Exemplar-Based Spectral Detail Compensation for Voice Conversion

Yu-Huai Peng[1], *Hsin-Te Hwang*[1], *Yi-Chiao Wu*[2], *Yu Tsao*[3], *Hsin-Min Wang*[1]

[1]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[2]Graduate School of Informatics, Nagoya University, Japan
[3]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
{roland0601,hwanght, whm}@iis.sinica.edu.tw, yu.tsao@citi.sinica.edu.tw

Abstract

Most voice conversion (VC) systems are established under the vocoder-based VC framework. When performing spectral conversion (SC) under this framework, the low-dimensional spectral features, such as mel-ceptral coefficients (MCCs), are often adopted to represent the high-dimensional spectral envelopes. The joint density Gaussian mixture model (GMM)-based SC method with the STRAIGHT vocoder is a well-known representative. Although it is reasonably effective, the loss of spectral details in the converted spectral envelopes inevitably deteriorates speech quality and similarity. To overcome this problem, we propose a novel exemplar-based spectral detail compensation method for VC. In the offline stage, the paired dictionaries of source spectral envelopes and target spectral details are constructed. In the online stage, the locally linear embedding (LLE) algorithm is applied to predict the target spectral details from the source spectral envelopes, and then, the predicted spectral details are used to compensate the converted spectral envelopes obtained by a baseline GMM-based SC method with the STRAIGHT vocoder. Experimental results show that the proposed method can notably improve the baseline system in terms of objective and subjective tests.

Index Terms: locally linear embedding, exemplar, Gaussian mixture model, vocoder, voice conversion

1. Introduction

Voice conversion (VC) transforms one type of speech to another, without changing the linguistic content. Speaker VC, which converts a source speaker's speech to a target speaker's speech, is a typical VC task. Most speaker VC systems perform spectral, prosodic, and excitation conversions under the vocoder-based VC framework. In this paper, we focus on spectral conversion (SC) using the STRAIGHT vocoder [1].

Numerous SC methods have been proposed during the last two decades. Among them, statistical methods operating on various low-dimensional spectral features, e.g., mel-ceptral coefficients (MCCs) [2], have been extensively studied due to the advantages of high computational efficiency and avoiding the curse of dimensionality problem [3-7]. The joint density Gaussian mixture model (GMM)-based method is a representative [4, 5]. The maximum likelihood parameter generation (MLPG) method for reducing discontinuity and the global variance (GV) and modulation spectrum (MS) methods for overcoming the over smoothing problem have been successfully applied to a GMM-based SC system [5, 6]. Although notable improvements have been achieved, the loss of spectral details in the high-dimensional STRAIGHT spectral envelopes reconstructed from the converted low-dimensional spectral features may still affect the similarity and speech quality of converted speech.

Several methods have been proposed to tackle this problem by directly conducting SC on the spectral envelopes. For instance, the deep neural network-based methods learned the mapping between the source and target spectral envelopes [8, 9]. The exemplar-based methods generated the converted spectral envelopes using the weighted linear combination of the target spectral envelope exemplars, where the weights were estimated by nonnegative matrix factorization (NMF) [10, 11] or a locally linear embedding (LLE) method [12]. Even with such efforts, the GMM-based SC method (using MCCs) is still very competitive today [13]. Our previous study also showed that the LLE-exemplar-based SC with the low-dimensional MCCs outperformed that with the high-dimensional spectral envelopes [14].

Another direction is to avoid using the STRAIGHT vocoder for waveform generation. For instance, the neural network-based vocoder has been proposed recently for VC and shown promising results [15]. However, it requires a huge amount of training data for high quality synthesis while the conventional vocoder does not require any training data. On the other hand, a vocoder-free VC framework has also been proposed [16]. Although notable improvements in speech quality in the intra-gender VC pairs have been achieved, the performance is comparable with the vocoder-based framework in terms of speech quality and similarity in the inter-gender VC pairs. Therefore, it is still worthwhile to further investigate the vocoder-based VC method.

