An exemplar-based approach to frequency warping for voice conversion

Tian, Xiaohai; Lee, Siu Wa; Wu, Zhizheng; Chng, Eng Siong; Li, Haizhou

2017

Tian, X., Lee, S. W., Wu, Z., Chng, E. S., & Li, H. (2017). An exemplar-based approach to frequency warping for voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10), 1863-1876. doi:10.1109/TASLP.2017.2723721

<https://hdl.handle.net/10356/89630>

<https://doi.org/10.1109/TASLP.2017.2723721>

© 2017 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: [<http://dx.doi.org/10.1109/TASLP.2017.2723721>].

Downloaded on 26 Apr 2023 19:24:13 SGT

An Exemplar-based Approach to Frequency Warping for Voice Conversion

Xiaohai Tian, Siu Wa Lee, Zhizheng Wu, Eng Siong Chng, *Senior Member, IEEE*, and Haizhou Li, *Fellow, IEEE*

Abstract—The voice conversion’s task is to modify a source speaker’s voice to sound like that of a target speaker. A conversion method is considered successful when the produced speech sounds natural and similar to the target speaker. This paper presents a new voice conversion framework in which we combine frequency warping and exemplar-based method for voice conversion. Our method maintains high-resolution details during conversion by directly applying frequency warping on the high-resolution spectrum to represent the target. The warping function is generated by a sparse interpolation from a dictionary of exemplar warping functions. As the generated warping function is dependent only on a very small set of exemplars, we do away with the statistical averaging effects inherited from Gaussian mixture models (GMM). To compensate for the conversion error, we also apply residual exemplars into the conversion process. Both objective and subjective evaluations on the VOICES database validated the effectiveness of the proposed voice conversion framework. We observed a significant improvement in speech quality over the state-of-the-art parametric methods.

Index Terms—voice conversion, exemplar, sparse representation, frequency warping, residual compensation

I. INTRODUCTION

VOICE conversion (VC) is a technique to modify one voice (source) to sound like other person (target) without changing the content. In a typical voice conversion system, a conversion function is first estimated from parallel utterances spoken by source and target speakers. At run-time, this function is then applied to convert the given source speech to the target. The success of the conversion depends on how effective the conversion function transforms the source speaker’s characteristics to the target. The characteristics capturing a speaker’s identity include formant structure and prosody. In this work, we focus on spectral mapping to transform the formant structure.

A. Related Works

There exists a number of parametric methods to realize both the linear and nonlinear conversion functions. For example, Gaussian mixture models (GMM) [1]–[3], linear regression models [4], and partial least squares regression [5] assume that source and target spectral features are linearly related.

However, dynamic kernel partial least squares regression (DKPLS) [6], neural network based approaches [7]–[10] and restricted Boltzmann machine (RBM) [11]–[13] assume non-linear relationship. To reduce the required amount of training data, maximum a posterior (MAP) adaptation [14] and eigen-voice technique [15] were also employed for voice conversion.

While the parametric methods could generate speech highly similar to the target, they are known to produce over-smoothed speech trajectories. This is because they usually use the minimum mean square error [1], [6], [7] or the maximum likelihood [3] function as the optimization criteria. As a result, statistical averaging arises and the converted speech fails to capture the desired details of temporal and spectral dynamics. Additionally, parametric methods generally employ low-resolution features such as the Mel cepstral coefficients (MCC) or line spectral frequencies (LSF) to avoid the curse of dimensionality. Such low-resolution features, however, are doomed to lose spectral details. These two aspects, statistical averaging and low-resolution features, cause the converted speech of most parametric methods to sound muffled [3], [16]. To enhance the temporal variation and alleviate the over-smoothing problem, global variance (GV) [3], [17] and modulation spectrum (MS) [18] were proposed. Although improvements were observed, the problem is only partially resolved as the low-resolution features are still applied in these solutions.

To preserve spectral details during conversion, a number of frequency warping-based methods were introduced. The frequency warping (FW) directly transforms the high-resolution source spectrum to that of the target speaker through a warping function on the frequency axis. In recent literature, the warping function is either realized by a single parameter, such as VTLN-based approaches [19]–[24], or represented as a piecewise linear function [16], [25], [26]. The studies [19], [20], [27] show that the single parameter VLTN approach may not adequately describe the mapping of spectral details. Hence, the piecewise linear warping function has become a mainstream solution.

The goal of piecewise linear warping function is to align a set of frequencies between the source and target spectrum by minimizing the spectral distance or maximizing the correlation between the converted and target spectrum. The set of frequencies can be all the frequency bins or some pivot frequencies. In [25], [28], dynamic frequency warping (DFW) was proposed to align all the frequency bins. As a variant of DFW, automatic mapping of formants (AMF) [16] was proposed to align the formants. In [26], the frequencies were aligned based on histograms of spectral peaks. In [29],

Xiaohai Tian and Eng Siong Chng are with the School of Computer Science and Engineering and Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), Nanyang Technological University, Singapore (e-mail: {xhtian, aseschng}@ntu.edu.sg). Siu Wa Lee is with the Institute for Infocomm Research, A*STAR, Singapore (e-mail: swylee@i2r.a-star.edu.sg). Zhizheng Wu is with the center for speech technology research, university of Edinburgh, United Kingdom (e-mail: zhizheng.wu@ed.ac.uk). Haizhou Li is with the Department of Electrical and Computer Engineering, National University of Singapore (e-mail: haizhou.li@nus.edu.sg).

a correlation-based frequency warping (CFW) method was studied to allow more flexible alignment between the source and target formants, and is therefore less sensitive to formant estimation errors.

Although frequency warping can somewhat transform the source spectrum to target, it is not perfect either. First, the same warping function is shared within an acoustic class [25], that leads to the discontinuity of converted spectra across classes. To mitigate this, the weighted frequency warping (WFW) [16] generates the warping function by interpolating warping functions from different acoustic classes. Second, frequency warping itself only transforms the frequency axis without considering the amplitude of the spectrum. To address this, amplitude scaling (AS) [24], [26], also known as residual compensation (RC) in [30]¹, were proposed. The amplitude difference between source and target spectrum is modelled for each acoustic class by a GMM. This also results in the statistical averaging which causes inaccurate target spectrum estimation [30].

Recently, exemplar-based sparse representation was studied [32]–[34]. It also shows the potential in achieving good performance in voice conversion task [35]–[39]. The approach generates the target speech using a weighted linear combination of a set of exemplars from the training data of the target speaker. The set of exemplars is also known as a dictionary in this reserved field. During dictionary construction, the paired source-target dictionary is established by aligning source to target spectra of a parallel corpus. At run-time conversion, the nonnegative matrix factorization (NMF) [40] is employed to estimate the combination weights. With the sparsity constraint on the combination weights, the over-smoothing problem is greatly reduced. As the converted speech is generated from the target exemplars, which are high-resolution features (i.e. spectrum), this approach produces relatively high quality speech.

In [37], it has been shown that the exemplar-based sparse representation did not perform as well as the GMM method with global variance. However, in the setting where only a small amount of training data is available, the GMM-based method will not be able to robustly model the high-resolution spectral mapping, and as such, the exemplar-based approach offers an effective solution to voice conversion [30] with limited training data.

B. Contributions of This Work

In [30], we introduced a voice conversion framework based on exemplar-based frequency warping and residual compensation. In this framework, instead of estimating the target spectral envelope directly [37], we use the exemplar-based technique to estimate the warping functions and residual vectors. In doing so, we avoid the statistical averaging problem inherited from GMM-based approaches [16] and provide two major advantages: 1) under exemplar-based frequency warping (EFW) framework, each warping function is generated from

a dictionary of warping function exemplars. Specifically, the warping function is realized as a weighted linear combination of exemplars with sparsity constraint. With such constraint, the over-smoothing problem of warping function in GMM-based framework [16] is reduced. 2) We propose an exemplar-based residual compensation (ERC) method, where the spectral residuals are computed frame by frame, as opposed to class by class in the GMM-based framework [16]. This provides a more accurate residual compensation.

In this work, we further the research in [30] with the following contributions. First, we provide more explanations and detailed algorithms of the exemplar-based frequency warping (EFW) framework [30]. Second, we provide an analytical study on EFW framework in the context of the state-of-the-art. 1) We present detailed evaluations with and without residual compensation under both exemplar- and GMM-based frequency warping frameworks. 2) We examine the effectiveness of proposed exemplar-based residual compensation (ERC) over GMM-based residual compensation (GRC). 3) We include the state-of-the-art baselines to further confirm the effectiveness of the proposed EFW framework.

