :

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

A more efficient way is to solve the linear system of equations $\mathbf{G}_t \mathbf{w}_t = \mathbf{1}$, and then rescale the weights to satisfy the constraint $\mathbf{1}^{\mathrm{T}}\mathbf{w}_t = 1$. The detailed derivations of the solution can be found in [23]. Finally, with the assumption that the source and target feature vector manifolds share a similar local geometry in two distinct spectral feature spaces, the converted spectral feature vector (composed of static and first- and second-order dynamic features) hence can be obtained by applying the reconstruction weights obtained in the source spectral feature space to the corresponding sub-target dictionary as

$$\hat{\mathbf{Y}}_t = \mathbf{B}_t \mathbf{w}_t = \sum_{k=1}^{K} \mathbf{w}_t(k) \mathbf{b}_{tk} , \qquad (4)$$

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

### *2.2.3. Parameter generation*

Since the LLE-exemplar-based SC method is performed in a frame-by-frame manner with the aim of minimizing the reconstruction error, two issues that are often encountered in other SC methods could also appear, namely, the discontinuity problem and the over-smoothing problem (which will be shown in section 3). In this study, we adopt the MLPG algorithm and the postfiltering-based GV compensation method to handle these two problems.

*1) MLPG Algorithm:* When incorporating the MLPG algorithm in the LLE-exemplar-based SC system, we have

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

where $\hat{\mathbf{y}}$ (a $DT$-by-1 vector) is the concatenated vector of the converted static spectral feature vector sequence obtained by the MLPG algorithm; $\mathbf{M}$ is a 3$DT$-by-$DT$ weighting matrix used for appending the dynamic features to the static ones; $\hat{\mathbf{Y}} = [\hat{\mathbf{Y}}_1^{\mathrm{T}}, \cdots, \hat{\mathbf{Y}}_t^{\mathrm{T}}, \cdots, \hat{\mathbf{Y}}_T^{\mathrm{T}}]^{\mathrm{T}}$ (a 3$DT$-by-1 vector) is the concatenated vector of the converted spectral feature vector sequence obtained by the LLE-exemplar-based SC system, i.e., $\hat{\mathbf{Y}}_t$ (for $t$=1~$T$) in $\hat{\mathbf{Y}}$ is given by (4); $\mathbf{U} = \mathrm{daig}[\mathbf{\Lambda}_1^{(\mathbf{Y})}, \cdots, \mathbf{\Lambda}_t^{(\mathbf{Y})}, \cdots, \mathbf{\Lambda}_T^{(\mathbf{Y})}]$ (a 3$DT$-by-3$DT$ matrix) is the global precision matrix, where $\mathbf{\Lambda}_1^{(\mathbf{Y})} = \cdots = \mathbf{\Lambda}_t^{(\mathbf{Y})} = \cdots = \mathbf{\Lambda}_T^{(\mathbf{Y})}$, $\mathbf{\Lambda}_t^{(\mathbf{Y})}$ (a 3$D$-by-3$D$ matrix) is assumed to be diagonal and is estimated from the precision/variance of the training data of the target speaker.

*2) Postfiltering-based GV compensation:* To overcome the over-smoothing problem, we further adopt the postfiltering-based GV compensation method. The converted static spectral feature vector $\hat{\mathbf{y}}$ given by (5) is processed by

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

where

$$\hat{v}(d) = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}_t(d) - \langle \hat{y}(d) \rangle\right)^2\ , \tag{7}$$

$$\langle \hat{y}(d) \rangle = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t(d)\ . \tag{8}$$

$\hat{y}_t(d)$ is the $d$-th element of the converted static spectral feature vector at frame $t$ obtained from the converted spectral static feature sequence $\hat{\mathbf{y}}$ given by (5). $\langle \hat{y}(d) \rangle$ and $\hat{v}(d)$ are the mean and variance of the converted static spectral feature vector. $\mu_v(d)$ is the $d$-th element of the mean vector of the target GV, which is obtained using the GVs of the target feature sequences calculated from individual utterances in the training data as described in [2]. Note that this method has been used as the initialization for the maximum likelihood global variance (MLGV) algorithm [2]. However, it has also been used in VC and speech synthesis for reducing the computational cost of the MLGV algorithm [22].

# 3. Experiments

## 3.1. Experimental Setup

The proposed LLE-exemplar-based SC system was evaluated using a parallel English speech corpus provided by the vcc2016 organizer [26]. The corpus was divided to training and evaluation sets. The training set comprises 5 source (3 females and 2 males) and 5 target speakers (2 females and 3 males), with 162 utterances for each speaker. The evaluation set comprises the same 5 source and 5 target speakers, with 54 utterances for each speaker. Speech signals were recorded at 16 kHz sampling rate, and the resolution per sample was 16 bits. We used the entire training set to build our system.