In this paper, we propose an exemplar-based spectral detail compensation method for any SC methods that worked on low-dimensional spectral features. We use the GMM-based SC method operating on the MCCs under the STRAIGHT vocoder as a case study. Specifically, the predicted spectral details are used to compensate the converted spectral envelopes obtained by a GMM-based SC method. The LLE algorithm is adopted for spectral detail prediction due to its satisfactory ability to handle the high-dimensional spectral features in LLE-based SC [12] and speech enhancement [17-19]. The remainder of this paper is organized as follows. The proposed spectral detail compensation method is described in Section 2. Experimental setup and results are presented in Section 3. Finally, Section 4 gives the conclusions.

2. The proposed spectral detail compensation method

2.1. System overview

Figure 1 gives an overview of the run-time conversion stage of the proposed spectral detail compensation method. Given a

source speech for conversion, the source spectral envelopes and MCCs (each source spectral envelope or MCC vector is composed of multi-dimensional static, delta, and delta-delta features) are extracted by the STRAIGHT vocoder first. Next, the GMM-based SC method is applied to convert the source MCCs to the converted spectral envelopes (reconstructed from the converted MCCs). Meanwhile, the LLE-based spectral detail prediction (LLE-based SDP) method followed by an additional "MLPG+GV" module is applied to predict the target spectral details from the source spectral envelopes. Finally, the predicted spectral details are used to compensate the converted spectral envelopes. The waveform can be reconstructed by the converted spectral envelopes and other acoustic features (described later) using the STRAIGHT synthesis method. In the following, we detail the offline and online stages of the LLE-based SDP method and the complete spectral detail compensation process.

2.2. Offline stage of the LLE-based SDP method

The offline stage of the LLE-based SDP method mainly involves the construction of the paired source and target dictionaries in the following steps: 1) preparing a parallel speech corpus consisting of the source and target speakers' voices; 2) extracting the MCCs and spectral envelopes from the source and target speakers' voices; 3) computing the reconstructed target spectral envelopes by reverting the target MCCs back to the spectral envelopes; 4) computing the target spectral details by subtracting the reconstructed target spectral envelopes from the target spectral envelopes; 5) computing the dynamic features of the source spectral envelopes and the target spectral details, and then appending the dynamic features to the corresponding static spectral envelopes and spectral details; 6) performing dynamic time warping (DTW) to align the source and target MCCs to obtain the frame alignment information; 7) applying the frame alignment information to the source spectral envelopes and the target spectral details to obtain the aligned source spectral envelopes and target spectral details; 8) constructing the paired source spectral envelope and target spectral detail dictionaries from the aligned source spectral envelopes and target spectral details.

In step 4), the target spectral details are computed as follows. Let $\mathbf{c} = [\mathbf{c}_1, \cdots, \mathbf{c}_n, \cdots, \mathbf{c}_N]$ and $\bar{\mathbf{c}} = [\bar{\mathbf{c}}_1, \cdots, \bar{\mathbf{c}}_n, \cdots, \bar{\mathbf{c}}_N]$ be the sequences of the target spectral envelope and reconstructed target spectral envelope vectors, respectively. N denotes the number of speech frames, and the dimensionality of each spectral envelope vector (i.e., $\{\mathbf{c}_n\}_{n=1}^{N}$ and $\{\bar{\mathbf{c}}_n\}_{n=1}^{N}$) is D. Then, the sequence of the target spectral detail vectors (denoted as $\mathbf{r} = [\mathbf{r}_1, \cdots, \mathbf{r}_n, \cdots, \mathbf{r}_N]$) is computed as $\mathbf{r} = \mathbf{c} - \bar{\mathbf{c}}$. After conducting steps 4) and 5), the covariance matrix of the 3D-dimensional target spectral detail vectors (composed of D-dimensional static, delta, and delta-delta features) and the global variances of individual elements of the target (static) spectral detail vectors are estimated to be used by the "MLPG+GV" module in the online stage. After conducting step 6), 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 pair is kept and used to construct the paired dictionaries. The necessity of this strategy has been confirmed in LLE-based SC [12, 14].