In the following section, we briefly present the conventional frequency warping techniques, the GMM-based frequency warping framework and its existing weakness. In section III, we then present our proposed voice conversion framework.

II. LIMITATIONS OF GMM-BASED WEIGHTED FREQUENCY WARPING FRAMEWORK

In voice conversion, a feature mapping function $\mathcal{F}(\cdot)$ is used to perform the transformation of the source feature vector $\mathbf{x}$ to the target feature $\mathbf{y}$. This can be expressed as

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e}, \quad (1)$$

where $\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}, \mathbf{e} \in \mathbb{R}^M$, M is the feature dimension, $\hat{\mathbf{y}}$ is the converted feature, and $\mathbf{e}$ is the approximation error. The mapping function can be estimated in different ways, i.e. statistical parametric mapping [1], [3], [6] and frequency warping [16], [25].

This work focuses only on the frequency warping techniques. In the following subsections, we first present a brief overview of three conventional frequency warping techniques used in our system and then highlight the existing weakness of the GMM-based frequency warping framework.

A. Conventional Frequency Warping

Piecewise-linear based frequency warping technique aims to align a pair of source-target spectrum with an invertible warping function $\mathbf{w}$ defined by multiple linear segments [25]. We estimate a target spectrum $\hat{\mathbf{y}}$ by warping the source spectrum $\mathbf{x}$ as follows

$$\hat{\mathbf{y}} = \mathcal{F}(\mathbf{x}) = \mathbf{w}(\mathbf{x}), \quad (2)$$

where $\mathbf{x}$ and $\hat{\mathbf{y}}$ are high-resolution spectra, $\mathbf{w} \in \mathbb{R}^M$ is a vector which also serves as the warping function to evaluate $\hat{\mathbf{y}}$ by $\mathbf{w}(\mathbf{x})$.

Frequency warping method is characterized by its optimization criterion. Here we review three well-known formulations.

¹As the main idea is to compensate the amplitude difference between the warped spectrum and the target spectrum rather than to scale the spectrum, we call the technique residual compensation in this work. Residual refers to the amplitude difference between the warped spectrum and the target spectrum. In some studies, it is also called spectrum differential [31].

1) *Dynamic Frequency Warping*: Dynamic frequency warping (DFW) was originally proposed in [25]. Given a pair of source-target spectrum, $\mathbf{x}$ and $\mathbf{y}$, the warping function is optimized to minimize the spectral distance over all frequency bins between source and target spectrum. If the search space is not constrained, one-to-many and many-to-one mapping of frequency bins can happen, which may result in an over-stretched or over-compressed alignment.

2) *Automatic Mapping of Formants*: Automatic mapping of formants (AMF) [16] is a variant of the piecewise linear frequency warping approach. AMF can be regarded as a special case of DFW. Instead of aligning all the frequency bins, AMF aligns only at the formant frequencies, which are calculated in advance from line spectral frequencies (LSF) vectors. Between the formant frequencies, linear alignment is assumed. In this way, the search space of AMF is reduced.

In practice, given two LSF vectors and their corresponding source and target spectra, the frequencies of poles (depicting the formant positions) are first calculated and kept in an ascending order. Then both source and target spectra are split into segments separated by the formant frequencies. Finally, dynamic programming is applied to search for the formant pairs by minimizing the spectral distance.

Operating on formant frequencies, AMF reduces the over-stretched or over-compressed problems caused by DFW to obtain a better warping function [30]. However, the AMF algorithm is highly dependent on an accurate formant detection.

3) *Correlation-based Frequency Warping*: In [29], we propose a correlation-based frequency warping (CFW) to improve AMF, where we allow for a more relaxed alignment between the source and target around the formants. Additionally, the correlation coefficient is employed as the optimization criterion to estimate the warping functions, instead of the spectral distance.

Fig. 1: Illustration of the key difference between AMF and CFW method to perform warping. The black dots and the red crosses at f_2 indicate the estimated and actual formant positions, respectively. (a) AMF algorithm assumes that source and target formants f_1 , f_2 and f_3 are given. The warping process searches for the best alignment between source and target formant sequences. (b) As CFW allows source-target formants to be aligned within a range of frequencies, it is less sensitive to inaccurate estimation of f_2 in the source spectrum.

The CFW algorithm is different to AMF as it allows more matching candidates to be around formant positions, as

shown in Fig. 1. Such a relaxation improves the warping as it makes the alignment less sensitive to formant estimation errors. Moreover, instead of minimizing the spectral distance, the warping function of CFW is defined by maximizing the correlation coefficient between the source and target spectra. During the warping process, the source and target spectra are split into segments by formant frequencies. The correlation coefficient between the l^{th} source-target spectral segments s_l and t_l , is calculated by $\rho(s_l, t_l) = \text{cov}(s_l, t_l) / \sqrt{\text{var}(s_l) \cdot \text{var}(t_l)}$, where $\text{cov}(s_l, t_l)$ denotes the covariance between s_l and t_l ; $\text{var}(\cdot)$ denotes the variance. The details of CFW warping process can be found in [41], [42] and formants estimation in [16].

B. Limitation and weakness of GMM-based Weighted Frequency Warping and Residual Compensation methods

In the original frequency warping method [25], the warping function is trained by spectral class. To avoid hard decisions, the weighted frequency warping (WFW) used the Gaussian mixture model (GMM) to derive frequency warping functions from soft partitions [16]. In this way, WFW generates smooth warping functions across frames to avoid abrupt change between acoustic classes. In [26], the GMM is also used to derive the residual spectrogram to further improve the converted speech quality.

1) *GMM-based Weighted Frequency Warping*: In the implementation of [16], two types of feature representations are used, the low-resolution features are typically Mel cepstral coefficients (MCC) or line spectral frequencies (LSF), and the high-resolution feature is the spectrum. The low-resolution features are used for GMM modeling and deriving the warping functions and the high-resolution features are used to perform run-time conversion to keep the spectral details.

Given the aligned source-target spectra pairs, we first learn a GMM with K Gaussian components. Then the K warping functions are derived from the mean vectors of corresponding Gaussian components, denoted as $\{\mathbf{w}_1, \dots, \mathbf{w}_k, \dots, \mathbf{w}_K\}$.

At run-time, for each input frame $\mathbf{x}$, its corresponding weighted warping function $\mathbf{w} \in \mathbb{R}^M$ is estimated by

$$\mathbf{w} = \sum_{k=1}^K p_k(\mathbf{x}) \mathbf{w}_k, \quad (3)$$

where $p_k(\mathbf{x})$ and $\mathbf{w}_k \in \mathbb{R}^M$ are the posterior probability and the corresponding warping function of $\mathbf{x}$ in mixture k .

2) *GMM-based Residual Compensation*: With $\mathbf{w}$, the warped spectrum $\hat{\mathbf{y}}$ can be obtained by Eq. (2). To compensate for the amplitude difference between $\hat{\mathbf{y}}$ and the reference target spectrum, a GMM-based residual compensation (GRC) technique was proposed in [26].

Let $\mathbf{x}_n$, $\mathbf{y}_n$ and $\hat{\mathbf{y}}_n$ be the n^{th} frame of the source, target and estimated target in training data, respectively. The residual spectrum $\mathbf{r}_n \in \mathbb{R}^M$ can be computed in log-scale by

$$\log \mathbf{r}_n = \log \mathbf{y}_n - \log \hat{\mathbf{y}}_n. \quad (4)$$

Then, the residual compensation vector $\mathbf{r} \in \mathbb{R}^M$ for input frame $\mathbf{x}$ is calculated in log-scale by

$$\log \mathbf{r} = \sum_{k=1}^K p_k(\mathbf{x}) \left[\sum_{n=1}^N \log \mathbf{r}_n \cdot \gamma_{n,k} \right], \quad (5)$$

where N is the number of aligned training frames, $\gamma_{n,k}$ is the occupation probability of the n^{th} frame belonging to the k^{th} Gaussian component. $\sum_{n=1}^N \log \mathbf{r}_n \cdot \gamma_{n,k}$ is the mean residual spectral vector of k^{th} Gaussian component in log-scale.