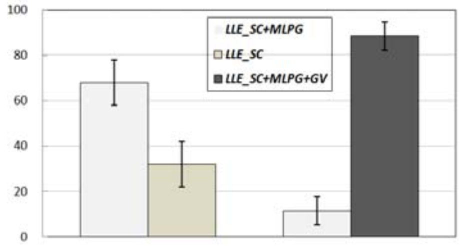


Figure 2: *Preference test results of naturalness. Error bars indicate 95% confidence intervals.*

The STRAIGHT [24] toolkit was used for feature extraction and waveform generation. For SC, each frame of speech signal was converted to a static feature vector with 513-dimensional spectral envelopes (we used the logarithmic magnitude spectra, LMS). The delta and delta-delta feature vectors were then appending to the static one to form the final spectral feature vector. Accordingly, the dimension of each final spectral feature vector was 1539. The LMS features among all frames were normalized to the same energy. Additionally, the first through 24-th Mel-cepstral coefficients (MCCs) (obtained by the FestVox toolkit [25]) extracted from the STRAIGHT spectral envelopes were used to align the source and target MCCs in order to obtain the source-target alignment. The paired dictionaries of the proposed SC system (described in Section 2) were constructed by using this alignment information. Moreover, the number of nearest neighbors, namely $K$ in (1), for the LLE algorithm was set to 1024 empirically.

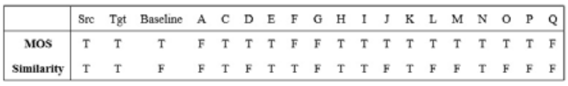
For prosodic conversion, we performed f0 (in log-scaled) conversion based on the liner transformation method, which is typically adopted in a conventional VC system [2]. Notably, the aperiodic components extracted by STRAIGHT are not converted in our VC system. Finally, the STRAIGHT synthesis was performed to generate speech waveforms by using the converted spectral features, converted f0 features, and source aperiodic components. Because we did not conduct voice activity detection (VAD) to exclude silence parts of speech utterances, VC was performed on not only speech but also silence parts of an input utterance. Moreover, the target statistics used for the MLPG and GV conversions (such as target GV) were all estimated from the target training utterances (including silence and speech parts).

## 3.2. Internal Evaluation Results

In the internal experiments, we compare three SC systems, namely, the LLE-exemplar-based SC system described in Section 2.2.2 (denoted as ***LLE_SC***), ***LLE_SC*** with the MLPG algorithm (denoted as ***LLE_SC+MLPG***), and ***LLE_SC+MLPG*** with the postfiltering-based GV compensation method (denoted as ***LLE_SC+MLPG+GV***). A preference listening test was conducted to evaluate the naturalness of the converted speech. We performed the male to male (i.e., SM1 to TM2) and female to male (i.e., SF1 to TM1) VC. The first 25 test sentences were chosen from the test set. We conducted an AB test, i.e., each pair of converted speeches by methods **A** and **B** were presented in a random order to five subjects. Figure 2 shows the overall average results of the preference test. From Figure 2, we can see that ***LLE_SC+MLPG*** outperforms ***LLE_SC.*** The result confirms the effectiveness of employing the MLPG algorithm in LLE-based SC. We also observe that ***LLE_SC+MLPG+GV*** achieves a significant gain over

| | Median | MAD | Mean | SD | # data points |
|---|---|---|---|---|---|
| **Source** | 5 | 0.00 | 4.62 | 0.69 | 1600 |
| **Target** | 5 | 0.00 | 4.57 | 0.74 | 1600 |
| **Baseline** | 1 | 0.00 | 1.48 | 0.72 | 1600 |
| **A** | 3 | 1.00 | 2.67 | 0.97 | 1600 |
| ***B** | **3** | **1.00** | **2.67** | **0.96** | **1600** |
| **C** | 1 | 0.00 | 1.30 | 0.63 | 1600 |
| **D** | 2 | 1.00 | 2.27 | 0.94 | 1600 |
| **E** | 2 | 1.00 | 2.43 | 1.02 | 1600 |
| **F** | 3 | 1.00 | 2.78 | 0.99 | 1600 |
| **G** | 3 | 1.00 | 2.78 | 1.02 | 1600 |
| **H** | 2 | 1.00 | 2.36 | 1.07 | 1600 |
| **I** | 1 | 0.00 | 1.51 | 0.75 | 1600 |
| **J** | 3 | 1.00 | 3.03 | 0.99 | 1600 |
| **K** | 3 | 1.00 | 3.25 | 1.08 | 1600 |
| **L** | 3 | 1.00 | 2.98 | 0.99 | 1600 |
| **M** | 2 | 1.00 | 2.10 | 0.92 | 1600 |
| **N** | 3 | 1.00 | 3.29 | 1.20 | 1600 |
| **O** | 3 | 1.00 | 2.97 | 1.07 | 1600 |
| **P** | 3 | 1.00 | 2.85 | 0.96 | 1600 |
| **Q** | 3 | 1.00 | 2.60 | 1.05 | 1600 |