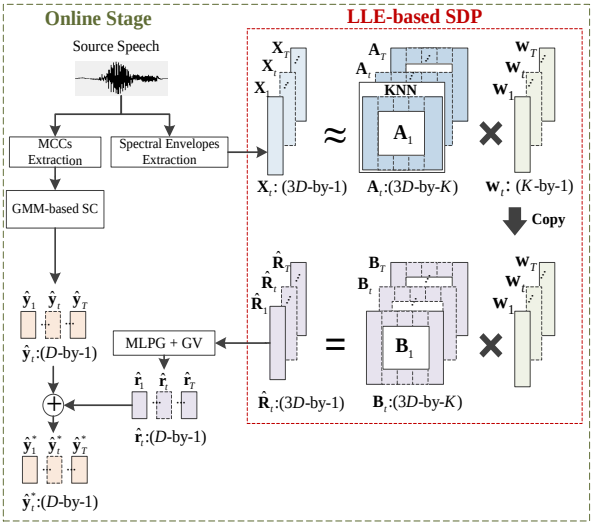

Figure 1: *Overview of the run-time conversion stage of the proposed spectral detail compensation method.*

2.3. Online stage of the LLE-based SDP method

Given a sequence of source spectral envelope vectors extracted from a source speech, the LLE-based SDP method is applied to convert each source spectral envelope vector to a corresponding target spectral detail vector *independently* in a frame-by-frame manner. Specifically, let $\mathbf{X}_t \in \mathcal{R}^{3D\times 1}$ be the source spectral envelope vector at frame t composed of the D-dimensional static $\mathbf{x}_t$, delta $\Delta^{(1)}\mathbf{x}_t$, and delta-delta $\Delta^{(2)}\mathbf{x}_t$ vectors, i.e., $\mathbf{X}_t = \left[\mathbf{x}_t^{\mathrm{T}}, \Delta^{(1)}\mathbf{x}_t^{\mathrm{T}}, \Delta^{(2)}\mathbf{x}_t^{\mathrm{T}}\right]^{\mathrm{T}}$, where the superscript T denotes transposition. The LLE-based SDP method has three steps. First, a locally linear patch is identified by finding K nearest neighbors (measured by the Euclidean distance) of $\mathbf{X}_t$ from the source dictionary. Second, the local geometry of the locally linear patch is characterized by a reconstruction weight vector estimated by minimizing the reconstruction error ε_t subject to the constraint $\mathbf{1}^{\mathrm{T}}\mathbf{w}_t = 1$ (for the purpose of translational invariance) at frame t:

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

where $\mathbf{A}_t \in \mathcal{R}^{3D\times K}$ is a matrix (a subset of the source 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. Solving $\mathbf{w}_t$ by minimizing ε_t subject to the constraint is a constrained least square problem, and the solution can be found in [12, 14, 20].

Third, with the assumption that the source spectral envelopes and the target spectral details share a similar local geometry in their respective feature spaces (manifolds), the predicted spectral detail vector $\hat{\mathbf{R}}_t \in \mathcal{R}^{3D\times 1}$ at frame t can be obtained by

$$\hat{\mathbf{R}}_t = \mathbf{B}_t\mathbf{w}_t \qquad (2)$$

where $\mathbf{w}_t$ is the reconstruction weight vector obtained in the second step; $\mathbf{B}_t \in \mathcal{R}^{3D\times K}$ is a matrix (a subset of the target dictionary) corresponding to $\mathbf{A}_t$, and is composed of *K* target spectral detail vectors, 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 spectral detail vector in $\mathbf{B}_t$ corresponding to (aligned with) $\mathbf{a}_{t,k}$.

Once the frame-by-frame prediction process is completed, a sequence of predicted spectral detail vectors $\hat{\mathbf{R}} \in \mathcal{R}^{3D\times T}$ is obtained as $\hat{\mathbf{R}} = \left[\hat{\mathbf{R}}_1,\cdots,\hat{\mathbf{R}}_t,\cdots,\hat{\mathbf{R}}_T\right]$, where *T* is the number of speech frames of the source speech.