A compensated warped spectrum is then obtained by

$$\tilde{\mathbf{y}} = \exp(\log \hat{\mathbf{y}} + \log \mathbf{r}). \quad (6)$$

Under the GMM-based framework, the estimation of warping function Eq. (3) and residual compensation vector Eq. (5) suffers from statistical averaging due to posterior probability $p_k(\mathbf{x})$, which is known to be a source of over-smoothing [37].

III. EXEMPLAR-BASED FREQUENCY WARPING WITH RESIDUAL COMPENSATION

Extending the work in [30], we propose an exemplar-based frequency warping framework (EFW) with residual compensation. The framework consists of two modules: a) dictionary construction; b) run-time conversion. Fig. 2 illustrates the workflow of the proposed framework. During dictionary construction, the source and target parallel utterances are first aligned to construct the paired source-target dictionary on the high-resolution spectrum. For each source feature vector, also known as exemplar in this work, we construct a warping function to warp the source to target spectrum. The residual error after the warping is retained and kept as an exemplar's residual vector. Hence, three dictionaries are established in the dictionary construction process, namely, the source, warping function and residual dictionaries. At run-time conversion, the input spectrum is first decomposed into a linear combination of exemplars from the source dictionary. The combination weights are constrained to be sparse to reduce statistical averaging. These weights are further used to generate an interpolated warping function and spectral residual for each input spectrum. Finally, the interpolated warping function and spectral residual are applied to the input spectrum to generate the output spectrum, which is then passed through a synthesis filter to reconstruct the target speech.

Our framework benefits from three unique properties, 1) we use sparse representation to avoid the over-smoothing issue in GMM-based framework; 2) the warping functions are applied on high-resolution spectrum and across multiple frames to maintain both spectral and temporal quality; 3) the exemplar-based residual compensation is applied to improve the high-resolution spectrum reconstruction process. Similar to the GMM-based frequency warping framework, we avoid hard clustering by estimating the warping functions over multiple frames to ensure a smooth spectral trajectory [16]. The algorithms and process in Fig. 2 are described in further detail next.

A. Dictionary Construction

Given a collection of source-target parallel speech utterances, we first extract the Mel cepstral coefficients (MCC) and spectrum features. Then, dynamic time warping (DTW) [43] is applied on the MCC features to align the parallel utterances. Using the found frame-wise alignment, we form the paired source-target spectra dictionary $[\mathbf{A}; \mathbf{B}] = [\mathbf{A}^T; \mathbf{B}^T]^T$, where

$\mathbf{A}$ and $\mathbf{B}$ are the source and target dictionaries of dimension $\mathbb{R}^{M \times N}$ respectively, M is the dimension of the spectra, and N is the number of aligned frames. Mathematically

$$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n, \dots, \mathbf{a}_N], \quad (7)$$

$$\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n, \dots, \mathbf{b}_N], \quad (8)$$

where $\mathbf{a}_n, \mathbf{b}_n \in \mathbb{R}^M$ are called exemplars, which are the n^{th} spectra that form the source-target pair. In this work, only the voiced frames are contained in $\mathbf{A}$ and $\mathbf{B}$.

From $[\mathbf{A}; \mathbf{B}]$, we can now derive the warping function dictionary $\mathbf{W}$. Specifically, between each paired exemplars $[\mathbf{a}_n; \mathbf{b}_n]$, a warping function $\mathbf{w}'_n \in \mathbb{R}^M$ can be determined by the frequency warping methods described in Section II-A, such as DFW, AMF and CFW. The set of N warping functions are then used to form a warping function dictionary $\mathbf{W}' \in \mathbb{R}^{M \times N}$.

As the warping functions are estimated from the source and target exemplars independently, these functions may vary widely across frames. It happens even within a continuous voiced segment. Abrupt changes of warping function across adjacent frames may lead to discontinuity in the converted spectrogram, which is not desirable. To reduce such effect, we post-process the warping function dictionary by taking the contextual information into consideration [37]. Specifically, for each segment, the warping functions of $\mathbf{W}'$ are smoothed by a moving average filter (a.k.a. temporal smoothing window). We denote the n^{th} smoothed warping function as $\mathbf{w}_n$. The set of N smoothed warping functions are then used to form a smoothed warping function dictionary $\mathbf{W} \in \mathbb{R}^{M \times N}$, which is denoted as

$$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n, \dots, \mathbf{w}_N]. \quad (9)$$

Fig. 3 shows an example of the warping function dictionaries before and after temporal smoothing. The smoothed warping function dictionary $\mathbf{W}$ is then used for the residual dictionary calculation.

The above warping function dictionary $\mathbf{W}$ establishes an approximate relationship between the source and target exemplars. When we apply the warping function on the source exemplar, there is an amplitude difference, which also called the residual, will be presented between the warped and the target spectrum. To improve the conversion performance, we retain these residual difference between the warped spectrum of n^{th} source exemplar to the reference target exemplar in log-scale, which is in a similar way as Eq. (4). Mathematically, this is

$$\log \mathbf{r}_n = \log \mathbf{b}_n - \log \hat{\mathbf{b}}_n, \quad (10)$$

where $\mathbf{b}_n$ is the n^{th} target exemplar, $\hat{\mathbf{b}}_n$ is the corresponding warped spectrum which is obtained by Eq. (2), and $\mathbf{r}_n \in \mathbb{R}^M$.

To maintain the nonnegative values in expressing the residual spectrum, we form the residual dictionary, $\mathbf{R} \in \mathbb{R}^{M \times N}$, with a set of N residual vectors as a nonnegative matrix

$$\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n, \dots, \mathbf{r}_N]. \quad (11)$$

Hence, at the completion of the dictionary construction stage, the source dictionary $\mathbf{A}$, warping function dictionary $\mathbf{W}$ and residual dictionary $\mathbf{R}$ are formed.

(a) Dictionary construction

(b) Run-time conversion

Exemplar-based frequency warping

Fig. 2: Block diagram of the exemplar-based frequency warping with residual compensation framework. a) During dictionary construction, a source dictionary $\mathbf{A}$, a warping function dictionary $\mathbf{W}$, and a residual dictionary $\mathbf{R}$ are built; b) at run-time conversion, the activation matrix $\mathbf{H}$ is estimated between the run-time speech sample $\mathbf{X}$ and the source dictionary $\mathbf{A}$. The activation matrix $\mathbf{H}$ is then shared by the warping function dictionary $\mathbf{W}$, and the residual dictionary $\mathbf{R}$ when generating the final converted speech.

Fig. 3: Illustration of the effects of temporal smoothing on warping function dictionaries estimated using DFW, AMF and CFW. (a), (c) and (e) shows a part of warping function dictionaries before smoothing, $\mathbf{W}'$, while (b), (d) and (f) are the respective warping function dictionaries after smoothing, $\mathbf{W}$.

B. Run-time Conversion

At run-time, the three constructed dictionaries, $\mathbf{A}$, $\mathbf{W}$ and $\mathbf{R}$, are used for voice conversion as shown in Fig. 2 (b).

Let $\mathbf{x} \in \mathbb{R}^M$ be a speech frame from the source speaker to be converted. We first decompose $\mathbf{x}$ into a linear combination of source exemplars in $\mathbf{A}$ as

$$\mathbf{x} \approx \sum_{n=1}^N h_n \mathbf{a}_n = \mathbf{A}\mathbf{h}, \quad (12)$$

where $\mathbf{h} \in \mathbb{R}^N$ is the activation vector, of which its n^{th} element h_n denotes the weight associated with exemplar $\mathbf{a}_n$ in $\mathbf{A}$.

To estimate $\mathbf{h}$, the nonnegative matrix factorization (NMF) technique [40], [44] is used.

$$\mathbf{h} = \arg \min_{\mathbf{h} \geq 0} d(\mathbf{x}, \mathbf{A}\mathbf{h}) + \lambda \|\mathbf{h}\|_1, \quad (13)$$

where $d(\mathbf{x}, \mathbf{A}\mathbf{h})$ is the Kullback-Leibler (KL) divergence between $\mathbf{x}$ and $\mathbf{A}\mathbf{h}$, λ is the sparsity penalty factor, and $\|\mathbf{h}\|_1$ is L_1 norm on the activation vector. The details of NMF can be found in [44] [37]. In our case, given $\mathbf{A}$ and $\mathbf{x}$, which are matrices of positive values, the estimated $\mathbf{h}$ is also of positive values, such that Eq. (13) is satisfied.