Table 1: *Mean Opinion Score (MOS).*



| | Src | Tgt | Baseline | A | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MOS** | T | T | T | F | T | T | T | F | F | T | T | T | T | T | T | T | T | T | F |
| **Similarity** | T | T | F | F | T | F | T | T | F | T | T | F | T | F | F | T | F | F | F |

Table 2: *Pairwise Wilcoxon signed rank tests.*

***LLE_SC+MLPG***. The result indicates that introducing the postfiltering-based GV compensation method further improves the quality of the converted speech.

## 3.3. External Evaluation Results

The external evaluation was conducted by the vcc2016 organizer in terms of naturalness and similarity of the converted speech [26, 27]. We built a VC system for each pair of source and target speakers using the entire 162 parallel utterance pairs; therefore, 25 VC systems were constructed. The converted voice samples were submitted to the vcc2016 organizer for performance evaluation. For the naturalness evaluation, the standard 5-point mean opinion score (MOS) test was adopted, where 5 stands for "completely natural" and 1 for "completely unnatural". For the similarity evaluation, a pairwise comparison between the converted speech and the natural speech from the reference (target) speaker was conducted, with 1 for "sounds like absolutely the same person" and 4 for "sounds like absolutely different person". 200 (52 males and 148 females) subjects took part in the evaluation.

### *3.3.1. MOS evaluation*

Table 1 shows the overall results of the MOS test for 18 systems (including the baseline GMM-based system implemented by the vcc2016 organizer). Our system is denoted as "**B**". The median, median absolute deviation (MAD), mean, and standard deviation (SD) of MOS were given for each system. Table 2 further shows Pairwise Wilcoxon signed rank tests (with alpha Bonferoni corrected) to determine the significant of the differences between two systems in view of the MOS and similarity tests. "T" and "F" denote that significance at 1% level is true and false, respectively.

In Table 1, our system "**B**" yields a notable gain in mean and median (mean: 2.67, and median: 3) over the baseline (mean: 1.48, and median: 1). From Table 2 (the MOS row), the result indicates that the difference between our system and the baseline system is significant. The results confirm that our system outperforms the baseline in terms of naturalness.

Table 1 shows that eight systems yield higher mean scores than our system in the MOS test. However, Table 2 shows that there is no significant difference between our system and four systems, among which "**F**" and "**G**" give higher mean scores than ours while "**A**" and "**Q**" give equal and lower mean scores, respectively, than ours. In other words, six systems are significantly better than our system in terms of naturalness. These results indicate that the performance of our system in terms of naturalness is above average among 18 systems.

| | M2M | M2F | F2F | F2M | All |
|---|---|---|---|---|---|
| **Source** | 14.47 | 3.29 | 18.24 | 4.05 | 10.00 |
| **Target** | 92.76 | 93.42 | 93.24 | 90.54 | 92.50 |
| **Baseline** | 42.76 | 74.34 | 68.24 | 53.38 | 59.67 |
| **A** | 67.76 | 61.18 | 70.95 | 66.89 | 66.67 |
| ***B** | **52.63** | **74.34** | **76.35** | **53.38** | **64.17** |
| **C** | 24.34 | 25.66 | 20.95 | 36.49 | 26.83 |
| **D** | 58.55 | 82.24 | 74.32 | 61.49 | 69.17 |
| **E** | 27.63 | 30.26 | 35.81 | 35.81 | 32.33 |
| **F** | 48.03 | 52.63 | 43.92 | 48.65 | 48.33 |
| **G** | 63.16 | 75.66 | 68.24 | 69.59 | 69.17 |
| **H** | 46.05 | 7.89 | 60.14 | 4.73 | 29.67 |
| **I** | 31.58 | 42.11 | 41.22 | 37.16 | 38.00 |
| **J** | 68.42 | 79.61 | 80.41 | 61.49 | 72.50 |
| **K** | 48.03 | 55.92 | 45.95 | 47.3 | 49.33 |
| **L** | 59.21 | 74.34 | 71.62 | 56.76 | 65.50 |
| **M** | 44.08 | 69.74 | 56.76 | 52.03 | 55.67 |
| **N** | 21.71 | 16.45 | 16.89 | 22.30 | 19.33 |
| **O** | 59.87 | 75.00 | 66.22 | 62.16 | 65.83 |
| **P** | 57.24 | 74.34 | 80.41 | 66.89 | 69.67 |
| **Q** | 42.11 | 59.21 | 62.84 | 49.32 | 53.33 |