2.4. Spectral detail compensation

As shown in Figure 1, to further enhance the predicted spectral details $\hat{\mathbf{R}} \in \mathcal{R}^{3D\times T}$, the "MLPG+GV" method is used (in the same way as it was used in LLE-based SC [12, 14]) to generate a final sequence of static spectral detail vectors $\hat{\mathbf{r}} \in \mathcal{R}^{D\times T}$, i.e., $\hat{\mathbf{r}} = \left[\hat{\mathbf{r}}_1,\cdots,\hat{\mathbf{r}}_t,\cdots,\hat{\mathbf{r}}_T\right]$, where $\hat{\mathbf{r}}_t \in \mathcal{R}^{D\times 1}$ is the final spectral detail vector at frame *t*.

Finally, the sequence of the converted spectral envelopes, i.e., $\hat{\mathbf{y}} = [\hat{\mathbf{y}}_1,\cdots,\hat{\mathbf{y}}_t,\cdots,\hat{\mathbf{y}}_T]$, obtained by the GMM-based SC method, and the sequence of the spectral detail vectors, i.e., $\hat{\mathbf{r}} = \left[\hat{\mathbf{r}}_1,\cdots,\hat{\mathbf{r}}_t,\cdots,\hat{\mathbf{r}}_T\right]$, given by the LLE-based SDP method and the "MLPG+GV" method, are added to generate the final sequence of the converted spectral envelope vectors. That is, $\hat{\mathbf{y}}^* = \hat{\mathbf{y}} + \hat{\mathbf{r}}$, where $\hat{\mathbf{y}}^* = [\hat{\mathbf{y}}_1^*,\cdots,\hat{\mathbf{y}}_t^*,\cdots,\hat{\mathbf{y}}_T^*]$ is the final sequence of the converted spectral envelope vectors, and $\hat{\mathbf{y}}_t^* \in \mathcal{R}^{D\times 1}$ is the final converted spectral envelope vector at frame *t*. The missing spectral details in the converted spectral envelopes $\hat{\mathbf{y}}$ are compensated by $\hat{\mathbf{r}}$.

3. Experiments

3.1. Experimental setup

Our experiments were conducted on the Sinica COSPRO speech corpus [21]. 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 16 kHz/16 bit format. Silence segments at the start and end of each utterance in the training set were discarded based on the segmentation information in the corpus.

The STRAIGHT vocoder [1] was employed to extract the smoothed spectral envelopes, aperiodicity components (AP), and pitch contours (F0), at 5 milliseconds steps. The FFT length was set to 1024; thus, the AP and spectral envelopes for each frame were 513-dimensional. The SC systems compared in the experiments operated on either spectral envelopes and/or MCCs (extracted from the spectral envelopes) to obtain the converted spectral envelopes. The same linear mean-variance transformation was used for F0 conversion. The source speaker's energy and AP were kept unmodified. Therefore, the converted spectral envelopes, converted F0, and source speaker's AP were passed to the STRAIGHT vocoder for waveform reconstruction. We compared two SC systems:

- ***GMM*** (Baseline): The baseline GMM-based SC method integrated with both MLPG and GV algorithms [5].
- ***GMM-SDC*** (Proposed): The proposed method that combines the baseline ***GMM*** system with the spectral detail compensation method.

For the baseline ***GMM*** system, the number of mixture components was set to 64, according to the objective scores and informal listening test measured on the development set. A cross-diagonal covariance matrix was used in the JDGMM. The spectral features were the first through 24th MCCs extracted from the STAIGHT spectral envelopes. The static, delta, and delta-delta features were used. Accordingly, the dimensionality of a final MCC vector was 72.

The proposed ***GMM-SDC*** system was built on top of the baseline ***GMM*** system. The spectral features used in the LLE-based SDP method were the 513-dimensional (log energy-normalized) spectral envelopes. Specifically, each frame of STRAIGHT spectral envelopes 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 spectral envelope value. Moreover, the static, delta, and delta-delta features were used. Accordingly, the dimensionality of a final spectral envelope vector was 1539. After SC, the log energy-normalized spectral envelopes were reverted back to the (linear) spectral envelopes, and the energy was compensated back to the spectral envelopes according to the energy normalizing factor. The number of nearest neighbors (*K* in (1) and (2)) was set to 1024, according to the computational complexity, objective scores, and informal listening test measured on the development set.