The activation vector $\mathbf{h}$ in Eq. (12) describes how an input speech frame $\mathbf{x}$ can be approximated by a sparse linear combination of source exemplars in $\mathbf{A}$. Because the source and target dictionaries, $\mathbf{A}$ and $\mathbf{B}$, are aligned with parallel content, the activation vector $\mathbf{h}$ can be used to construct $\tilde{\mathbf{y}}$ [35], [37],

$$\tilde{\mathbf{y}} = \mathbf{B}\mathbf{h}. \quad (14)$$

Instead of constructing $\tilde{\mathbf{y}}$ from $\mathbf{B}$ directly, in this work, we estimate the converted spectrum $\hat{\mathbf{y}}$ and residual spectrum $\hat{\mathbf{r}}$ separately. Noted that we have the converted dictionary $\hat{\mathbf{B}}$ and residual dictionary $\mathbf{R}$ in Fig. 2. From Eq. (10) and Eq. (11), the target dictionary $\mathbf{B}$ can be expressed as

$$\log \mathbf{B} = \log \hat{\mathbf{B}} + \log \mathbf{R}, \quad (15)$$

which can be rewritten in linear-scale as

$$\mathbf{B} = \hat{\mathbf{B}} \circ \mathbf{R}, \quad (16)$$

where, $\circ$ denotes the element-wise multiplication. Then, substituting Eq. (16) into Eq. (14) yields

$$\tilde{\mathbf{y}} = (\hat{\mathbf{B}} \circ \mathbf{R})\mathbf{h}. \quad (17)$$

Due to the sparsity constraint of $\mathbf{h}$, Eq. (17) can be further approximated as

$$\tilde{\mathbf{y}} = (\hat{\mathbf{B}} \circ \mathbf{R})\mathbf{h} \approx (\hat{\mathbf{B}}\mathbf{h}) \circ (\mathbf{R}\mathbf{h}). \quad (18)$$

We now express Eq. (18) in log-scale as

$$\log \tilde{\mathbf{y}} \approx \log(\hat{\mathbf{B}}\mathbf{h}) + \log(\mathbf{R}\mathbf{h}), \quad (19)$$

where, $\hat{\mathbf{B}}\mathbf{h}$ denotes the warped spectrum $\hat{\mathbf{y}}$, and $\mathbf{R}\mathbf{h}$ denotes the residual vector $\hat{\mathbf{r}}$. Next we justify the run-time implementation of Eq. (19).

1) *Exemplar-based Frequency Warping*: The warped spectrum $\hat{\mathbf{y}}$ in Eq. (19) can be further written as a linear combination of exemplars of converted dictionary

$$\hat{\mathbf{y}} = \hat{\mathbf{B}}\mathbf{h} = \sum_{n=1}^N h_n [\mathbf{w}_n(\mathbf{a}_n)], \quad (20)$$

where $\mathbf{w}_n(\mathbf{a}_n) \in \mathbb{R}^M$ denotes the n^{th} exemplar in $\hat{\mathbf{B}}$.

As illustrated in Section III-A, the warping function dictionary $\mathbf{W}$ preserves the relationship between $\mathbf{A}$ and $\mathbf{B}$. Therefore, given a speech frame $\mathbf{x}$, it is logical to assume that the same activation vector $\mathbf{h}$ can also be used to interpolate the $\mathbf{W}$ to obtain the warping function $\hat{\mathbf{w}} \in \mathbb{R}^M$ in a similar way as Eq. (3)

$$\hat{\mathbf{w}} = \sum_{n=1}^N h_n \mathbf{w}_n = \mathbf{W}\mathbf{h}. \quad (21)$$

Then, Eq. (20) can be rewritten as

$$\hat{\mathbf{y}} \approx \hat{\mathbf{w}} \left(\sum_{n=1}^N h_n \mathbf{a}_n \right) = \hat{\mathbf{w}}(\mathbf{x}). \quad (22)$$

As each speech frame requires an independent factorization and warping process, for a given input utterance with Q frames of spectrogram, i.e. $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_q, \dots, \mathbf{x}_Q] \in \mathbb{R}^{M \times Q}$, we can construct its corresponding activation matrix $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_q, \dots, \mathbf{h}_Q] \in \mathbb{R}^{N \times Q}$. Mathematically, this is

$$\mathbf{X} \approx \mathbf{A}\mathbf{H}. \quad (23)$$

Thus, the warping functions $\hat{\mathbf{W}}$ of $\mathbf{X}$ is calculated by

$$\hat{\mathbf{W}} = \mathbf{W}\mathbf{H}. \quad (24)$$

Finally, we apply the warping functions $\hat{\mathbf{W}} = [\hat{\mathbf{w}}_1, \dots, \hat{\mathbf{w}}_q, \dots, \hat{\mathbf{w}}_Q] \in \mathbb{R}^{M \times Q}$ on the source spectrogram $\mathbf{X}$ to obtain the warped spectrogram $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_q, \dots, \hat{\mathbf{y}}_Q] \in \mathbb{R}^{M \times Q}$.

2) *Exemplar-based Residual Compensation*: To improve the quality of generated $\hat{\mathbf{Y}}$, the exemplar-based residual compensation (ERC) is applied to compensate the amplitude difference. In view of the intrinsic relationship between the source dictionary $\mathbf{A}$ and residual dictionary $\mathbf{R}$, we use the same activation vector $\mathbf{h}$ in Eq. (12) to derive a spectral residual for input speech frame $\mathbf{x}$. For a given spectrogram $\mathbf{X}$, the residual spectrogram, as expressed in the second term of Eq. (19), can be calculated as

$$\hat{\mathbf{R}} = \mathbf{R}\mathbf{H}. \quad (25)$$

Finally, the converted spectrogram $\tilde{\mathbf{Y}} = [\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_Q] \in \mathbb{R}^{M \times Q}$ is obtained by combining the residual spectrogram $\mathbf{R} \in \mathbb{R}^{M \times Q}$ and the warped spectrogram $\hat{\mathbf{Y}} \in \mathbb{R}^{M \times Q}$ in the log-scale

$$\tilde{\mathbf{Y}} = \exp(\log \hat{\mathbf{Y}} + \log \hat{\mathbf{R}}). \quad (26)$$

C. Exemplar-based Frequency Warping vs. Exemplar-based Sparse Representation

Both the proposed exemplar-based frequency warping method and the exemplar-based sparse representation [37] originate from the exemplar-based idea. Their differences lie in the conversion function and the process.

During dictionary construction, the exemplar-based sparse representation framework [37] establishes a paired source-target dictionary $[\mathbf{A}; \mathbf{B}]$, that defines the frame-to-frame mapping correspondence, while the exemplar-based frequency warping framework establishes a warping function dictionary $\mathbf{W}$ in addition to $\mathbf{A}$ and $\mathbf{B}$.

At run-time conversion, the exemplar-based sparse representation framework [37] estimates the linear combination weights of exemplars $\mathbf{H}$ from the source dictionary. $\mathbf{H}$ is then applied to the target dictionary to realize the conversion. Although the exemplar-based frequency warping framework also estimates $\mathbf{H}$ from the source dictionary, $\mathbf{H}$ is applied to the warping function dictionary to compose $\hat{\mathbf{W}}$, which is then used to generate $\hat{\mathbf{Y}}$.

For residual compensation, the exemplar-based sparse representation framework [37] utilizes two residual dictionaries from both the source and target dictionary. A linear transformation is estimated between these two residual dictionaries. At run-time conversion, the source residual spectrogram is mapped to obtain the target residual spectrogram [37] with the linear transformation. In the exemplar-based frequency warping framework, we only obtain one residual dictionary $\hat{\mathbf{R}}$ during dictionary construction (See Fig. 2). At run-time conversion, we propose a new strategy to obtain the target residual spectrogram by interpolating the residual dictionary $\mathbf{R}$ with the same activation matrix $\mathbf{H}$ used for the warping function.

IV. EXPERIMENTS

In this section, we conducted two sets of experiments to validate the proposed framework. We first conducted experiments to find the optimal configuration. Then, we evaluated the proposed EFW framework with the optimal configuration by comparing it against the state-of-the-art baselines.

TABLE I: The summary of three categories of systems in seven groups. In category (I), we report the state-of-the-art parametric approaches. In category (II), we report a cluster of GMM-based frequency warping techniques with and without residual compensation. In category (III), we compare the proposed EFW approaches with the GMM counterpart in category (II).