Table 3: *Similarity score (%).*

### *3.3.2. Similarity evaluation*

Table 3 shows the results of the similarity test for 18 systems. "**M2M**", "**M2F**", "**F2F**", "**F2M**" and "**All**" denotes "male to male VC", "male to female VC", "female to female VC", "female to male VC", and the overall performance, respectively. The similarity score is the percentage of converted samples judged to be the same as the corresponding target speaker. The higher the similarity score, the better the VC system achieves the target speaker identity.

From Table 3 and Table 2 (the similarity row), we first observe that, although our system yields a higher overall score (i.e., 64.17%) than the baseline system (i.e., 59.67%), the gain is not significant. Seven systems yield higher overall scores than our system in the similarity test. Table 2 shows that there is no significant difference between our system and ten systems, among which seven systems give higher scores than ours (the highest score is 72% by "**J**") while the remaining three systems give lower scores than ours (the lowest score is 53% by "**Q**"). Our system yields higher scores in the M2F and F2F cases than the F2M and M2M cases. The reason is worth further study.

# 4. Conclusions

This paper has presented a novel exemplar-based SC system. The main contributions are threefold. First, we investigate the effectiveness of a manifold learning algorithm, i.e., the LLE algorithm, for the exemplar-based SC. Second, the MLPG algorithm is employed in our LLE-exemplar-based SC system to address the discontinuity problem existing in exemplar-based SC systems. Third, a postfiltering-based GV compensation method is adopted to improve the quality of the converted speech. We participated in the vcc2016 evaluation. The evaluation results provided by the vcc2016 organizer demonstrate that our VC system achieves satisfied results. In the future, we will study other spectral features, e.g., MCCs, and investigate whether the proposed spectral conversion framework is also effective in prosodic conversion.

# 5. References

[1] 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.

[2] T. Toda, A. W. 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.

[3] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, “Modulation spectrum-constrained trajectory training algorithm for GMM-based voice conversion,” *Proc. ICASSP*, 2015.

[4] H. T. Hwang, Y. Tsao, H. M. Wang, Y. R. Wang and S. H. Chen, “Incorporating global variance in the training phase of GMM-based voice conversion,” *Proc. APSIPA*, 2013.

[5] 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, July. 2010.

[6] 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.

[7] S. Desai, A. W. 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, 2010.

[8] 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, 2014.

[9] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, “Voice conversion in high-order eigen space using deep belief nets,” *Proc. INTERSPEEH*, 2013.

[10] 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," *Proc. APSIPA*, 2015.

[11] R. Takashima, T. Takiguchi, and Y. Ariki, “Exemplar-based voice conversion in noisy environment,” *Proc. Spoken Language Technology Workshop (SLT)*, 2012.

[12] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, “Exemplar based voice conversion using non-negative spectrogram deconvolution,” *Proc. 8th ISCA Speech Synth. Workshop (SSW8)*, 2013.

[13] 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, 2014.

[14] 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.

[15] 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.

[16] G. Hinton and S. Roweis, “Stochastic neighbor embedding,” *Advances in neural information processing systems*, vol. 15, pp. 833–840, 2002.

[17] 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.

[18] 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.

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

[20] H. Chang, D.Y. Yeung, and Y. Xiong, “Super-resolution through neighbor embedding,” *Proc. CVPR, 2004.*

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

[22] H. Silén, E. Helander, J. Nurminen, M. Gabbouj, “Ways to implement global variance in statistical speech synthesis,” *Proc. INTERSPEECH*, 2012.

[23] 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.

[24] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based $F_0$ extraction: Possible role of a repetitive structure in sounds,” *Speech Commun.*, vol. 27, no. 3-4, pp.187-207, 1999.

[25] Festvox. Available: http://www.festvox.org/download.html.

[26] T. Toda, L. H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu and J. Yamagishi , “The Voice Conversion Challenge 2016,” *Proc. INTERSPEECH*, 2016.

[27] M. Wester, Z. Wu and J. Yamagishi, “Analysis of the Voice Conversion Challenge 2016 Evaluation Results,” *Proc. INTERSPEECH*, 2016.