3.2. Objective evaluations

The objective evaluation was conducted on the test set in terms of the modulation spectrum (MS) [6]. The MS of an acoustic feature sequence is defined as the log-scaled power spectrum of the sequence. Therefore, the MS can be used to measure the temporal fluctuation of the feature sequence. Different from the previous study [6], we measured the MS of each dimension of the STRAIGHT spectral envelope sequence rather than the MS of each dimension of the MCC vector sequence since the proposed method is applied to compensate missing details in the converted spectral envelopes. Figure 2 shows the average modulation spectra of the 50th, 250th, and 450th dimensions (frequency bins) of the sequences of STRAIGHT spectral envelopes for 43 natural target speech utterances and the corresponding converted spectral envelopes by ***GMM*** and ***GMM-SDC***, respectively. From Figure 2, we observe that the MSs of the 250th and 450th dimensions (i.e., the higher frequency bins) of the converted spectral envelopes by ***GMM-SDC*** are closer to the MSs of the natural target speech. On the other hand, the MSs of the 50th dimension (i.e., the lower frequency bin) of the converted spectral envelopes obtained by ***GMM*** and ***GMM-SDC*** are similar and far apart from the MS of the natural target speech. This result implies that introducing the spectral detail compensation method to the baseline GMM-based SC system can effectively enhance the MSs of higher frequency bins. Figure 3 shows the

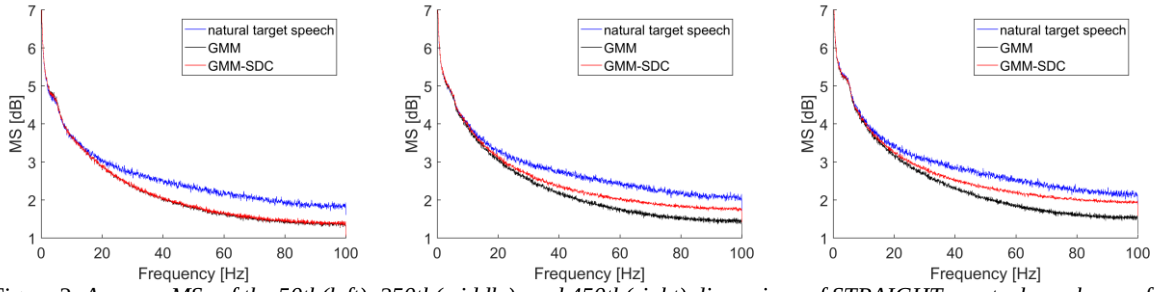

Figure 2: *Average MSs of the 50th(left), 250th(middle), and 450th(right) dimensions of STRAIGHT spectral envelopes of natural target speech and the converted speech by **GMM** and **GMM-SDC**.*

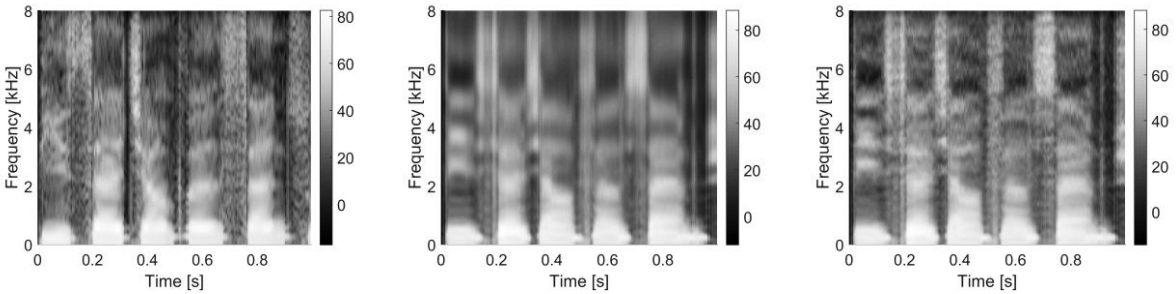

Figure 3: *STRAIGHT spectrograms of a natural target speech (left) and the converted speeches by **GMM** (middle) and **GMM-SDC** (right).*

STRAIGHT spectrograms of a natural target speech and the converted speeches by ***GMM*** and ***GMM-SDC***. From the figure, we observe that more spectral details are revealed in the converted speech by ***GMM-SDC***, particularly in the higher frequency bins (4~8kHz).