Category	Group	Conversion method	Weighted frequency warping	Residual compensation	Number of Gaussians	Smoothing window for WF dictionary	Feature for training	Feature for conversion	LSD ratio (%)
(I) Parametric approaches	(a)	ML-GMM	N/A	N/A	64	N/A	24-dim MCCs with delta	24-dim MCCs with delta	66.38
		ML-GMM-GV			N/A		24-dim MCCs	24-dim MCCs	66.60
		DKPLS			N/A		24-dim MCCs	24-dim MCCs	66.21
(II) GMM-based frequency warping approaches	(b)	GAMF	GMM (Eq. (3))	N/A	32	N/A	14-dim LSFs	513-dim spectrum	97.13
		GDFW							86.34
		GCFW							96.79
	(c)	GAMF-GRC	GMM (Eq. (3))	GMM (Eq. (5))	32	N/A	14-dim LSFs	513-dim spectrum	71.96
		GDFW-GRC							72.25
		GCFW-GRC							71.08
	(d)	GAMF-ERC	GMM (Eq. (3))	Exemplar (Eq. (25))	32	N/A	14-dim LSFs	513-dim spectrum	61.38
		GDFW-ERC							64.73
		GCFW-ERC							61.60
(III) Exemplar-based frequency warping approaches	(e)	EAMF	Exemplar (Eq. (24))	N/A	N/A	15	513-dim spectrum	513-dim spectrum	94.99
		EDFW							94.77
		ECFW							96.57
	(f)	EAMF-GRC	Exemplar (Eq. (24))	GMM (Eq. (5))	32	15	513-dim spectrum	513-dim spectrum	71.80
		EDFW-GRC							73.59
		ECFW-GRC							70.26
	(g)	EAMF-ERC	Exemplar (Eq. (24))	Exemplar (Eq. (25))	N/A	15	513-dim spectrum	513-dim spectrum	60.50
		EDFW-ERC							63.71
		EFW(ECFW-ERC)							58.70

A. Database

The VOICES database [45] was used in all the experiments to report the systems performance. The data from two male (*jal* and *jcs*) and two female (*leb* and *sas*) speakers were used for experiments, that make up 12 speaker pairs, 4 intra-gender and 8 inter-gender conversions. 20 parallel utterances of each speaker were used as training set, other 10 utterances were used as development set, and another 20 utterances were used as evaluation set. All speech signals were sampled at 16kHz and the average length is around 4 seconds. Hence, the total length of 20 training utterances was around 80 seconds.

We used STRAIGHT to extract the 513-dimensional spectrum [46], aperiodicity measures and $\log F_0$ with 5 ms frame shift. The Mel cepstral coefficients (MCC) in 25 dimensions and the line spectral frequencies (LSF) in 14 dimensions were also extracted from the STRAIGHT spectrum.

B. System Setup

We validated the proposed ideas by comparing against three categories of the state-of-the-art techniques. To understand how the component modules work, we also reported the contrastive results on their different combinations. Specifically, twenty one combinations of methods in seven groups were studied under the three categories as summarized in Table I, that includes the GMM method with global variance (GV) enhancement and dynamic kernel partial least square (DKPLS) method. Note that the exemplar-based sparse representation method does not outperform the GMM method with global variance (GV) enhancement [37]. Therefore, exemplar-based sparse representation method was not included as a baseline.

(I) *Parametric approaches*: We implemented the joint-density Gaussian mixture model with maximum likelihood parameter conversion [3] (ML-GMM) and its variant with global variance enhancement (ML-GMM-GV) [47]. ML-GMM-GV used global variance that is derived from the target speaker to reduce the muffled

effects. Both static and its dynamic features were used in these implementations. We also implemented the dynamic kernel partial least squares regression (DKPLS) method [6]. We used the same configuration as in [6], where the context window size is set to 3, the number of latent components and reference vectors are 35 and 200 respectively. ML-GMM-GV and DKPLS represent the state-of-the-art of parametric approaches.

(II) *GMM-based frequency warping approaches*: We implemented the weighted frequency warping method (WFW) [16], where we used LSF features to train GMM and automatic mapping of formants (AMF) to derive the warping functions. We call the first technique GAMF. We also implemented two variants of WFW where AMF was replaced by dynamic frequency warping (DFW) [25] and correlation-based frequency warping (CFW) [29], that we call GDFW and GCFW, respectively. We also studied the effect of GMM-based residual compensation (Eq. (5)) and exemplar-based residual compensation (Eq. (25)), denoted as GRC and ERC, in conjunction with the above three frequency warping schemes, respectively.

(III) *Exemplar-based frequency warping approaches*: We implemented the proposed exemplar-based frequency warping method (EFW), with three types of warping functions, AMF, DFW and CFW. We call these three setups of EFW as EAMF, EDFW and ECFW, respectively. We also studied the effect of GRC and ERC, in conjunction with the above three frequency warping schemes, respectively.

All the conversion methods shared the same frame alignment. We first obtained the alignment by applying the dynamic time warping (DTW) over MCC features. In the aligned frame pairs, some source and target frames were repeated due to the warping effect. We only retained the frame pairs that have both distinctive source or target frames in the dictionary and remove the rest. Note that in parametric approaches, we used all the aligned frames for training. However, in frequency warping approaches, we involved only the voiced frames in

establishing the source-target dictionaries and estimating the warping functions. For different speaker pairs, the number of exemplars ranges between 6,700 and 7,100.

At run-time conversion, the aperiodicity coefficients were carried over from source to target speakers, while the F_0 was converted by a global linear transformation in log-scale [48].

We empirically set the sparsity penalty factor λ to 0.1. To verify the sparsity of the activation weights, we calculated the average top weights over the development utterances for all speaker pairs. For each speaker pair, we sorted the activation weight vectors of frames in a descending order. Then, we averaged the activation weights over all the frames of the development utterances. We observed that, for all the speaker pairs, the top 100 averaged weights accounted for over 75% overall weights. This confirms the effectiveness of the sparsity constraint and is consistent with the results in [37].

As suggested in [37], [49], a spectral compression factor p was used in EFW approaches to maintain the nonnegative nature and emphasize the low intensity frequency bands of the spectral features. From the results on the development data, we empirically set p to 0.4.

C. Evaluation Metrics

Both objective and subjective evaluations were conducted to perform a systematic comparison. Specifically, the objective evaluation was used to measure the spectral distortion between the converted and reference target speech, while the subjective evaluation was used to examine the naturalness and speaker identity respectively.

1) *Objective Evaluation:* The log spectral distortion (LSD) ratio [50] was employed to measure the distortion. The distortion in the q^{th} frame of log spectrum was calculated as

$$d(\mathbf{x}_q, \mathbf{y}_q) = \sum_{i=1}^M (\log \mathbf{x}_{q,i} - \log \mathbf{y}_{q,i})^2, \quad (27)$$

where, M is the total number of the frequency bins. A distortion ratio between the converted-to-target distortion and the source-to-target distortion was defined as

$$\text{LSD} = \frac{\sum_{q=1}^Q d(\tilde{\mathbf{y}}_q, \mathbf{y}_q)}{\sum_{q=1}^Q d(\mathbf{x}_q, \mathbf{y}_q)} \times 100\%, \quad (28)$$

where, $\mathbf{x}_q$ and $\mathbf{y}_q$ denote the source and target spectrum respectively. $\tilde{\mathbf{y}}_q$ is the converted spectrum. The average LSD ratio over all evaluation pairs was reported. A lower LSD ratio indicates a smaller spectral distortion. As the unvoiced frames were not transformed in frequency warping approaches, only voiced frames were involved in the LSD ratio calculation.

2) *Subjective Evaluation:* We also conducted listening tests as the subjective evaluation to confirm the findings in the objective evaluation. We reported the assessment in terms of speech quality and speaker similarity.

Amazon Mechanical Turk (AMT), a widely used crowdsourcing platform [51]–[53], was used in all the listening tests. The evaluation set was constructed by 12 speaker pairs with 20 converted sentences for each pair. In the test, 20 sentences were randomly selected from the 240 testing samples. To prevent cheating, three golden standard pairs were randomly

mixed with the testing samples [53], [37]. 20 paid subjects participated in all the listening tests.

To assess the speech quality, AB preference tests were performed. For each AB test pair, one was from the evaluating method, and another was from its counterpart. They were presented in a random order. A listener was asked to listen to both samples and decide which sample is better in terms of quality.

We followed a similar protocol in [6], [37] to conduct XAB test to assess the speaker similarity for each voice conversion method. Similar to the AB preference test, we randomly selected 20 pairs from the 240 paired samples. In each pair, X was the converted sample, A and B were either the source or target sample. Note that X, A and B shared the same language content. The listeners were asked to listen to the sample X first and then A and B samples were presented in a random order. After listening, the listener should decide the reference sample X is a) closer to A, b) closer to B or c) neither of above, in terms of speaker individuality. Note that the higher the identification rate of a given voice conversion system, the better the similarity to the target speaker is.

T-test with a 95% confidence interval was used in all the subjective evaluations to measure the significance of the differences between different voice conversion methods.

D. Validate the configurations of EFW framework

In this section, we examined four configurations of EFW framework, including the effects of dictionary size, smoothing the warping function dictionary, different frequency warping methods and exemplar-based residual compensation.

Fig. 4: Spectral distortion as a function of dictionary size.

1) *Effects of Dictionary Size:* Addressing the problem of limited training data, we would like to know the effects of different training data size. Note that 20 training utterances offer about 7,000 exemplars in the dictionaries. We studied the effects by reducing the number of exemplars in the dictionaries, that is equivalent to reducing the number of training utterances.

Fig. 4 presents the LSD ratios as a function of the number of the exemplars in the dictionary from the total of about 7,000 to a small sub-sample of 500. It is observed that the LSD ratio decreases as the number of exemplars increases in both development and evaluation data sets.

According to Section III-B, activation matrix estimation is the main computational loads of the EFW framework during

run-time conversion. It is observed in the experiments that the computational cost of this module is almost linear correlated with the size of source and target dictionaries, i.e. the system becomes approximately two times faster, while the dictionaries reduce to half size.

To compare fairly with baseline methods, all exemplars were used in EFW framework in the following experiments.

2) *Effects of Smoothing the Warping Function Dictionary:* As discussed in Section III-A, inter-frame discontinuity may appear in the warping function dictionary $\mathbf{W}'$. Here we proposed a temporal smoothing window to address this problem. To validate our proposal, we conducted experiments using ECFW-ERC to assess the effectiveness of the smoothed warping function dictionary.

Fig. 5: (a) Spectral distortion as a function of smoothing window size for warping function dictionary. (b) Mean square error (MSE) of delta feature as a function of smoothing window size for warping function dictionary.

Objective evaluation. Firstly, we examined the effect of the temporal smoothing window size for the warping function dictionary. The experiments were conducted on both development and evaluation sets. The LSD ratio was measured with different smoothing window sizes from 1 (5ms) to 19 (95ms). As shown in Fig. 5 (a), experiments on both the development and evaluation sets show LSD ratios improve as the smoothing window size increases. After a window size of 15 (75ms), we observe that the LSD ratio plateaus out. As shown in Fig. 5 (b), we also observe the similar trade in delta feature measurement. Therefore, we set the window size as 15 (75ms) in the rest of the experiments.

TABLE II: Spectral distortion of exemplar-based frequency warping methods with and without temporal smoothing (window size 1 vs. 15) on the evaluation set.

Conversion method	Window size	LSD ratio (%)	Window size	LSD ratio (%)
EAMF	1	95.28	15	94.99
EDFW	1	95.20	15	94.77
ECFW	1	96.73	15	96.57
EAMF-ERC	1	63.46	15	60.50
EDFW-ERC	1	70.32	15	63.71
ECFW-ERC	1	61.00	15	58.70

Secondly, we examined the performance of the smoothed warping function dictionary under EFW framework across different frequency warping methods. As shown in Table II, by using the smoothed warping function dictionary, the LSD ratios of all the three variants of EFW methods decrease.

This confirms the need of the temporal smoothing in warping function dictionary construction.

Fig. 6: Preference t-test result of speech quality with 95% confidence intervals for ECFW-ERC with and without smoothing the warping function dictionary.

Subjective evaluation. We also conducted an AB preference evaluation to validate the effectiveness of the smoothed warping function dictionary using ECFW-ERC. We set the smoothing window size to 1 for no inter-frame smoothing (win 1), and 15 for inter-frame smoothing (win 15). As shown in Fig. 6, the ECFW-ERC (win 15) achieves significantly better speech quality than the ECFW-ERC (win 1), with the preference scores of $(55 \pm 4.5)\%$ and $(45 \pm 4.5)\%$ respectively. Both objective and subjective results confirm the effectiveness of smoothing the warping function dictionary in EFW framework. Hereafter, all the experiments were reported with the smoothed warping function dictionary whenever it is used.

3) *CFW vs. AMF and DFW:* We compared the performance of different frequency warping methods, AMF, DFW and CFW, with and without residual compensation. The objective tests were reported in Table I group (b), (c), (e) and (g).

Objective evaluation. We observe that under GMM-based frequency warping framework the three frequency warping techniques produce similar results within group (b) and (c) of Table I, with GDFW achieving the lowest LSD ratio without residual compensation, and GCFW-GRC achieving the lowest LSD ratio with residual compensation. All results suggest that residual compensation contributed significantly while correlation-based frequency warping with residual compensation (GCFW-GRC) achieves the best overall LSD ratio at 71.08%. We apply GMM-based residual compensation (GRC) in group (c) that is formulated in Eq. (5) as part of the GMM-based frequency warping framework [26]. Eq. (3) and Eq. (5) share the same posterior probability of Gaussian components.

A similar pattern is observed within group (e) and (g) of Table I. We apply exemplar-based residual compensation (ERC) in group (g) which is a part of the EFW framework as illustrated in Fig. 2, where Eq. (24) and Eq. (25) share the same activation matrix. In group (e), EDFW achieves the lowest LSD ratio of 94.77% without residual compensation. In group (g), ECFW-ERC achieves the overall lowest LSD ratio of 58.70%.

Results in group (c) and (g) confirm the effectiveness of correlation-based frequency warping.

Subjective evaluation. Two sets of subjective evaluations were performed to assess the speech quality between correlation-based frequency warping method (CFW) and other frequency warping methods, AMF and DFW.

Firstly, we performed AB preference test to assess the speech quality between CFW and other frequency warping methods, AMF and DFW, without residual compensation in group (b) and (e) of Table I. Fig. 7 presents the preference

Fig. 7: Preference test results of speech quality with 95% confidence intervals for GMM-based and exemplar-based frequency warping methods without residual compensation. (a) GAMF vs. GCFW, (b) GDFW vs. GCFW, (c) EAMF vs. ECFW and (d) EDFW vs. ECFW.

Fig. 8: Preference test results of speech quality with 95% confidence intervals for GMM-based and exemplar-based frequency warping methods with residual compensation. (a) GAMF-GRC vs. GCFW-GRC, (b) GDFW-GRC vs. GCFW-GRC, (c) EAMF-ERC vs. ECFW-ERC and (d) EDFW-ERC vs. ECFW-ERC.

scores with 95% confidence intervals. Fig. 7 (a) and Fig. 7 (b) suggest that the speech quality of GCFW outperforms that of GAMF and GDFW under the GMM-based frequency warping framework. Similar results are also observed in Fig. 7 (c) and Fig. 7 (d), which suggest that ECFW outperforms EAMF and EDFW under the exemplar-based frequency warping framework as well.

Secondly, we performed AB preference test to assess the speech quality between CFW and other frequency warping methods with residual compensation in group (c) and (g) of Table I. We observe in Fig. 8 that CFW consistently outperforms AMF and DFW in all the tests, although the difference between EAMF-ERC ($(47.5 \pm 3.35)\%$) and ECFW-ERC ($(52.5 \pm 3.35)\%$) is not significantly. The results are also consistent with those reported in the objective evaluation (See Table I).

4) *GMM-based vs. Exemplar-based Residual Compensation*: We consider that the exemplar-based residual compensation provides a more detailed modelling than the GMM-based residual compensation. To validate the idea, we examined the two residual compensation methods in combination with both

GMM-based and exemplar-based frequency warping frameworks, and reported the objective tests in group (c), (d), (f) and (g) of Table I.

Objective evaluation. We observe that the exemplar-based residual compensation method (ERC) outperforms its GMM-based counterpart (GRC) across all experiments, by around 10% in terms of LSD ratio reduction.

Fig. 9: Preference test results of speech quality with 95% confidence intervals for GMM-based and exemplar-based residual compensation with GCFW and ECFW systems.