3.3. 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 pair, eight sentences were randomly selected from the test set, thereby resulting in 64 (8x8) test sentences. Ten Chinese-native listeners were recruited to conduct the naturalness and speaker similarity tests.

First, we conducted a preference test to evaluate the naturalness of the converted speech. Specifically, 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 1 shows the overall average results of the preference test. We can see that ***GMM-SDC*** yields remarkable gains over the baseline ***GMM*** system in terms of naturalness. The result is consistent with those of the objective tests in Figures 2 and 3.

Next, we conducted an ABX test for each system independently to evaluate the speaker similarity performance. The natural source and target speeches were presented to the listeners in a random order as A and B, and the corresponding converted speech was presented as X. The same sentence was used for A and B, and a different one was used for X to prevent the listeners from evaluating only a specific prosodic pattern of each utterance [5]. Listeners were asked to judge whether utterance X sounded 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 [11]. The third row of Table 1 shows the overall

Table 1: *Test results (%) of naturalness and speaker similarity. P is the P-value of the t-test*

	GMM	***GMM-SDC***	*p*
Naturalness	20.78	**79.22**	0.000
Similarity	86.88	**87.81**	0.975

average results of the ABX test. We note that there are no significant differences between ***GMM*** and ***GMM-SDC*** in terms of similarity. The results of both objective and subjective tests confirmed that the proposed spectral detail compensation method could effectively reintroduce the spectral details to the converted spectral envelopes obtained by a GMM-based SC system, thereby obtaining better naturalness while maintaining speaker similarity.

4. Conclusions

In this paper, we have proposed an exemplar-based spectral detail compensation method for improving spectral conversion methods operating on low-dimensional spectral features. The key idea of the proposed method is to compensate the loss of spectral details in the (STRAIGHT) spectral envelopes reconstructed from the converted low-dimensional spectral features. The GMM-based SC (using MCCs) method was adopted as a case study. Experimental results revealed that the proposed method could effectively reintroduce the spectral details to the converted spectral envelopes obtained by the GMM-based SC method, thus resulting in notably better naturalness. It is expected that the proposed method may improve any SC methods operating on low-dimensional spectral features using the STRAIGHT vocoder.

5. Acknowledgements

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

6. References

[1] 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.

[2] 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.

[3] 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.

[4] A. Kain, and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, 1998, pp. 285-288.

[5] 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.

[6] 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.

[7] Z. Wu, E. S. Chng, and H. Li, "Conditional restricted Boltzmann machine for voice conversion," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process. (ChinaSIP)*, 2013, pp. 104–108.

[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, Dec. 2014.

[9] 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," in *Proc. INTERSPEEH*, 2017.

[10] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, 2012, pp. 313-317.

[11] 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.

[12] 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.

[13] K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda, " The NU-NAIST voice conversion system for the Voice Conversion Challenge 2016," in *Proc. INTERSPEECH*, 2016, pp. 1667-1671.

[14] H. T. Hwang, Y. C. Wu, Y. H. Peng, C. C. Hsu, Y. Tsao, H. M. Wang, Y. R. Wang, and S. H. Chen, "Voice conversion based on locally linear embedding," to appear in *Journal of Information Science and Engineering*.

[15] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," in *Proc. INTERSPEECH*, 2017, pp. 1138-1142.

[16] K. Kobayashi, T. Toda, and S. Nakamura, "F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential," in *Proc. SLT*, 2016, pp. 693-700.

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

[18] Y. C. Wu, H. T. Hwang, S. S. Wang, C. C. Hsu, Y. Tsao, and H. M. Wang, "A post-filtering approach based on locally linear embedding difference compensation for speech enhancement," in *Proc. INTERSPEECH*, 2017.

[19] H. T. Hwang, Y. C. Wu, S. S. Wang, C. C. Hsu, Y. Tsao, H. M. Wang, Y. R. Wang, and S. H. Chen, "Locally linear embedding based post-filtering for speech enhancement," to appear in *Journal of Information Science and Engineering*.

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

[21] 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.