Subjective evaluation. We further performed AB preference test to compare the two residual compensation methods under GMM-based and exemplar-based frequency warping frameworks. Note that CFW achieves consistently good results in previous experiments. Here we only implemented GCFW and ECFW in this study. We observe in Fig. 9 (a) that the GCFW-GRC outperforms GCFW-ERC which contradicts the objective evaluation results in group (c) and (d) of Table I, where GCFW-ERC has a lower LSD ratio. We understand that subjective evaluation provides us a different perspective from objective evaluation. We also observe in Fig. 9 (b) that ECFW-ERC outperforms ECFW-GRC, but their preference scores fall into each other's confidence intervals ($(51 \pm 4.16)\%$ and $(49 \pm 4.16)\%$), which means they are not statistically significant. The subjective evaluation seems favor the implementation where the warping function and the residual compensation share the same linear combination weights. Specifically, we have the posterior probability of Gaussian components as the linear combination weights in the GMM-based framework, and the activation matrix in the exemplar-based framework.

E. EFW Framework vs. State-of-the-Arts

In this section, we validated the proposed EFW framework with optimal configuration against the baseline methods in terms of both speech quality and speaker similarity.

1) *Speech Quality Evaluation*: In this paper, we advocated the use of 1) exemplar-based frequency warping (EFW) method, and 2) the exemplar-based residual compensation (ERC) method. We have confirmed that the exemplar-based residual compensation is superior to the GMM-based baselines. In this section, we conducted objective and subjective evaluations to assess the speech quality between the proposed EFW method with the state-of-the-art baselines.

Objective evaluation. From Table I, we clearly see that all EFW methods (Table I group (g)) outperform their GMM-based counterparts (Table I group (c)), with ECFW-ERC giving the lowest LSD ratio of 58.70%. The results suggest that exemplar-based frequency warping and residual compensation

methods offer a clear advantage over GMM-based methods individually and collectively.

It is worth noting that the ECFW-ERC implementation outperforms the state-of-the-art parametric approaches, ML-GMM, ML-GMM-GV and DKPLS, as reported in Table I.

Subjective evaluation. Then we conducted subjective evaluations to assess the speech quality between the proposed EFW method with the state-of-the-art baselines.

Fig. 10: Preference test results of speech quality with 95% confidence intervals for GMM-based and exemplar-based frequency warping frameworks.

First, subjective evaluations were conducted to assess the speech quality between the GMM-based and the exemplar-based frequency warping frameworks. In Section IV-D4, experiments suggest that it is advantageous for warping function and residual compensation to share the same framework, either GMM-based or exemplar-based. Here we only evaluated the three typical comparative pairs, GAMF-GRC vs. EAMF-ERC, GDFW-GRC vs. EDFW-ERC and GCFW-GRC vs. ECFW-ERC. We observe in Fig. 10 that for all the three frequency warping methods, AMF, DFW and CFW, the exemplar-based framework performs significantly better than GMM-based framework. These results also resonate with the objective evaluations in Table I. Hence, we conclude that exemplar-based frequency warping framework is superior to GMM-based frequency warping framework in terms of speech quality.

Finally, we compared the ECFW-ERC with the statistical parametric methods.

Fig. 11: Preference test results of speech quality with 95% confidence intervals for ECFW-ERC vs. parametric methods ML-GMM, ML-GMM-GV and DKPLS.

In [37], exemplar-based sparse representation method was benchmarked against ML-GMM, ML-GMM-GV and DKPLS. Similarly, we would like to conduct AB preference tests between the proposed exemplar-based frequency warping method

with these three baselines. The preference scores with 95% confidence intervals reported in Fig. 11 clearly show that the proposed ECFW-ERC outperforms ML-GMM, ML-GMM-GV and DKPLS by a large margin, while the exemplar-based sparse representation is not superior to ML-GMM-GV [37].

We are glad to see that all results in Fig. 7, 8, 10 and 11 resonate with those in Table I in the sense that better subjective test results come with lower LSD ratios. Fig. 9 (a) does present a different view that we consider it as an outlier.

Fig. 12: Comparison of speaker identification rates with 95% confidence intervals for different voice conversion methods.

2) Speaker Similarity Evaluation: We conducted XAB test to assess the speaker similarity for each voice conversion method. From the objective and subjective evaluations, we note that residual compensation helps consistently. In this test, we only involved systems with residual compensation. Fig. 12 reports the results of identification rate for the voice conversion methods of group (a), (c) and (g) in Table I. We note that the EDFW-ERC achieves the lowest identification rate of $(67.50 \pm 9.8)\%$, which is significantly lower than GDFW-GRC and ML-GMM-GV with the identifications of $(83.75 \pm 5.43)\%$ and $(85 \pm 7.17)\%$, respectively. While the identification rates of other methods fall into each other's confidence intervals. In other words, they are not significantly different².

To summarize, the subjective evaluation confirms that the proposed exemplar-based frequency warping framework achieves a significantly higher quality of speech than the state-of-the-art baselines, while maintaining the similar speaker identification rate.

V. CONCLUSIONS

In this paper, we propose a voice conversion framework with exemplar-based frequency warping and residual compensation. Both objective and subjective results confirm the effectiveness of the proposed methods. We summarize our findings here,

- The proposed exemplar-based frequency warping (EFW) framework benefits from both exemplar-based sparse representation and the frequency warping to improve spectral details.
- The exemplar-based residual compensation (ERC) that shares the same activation matrix with the warping func-

²Converted samples are available via: <http://listeningtestntnu.azurewebsites.net/voiceconversion/EFW>

tion reduces the spectral distortion and enhances the converted speech quality.

- The ECFW-ERC achieves significantly higher speech quality than the state-of-the-art methods (i.e. the ML-GMM-GV and DKPLS) while maintaining similar speaker identification rate.

In this work, we have shown that the exemplar-based frequency warping framework is suitable when the amount of training data is very limited, i.e. 20 sentences. While this study is focused on spectral transformation, we consider extending the proposed framework to prosodic transformation in the future work.

ACKNOWLEDGMENT

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative. Prof. Haizhou Li's contribution is supported by Ministry of Education, Singapore AcRF Tier 1 NUS Start-up Grant FY2016 Non-parametric approach to voice morphing.

The authors also would like to thank Mr. Hy Quy Nguyen, Mr. Zhi Hao Lim and Dr. Minghui Dong for their contributions to the discussions and valuable comments that have improved our manuscript.

REFERENCES

- [1] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, pp. 131–142, 1998.
- [2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1998.
- [3] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 2222–2235, 2007.
- [4] V. Popa, H. Silen, J. Nurminen, and M. Gabbouj, "Local linear transformation for voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.
- [5] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 912–921, 2010.
- [6] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 806–817, 2012.
- [7] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.
- [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 Transactions on Speech and Audio Processing*, pp. 1859–1872, 2014.
- [9] F.-L. Xie, Y. Qian, Y. Fan, F. K. Soong, and H. Li, "Sequence error (SE) minimization training of neural network for voice conversion," in *INTERSPEECH*, 2014.
- [10] H. Q. Nguyen, S. W. Lee, X. Tian, M. Dong, and E. S. Chng, "High quality voice conversion using prosodic and high-resolution spectral features," *Multimedia Tools and Applications*, pp. 5265–5285, 2016.
- [11] Z. Wu, E. S. Chng, and H. Li, "Conditional restricted boltzmann machine for voice conversion," in *China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, 2013.
- [12] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion using RNN pre-trained by recurrent temporal restricted boltzmann machines," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, pp. 580–587, 2015.
- [13] T. Nakashika, T. Takiguchi, and Y. Minami, "Non-parallel training in voice conversion using an adaptive restricted boltzmann machine," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 2032–2045, 2016.
- [14] C.-H. Lee and C.-H. Wu, "MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training," in *INTERSPEECH*, 2006.
- [15] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on gaussian mixture model," in *INTERSPEECH*, 2006.
- [16] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 922–931, 2010.
- [17] H. Benisty and D. Malah, "Voice conversion using GMM with enhanced global variance," in *INTERSPEECH*, 2011.
- [18] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modulation spectrum-based post-filter for GMM-based voice conversion," in *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, 2014.
- [19] D. Sundermann and H. Ney, "VTLN-based voice conversion," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2003.
- [20] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2003.
- [21] M. Eichner, M. Wolff, and R. Hoffmann, "Voice characteristics conversion for TTS using reverse VTLN," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.
- [22] A. Přibilová and J. Přibíl, "Non-linear frequency scale mapping for voice conversion in text-to-speech system with cepstral description," *Speech Communication*, pp. 1691–1703, 2006.
- [23] R. Vich and M. Vondra, "Pitch synchronous transform warping in voice conversion," in *Cognitive Behavioural Systems*, 2012.
- [24] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 556–566, 2013.
- [25] H. Valbret, E. Moulines, and J.-P. Tubach, "Voice transformation using PSOLA technique," *Speech Communication*, pp. 175–187, 1992.
- [26] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 1313–1323, 2012.
- [27] M. Vondra and R. Vich, "Speech identity conversion," in *Nonlinear Speech Modeling and Applications*, 2005.
- [28] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.
- [29] X. Tian, Z. Wu, S. W. Lee, and E. S. Chng, "Correlation-based frequency warping for voice conversion," in *9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2014.
- [30] X. Tian, Z. Wu, S. W. Lee, N. Q. Hy, E. S. Chng, and M. Dong, "Sparse representation for frequency warping based voice conversion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.
- [31] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," *INTERSPEECH*, 2014.
- [32] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 1066–1074, 2007.
- [33] B. King, C. Févotte, and P. Smaragdis, "Optimal cost function and magnitude power for NMF-based speech separation and music interpolation," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2012.
- [34] D. Kitamura, H. Saruwatari, H. Kameoka, Y. Takahashi, K. Kondo, and S. Nakamura, "Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 654–669, 2015.
- [35] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Spoken Language Technology workshop (SLT)*, 2012.
- [36] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in *8th ISCA Speech Synthesis Workshop*, 2013.

- [37] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1506–1521, 2014.
- [38] R. Aihara, T. Fujii, T. Nakashika, T. Takiguchi, and Y. Ariki, "Small-parallel exemplar-based voice conversion in noisy environments using affine non-negative matrix factorization," *EURASIP J. Audio, Speech and Music Processing*, p. 32, 2015.
- [39] Z. Jin, A. Finkelstein, S. DiVerdi, J. Lu, and G. J. Mysore, "Cute: A concatenative method for voice conversion using exemplar-based unit selection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.
- [40] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001.
- [41] N.-P. V. Nielsen, J. M. Carstensen, and J. Smedsgaard, "Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping," *Journal of Chromatography A*, pp. 17–35, 1998.
- [42] G. Tomasi, F. van den Berg, and C. Andersson, "Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data," *Journal of Chemometrics*, pp. 231–241, 2004.
- [43] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 43–49, 1978.
- [44] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, pp. 2067–2080, 2011.
- [45] A. B. Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School of Science and Eng., Portland, Oregon, USA, 2001.
- [46] 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 communication*, pp. 187–207, 1999.
- [47] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion," in *INTERSPEECH*, 2012.
- [48] D. T. Chappell and J. H. Hansen, "Speaker-specific pitch contour modeling and modification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1998.
- [49] B. King, C. Févotte, and P. Smaragdis, "Optimal cost function and magnitude power for NMF-based speech separation and music interpolation," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2012.
- [50] H. Ye and S. Young, "High quality voice morphing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.
- [51] M. K. Wolters, K. B. Isaac, and S. Renals, "Evaluating speech synthesis intelligibility using amazon mechanical turk," in *7th ISCA Speech Synthesis Workshop (SSW7)*, 2010.
- [52] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, "Crowdmoss: An approach for crowdsourcing mean opinion score studies," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.
- [53] S. Buchholz and J. Latorre, "Crowdsourcing preference tests, and how to detect cheating," in *INTERSPEECH*, 2011.

Xiaohai Tian received the B.Eng. degree in Software Engineering from the Northwestern Polytechnical University, Xi'an, China, in 2006, and M.Eng. degree in computer science from the same University in 2011. He is currently pursuing the Ph.D. degree at the School of Computer Science and Engineering, Nanyang Technological University, Singapore.

His current research interests are in voice conversion, speech synthesis and anti-spoofing.

Siu Wa Lee received her B.Eng. (Hons) First Class, M.Phil. and Ph.D. degrees in Electronic Engineering from The Chinese University of Hong Kong (CUHK). During her postgraduate studies, she worked in ATR Institute International, Kyoto and Microsoft Research Asia, Beijing as a visiting student. After PhD graduation, she was a Post-doctoral Fellow in the Department of Electronic Engineering, CUHK. From September 2016 onwards, she is working in Office of the Communications Authority (OFCA), Hong Kong Special Administrative Region Government, Hong Kong. Prior to that, she was with the Department of Human Language Technology, Institute for Infocomm Research (I2R), A*STAR, Singapore as a scientist. Her research interests include radio communication, spectral planning, spectrum regulatory affairs, speech signal processing, speech synthesis, voice conversion, speech separation, etc. She was awarded the Best Show & Tell Paper Award in INTERSPEECH 2014. She is a member of IEEE, International Speech and Communication Association (ISCA) and a corporate member of The Hong Kong Institution of Engineers.

Zhizheng Wu has been a Research Scientist at Apple Inc. since May 2016, prior to which he was a research fellow at CSTR from May 2014 to April 2016. He received his Ph.D. from Nanyang Technological University, Singapore. During his studies he joined Microsoft Research Asia (2007 - 2009) and the University of Eastern Finland (2012) as a visiting scientist and received the best paper award at the Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2012. He co-organised the first Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015) at Interspeech 2015 and the first Voice Conversion Challenge (VCC 2016) as a special session at Interspeech 2016. He delivered a tutorial on Spoofing and Anti-Spoofing: A Shared View of Speaker Verification, Speech Synthesis and Voice Conversion at APSIPA ASC 2015. He is the principal architect of the open-source speech synthesis system, Merlin.

Eng Siong Chng is currently an Associate Professor in the School of Computer Engineering, Nanyang Technological University, Singapore. He received his B.Eng. (Hons) in Electrical and Electronics Engineering from the University of Edinburgh, U.K in 1991, and Ph.D. from the same University in 1996.

Prior to joining NTU in 2003, he has worked in the following research centres and companies: 1996, Institute of Physics and Chemical Research, Riken, as a post-doc working in the area of signal processing and classification, 1996-1999, Institute of System Science as a research staff to transfer the Apple-ISS speech and handwriting technologies to ISS, 1999-2000, Lernout and Hauspie as a senior researcher in speech recognition, and 2001-2002, Knowles Electronics as a manager for the Intellisonic microphone array research and from 2003 onwards: NTU's School of Computer Science and Engineering. His research interests are in pattern recognition, signal, and speech processing. He is currently leading the speech and language technology program in at the School of Computer Science and Engineering, NTU.

Eng Siong is a senior member of IEEE since 2005.

Haizhou Li (M'91-SM'01-F'14) received the B.Sc., M.Sc., and Ph.D. degree in electrical and electronic engineering from South China University of Technology, Guangzhou, China in 1984, 1987, and 1990 respectively. Dr Li is currently a Professor at the Department of Electrical and Computer Engineering, National University of Singapore (NUS). He is also a Conjoint Professor at the University of New South Wales, Australia. His research interests include automatic speech recognition, speaker and language recognition, and natural language processing.

Prior to joining NUS, he taught in the University of Hong Kong (1988-1990) and South China University of Technology (1990-1994). He was a Visiting Professor at CRIN in France (1994-1995), Research Manager at the Apple-ISS Research Centre (1996-1998), Research Director in Lernout & Hauspie Asia Pacific (1999-2001), Vice President in InfoTalk Corp. Ltd. (2001-2003), and

the Principal Scientist and Department Head of Human Language Technology in the Institute for Infocomm Research, Singapore (2003-2016).

Dr Li is currently the Editor-in-Chief of IEEE/ACM Transactions on Audio, Speech and Language Processing (2015-2018), a Member of the Editorial Board of Computer Speech and Language (2012-2018). He was an elected Member of IEEE Speech and Language Processing Technical Committee (2013-2015), the President of the International Speech Communication Association (2015-2017), the President of Asia Pacific Signal and Information Processing Association (2015-2016), and the President of Asian Federation of Natural Language Processing (2017-2018). He was the General Chair of ACL 2012 and INTERSPEECH 2014.

Dr Li is a Fellow of the IEEE. He was a recipient of the National Infocomm Award 2002 and the President's Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation.