Exemplar-based sparse representation with residual compensation for voice conversion

Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, *Senior member, IEEE,* and Haizhou Li, *Fellow, IEEE*

***Abstract*—We propose a nonparametric framework for voice conversion, that is, exemplar-based sparse representation with residual compensation. In this framework, a spectrogram is reconstructed as a weighted linear combination of speech segments, called exemplars, which span multiple consecutive frames. The linear combination weights are constrained to be sparse to avoid over-smoothing, and high-resolution spectra are employed in the exemplars directly without dimensionality reduction to maintain spectral details. In addition, a spectral compression factor and a residual compensation technique are included in the framework to enhance the conversion performances. We conducted experiments on the VOICES database to compare the proposed method with a large set of state-of-the-art baseline methods, including the maximum likelihood Gaussian mixture model (ML-GMM) with dynamic feature constraint and the partial least squares (PLS) regression based methods. The experimental results show that the objective spectral distortion of ML-GMM is reduced from 5.19 dB to 4.92 dB, and both the subjective mean opinion score and the speaker identification rate are increased from 2.49 and 73.50 % to 3.15 and 79.50 %, respectively, by the proposed method. The results also show the superiority of our method over PLS-based methods. In addition, the subjective listening tests indicate that the naturalness of the converted speech by our proposed method is comparable with that by the ML-GMM method with global variance constraint.**

***Index Terms*—Voice conversion, exemplar, sparse representation, nonnegative matrix factorization, residual compensation**

I. Introduction

Generally speaking, a speech signal carries threefold information: voice timbre, prosody, and language content. *Voice conversion* is a technique to manipulate one speaker's (source) voice timbre and/or prosody to impersonate another speaker (target) without changing the language content. As *spectral attributes*, which relate to voice timbre, play an important role in characterizing speaker individuality [1], *spectral mapping* has been intensively studied as an acoustic feature mapping problem, which is also the focus of this work.

Spectral mapping has many practical uses. The most popular one would be in speech synthesis, where voice conversion techniques are used to create a personalized text-to-speech (TTS) system with a small amount of speech samples from a specific speaker [2]. Such mapping techniques are also handy in many other applications, such as narrowband-to-wideband conversion [3], single-channel speech enhancement [4], noise robust feature compensation [5], [6], acoustic-to-articulatory inversion mapping [7], and body-transmitted speech enhancement [8].

A large number of statistical parametric approaches have attempted to achieve a robust spectral mapping. It was proposed that source-target feature mapping can be established through vector quantization (VQ) [9], where the source-target feature pairs are used to estimate a mapping codebook during training. At runtime, the codebook is used to select a sequence of target centroid features to generate converted feature trajectories. The discrete nature of VQ inevitably leads to glitches in the converted voices that sound unnatural. To alleviate such an incoherence, fuzzy VQ was proposed in [10] using continuous-values weight vectors to interpolate the centroid vectors in the codebook for generation. *Gaussian mixture model* (GMM) based approaches were also proposed in [2], [11], [12], [13] to implement a weighted local linear conversion function for continuous spectral mapping. At the same time, other methods based on linear transformation, such as perceptually weighted linear transformation [14], maximum likelihood linear transformation [15], partial least squares (PLS) regression [16] and local linear transformation [17], have also been proposed assuming that the source and target features are linearly correlated. Alternatively, nonlinear methods have also been proposed assuming a nonlinear relationship between the source and target features. Such methods include artificial neural network [18], [19], support vector regression [20], kernel PLS regression [21] and restricted Boltzmann machine [22], just to name a few.

While the above statistical parametric approaches convert speaker identity well, they are often at the cost of degraded speech quality. As an alternative to statistical approaches which typically shift source spectra to match those of the target through mapping functions, vocal tract length normalization (VTLN) techniques were proposed in [23], [24] to warp the frequency axis of a source spectrum to match that of the target. Similar ideas, such as weighted frequency warping [25], dynamic frequency warping [26] and bilinear frequency warping [27], were also proposed. As these frequency warping approaches are able to keep the spectral details, they produce converted speech of perceptually higher quality than statistical parametric approaches. However, there is a trade-off between speech quality and identity conversion performance. It is reported that frequency warping techniques offer inferior speaker identity conversion quality to that of statistical approaches [25], [27].

Z. Wu and E. S. Chng are with the School of Computer Engineering, Nanyang Technological University, Singapore 639798, and also with Temasek Lab@NTU, Nanyang Technological University, Singapore. (Email: wuzz@ntu.edu.sg; aseschng@ntu.edu.sg)

T. Virtanen is with the Department of Signal Processing, Tampere University of Technology, Tampere, Finland. (Email: tuomas.virtanen@tut.fi)

H. Li is with the Institute for Infocomm Research, Singapore 138632, and also with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (Email:hli@i2r.a-star.edu.sg)

While balancing speech quality and speaker individuality of converted speech, the *robustness* of statistical parametric approaches is limited by the fact that they attempt to predict speech trajectories from model parameters. Robustness refers to the ability that a system is capable to handle new training data without much tuning. In the parametric approaches, when new training data arrives, the model parameters for prediction are required to be re-optimized to reduce the mismatch between the 'old' model and new data. It also refers to the ability that a method is reliable to handle various training scenarios, such as limited training samples and high-dimensional features. When there are too many parameters and too few training samples, over-fitting occurs [16]. Over-fitted model usually has poor predictive performance. It is because the model parameters are estimated by maximizing the performance on the training data.

In [28], we proposed a nonparametric spectral mapping framework, namely exemplar-based sparse representation, for voice conversion. In this framework, each speech segment is reconstructed as a weighted linear combination of a set of basis *exemplars*. An exemplar is defined as a speech segment spanning multiple frames extracted from the training data, while the set of linear combination weights compose an *activation* vector, which is constrained to be *sparse*. In the work, we assumed that a collection of acoustically aligned source and target exemplars, called *coupled dictionary* could share the same activation vector to generate the converted spectrograms. Due to the nonnegative nature of a spectrogram, in practice, both nonnegative matrix factorization (NMF) [29] with sparsity constraint [30] and nonnegative matrix deconvolution (NMD) [31] were examined to estimate the activations. Similar work based on sparse representation was conducted in [32] for noise-robust voice conversion, where NMF with sparsity constraint was employed to find the activations.

The exemplar-based sparse representation framework describes speech observations by a dictionary. There are three advantages of using this framework for voice conversion: a) it is straightforward to construct the dictionary by using speech segments directly from the training data; b) it allows us to model high-dimensional spectra directly to maintain the spectral details; and c) the generation of converted spectrogram is as simple as combining a set of basis speech segments without mapping or modification. In addition, by constraining the activation vector to be sparse, we avoid the over-smoothing problem during the linear combination of exemplars. It has been confirmed that the exemplar-based sparse representation framework offers high quality voice conversion [28].

In this work, we extend the exemplar-based sparse representation framework [28] through residual compensation. First, we examine the temporal constraint for both low-resolution and high-resolution features. High-resolution features usually produce accurate activation weights but at a high computational cost, while low-resolution features are computational efficient in capturing the temporal structure of speech signals. We present a detailed comparison between low-resolution and high-resolution features for voice conversion performance.

Second, inspired by the work in source separation [33], the dynamic range of the spectrogram amplitude affects the performance considerably. We know that, in the original spectrogram, high-frequency bands have low-intensity observations, while low-frequency bands have high-intensity observations. As a result, some important but low-intensity observations are not given adequate attention. Addressing this problem, we introduce a spectral compression factor to balance the intensity between high- and low-frequency bands.

Last, during the estimation of activations there exists inevitably some modelling error between the source spectrogram and the modeled spectrogram. Such residuals usually contain spectral details which may affect the converted speech quality. We hence propose a residual compensation technique to reimburse the source model residual for the converted spectrogram to enhance the speech quality.

The main contributions of this work are threefold:

- An exemplar-based sparse representation with residual compensation framework for voice conversion is proposed, allowing us to model high-resolution spectra directly.
- A spectral compression method is investigated to emphasize important but low-intensity observations.
- A residual compensation technique is introduced to enhance the converted speech quality.

II. Problem Statement

A. General voice conversion framework

The goal in spectral mapping for voice conversion is to map source speaker's features to match those of the target speaker. Mathematically, spectral mapping is written as

$$\mathbf{y} = \mathcal{F}(\mathbf{x}), \tag{1}$$

where $\mathbf{x} \in \mathcal{R}^{d\times 1}$ and $\mathbf{y} \in \mathcal{R}^{d\times 1}$ are the source and target features, respectively, d is the dimensionality of features, and $\mathcal{F}(\cdot)$ is the conversion function to map the features.

A typical voice conversion system consists of two stages: offline training and runtime conversion. During the offline training, parallel utterances from a source speaker X and a target speaker Y are first aligned using dynamic time warping (DTW) algorithm as

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_N] \tag{2}$$

$$\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n, \ldots, \mathbf{y}_N], \tag{3}$$

where $\mathbf{X} \in \mathcal{R}^{d\times N}$ and $\mathbf{Y} \in \mathcal{R}^{d\times N}$ are the paired source and target data, respectively, N is the number of frames in training, and $\mathbf{x}_n$ and $\mathbf{y}_n$ are the paired source and target features, respectively. If a parallel dataset is not available, nonparallel alignment techniques [34] can be applied to find the frame alignment. The parallel data $\mathbf{Z} = [\mathbf{X}; \mathbf{Y}]$, where each column $\mathbf{z}_n = [\mathbf{x}_n; \mathbf{y}_n]$ is a joint vector, is employed to estimate the conversion function $\mathcal{F}(\cdot)$.

At runtime, given a source feature $\mathbf{x}$ or a sequence of source features $\mathbf{X}$, the conversion function $\mathcal{F}(\cdot)$ is adopted to predict a target feature $\hat{\mathbf{y}} = \mathcal{F}(\mathbf{x})$.

B. Limitations

Aside from the issue of robustness, the performance of statistical parametric approaches is also limited by the statistical average nature and the use of low-resolution features. We discuss the two limitations in this section.

1) Statistical average: In the statistical parametric methods, conversion functions are optimized by minimizing mean square error of between reference target and converted speech [11], or maximizing joint likelihood between source and target speech in the training stage [12]. This inevitably leads to conversion functions that model the average properties of spectra, but discard spectral details.

We take the joint-density Gaussian mixture model (JD-GMM) method as a case study to show the averaging effect. During the parameter estimation process of JD-GMM, the joint mean vector is calculated as a weighted linear combination of all the training samples, which is the so called *statistical average*. The averaging process will remove some spectral details which cannot be recovered during synthesis.

Moreover, if the correlation between source and target features is low, the value of the cross covariance elements will be extremely small. As a result, only the mean vector will contribute to the converted speech [35]. At the same time, during conversion, the mean vectors are fixed for every input feature, thus the variation of the generated parameter trajectories will be low.

2) Low-resolution features: In conventional statistical parametric approaches, low-resolution features such as mel-cepstral coefficients (MCCs) [36] and line spectrum pair (LSP) [37] are commonly used to represent high-resolution spectra, which are the spectra/spectral envelopes extracted from the discrete Fourier transform or linear predictive coding. The use of low-resolution features is to reduce the feature dimensionality for computational efficiency. We argue that the use of such low-resolution features loses spectral details and results in converted spectrum that is smoothed. Fig. 1 shows a comparison of an original spectral envelope and a reconstructed spectral envelope from 24-order MCCs. It is observed that after reconstruction from low-resolution features, the spectral details are lost, especially in high-frequency bands. There are partial evidences showing that high-resolution feature representations produce synthetic speech with better quality than low-resolution features [38], [39].

In general, due to the statistical average and the use of low-resolution feature representations, conventional statistical parametric approaches such as JD-GMM suffer from the over-smoothing problem, and generate muffled-sounding speech.

III. Exemplar-based sparse representation with residual compensation

To overcome the limitations of statistical parametric approaches, we propose an alternative nonparametric framework, specifically, an exemplar-based sparse representation with residual compensation method, where high-resolution spectra are directly used to synthesize the converted speech. The proposed framework is described in this section.

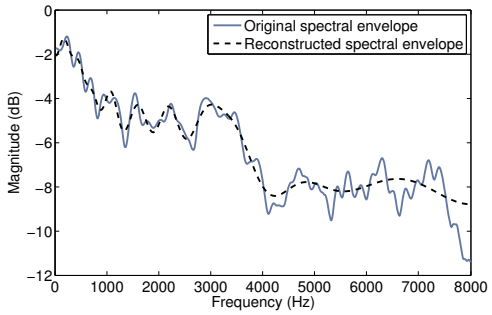

Fig. 1: Illustration of the smoothing effect of low-resolution features. The dashed line is reconstructed from 24-order MCCs, which are computed from the solid line.

A. Basic exemplar-based sparse representation

The idea of exemplar-based sparse representation is to describe a magnitude spectrum as a linear combination of a set of basis spectra, namely, exemplars. Mathematically, it is written as

$$\mathbf{x}^{(\mathrm{DFT})} \approx \sum_{n=1}^{N} \mathbf{a}_n^{(\mathrm{DFT})} \cdot h_n = \mathbf{A}^{(\mathrm{DFT})}\mathbf{h} \quad \text{subject to} \quad \mathbf{h} \geq 0, \tag{4}$$

where $\mathbf{x}^{(\mathrm{DFT})} \in \mathcal{R}^{F\times 1}$ represents the high-resolution spectrum of one speech frame, F is the dimensionality of high-resolution spectra, N is the number of exemplars in a dictionary, $\mathbf{A}^{(\mathrm{DFT})} = [\mathbf{a}_1^{(\mathrm{DFT})}, \mathbf{a}_2^{(\mathrm{DFT})}, \cdots, \mathbf{a}_N^{(\mathrm{DFT})}] \in \mathcal{R}^{F\times N}$ is the fixed dictionary of exemplars built from the source training set, $\mathbf{a}_n^{(\mathrm{DFT})}$ is the n^{th} source exemplar which has the same dimensionality as $\mathbf{x}^{(\mathrm{DFT})}$, $\mathbf{h} = [h_1, h_2, \cdots, h_N] \in \mathcal{R}^{N\times 1}$ is the activation vector and h_n is the nonnegative weight, or activation, of the n^{th} exemplar.

Each observation is modeled independently, and a spectrogram of each source utterance can therefore be represented as

$$\mathbf{X}^{(\mathrm{DFT})} \approx \mathbf{A}^{(\mathrm{DFT})}\mathbf{H}, \tag{5}$$

where $\mathbf{X}^{(\mathrm{DFT})} \in \mathcal{R}^{F\times M}$ is the source spectrogram, M is the number of frames in a source utterance and $\mathbf{H} \in \mathcal{R}^{N\times M}$ is the activation matrix, each column vector of which is an activation vector in Eq. (4).

To generate a converted spectrogram, we assume that paired source-target dictionaries $\mathbf{A}^{(\mathrm{DFT})}$ and $\mathbf{B}^{(\mathrm{DFT})}$ with acoustically aligned exemplars can share the same activation matrix $\mathbf{H}$. Note that each column vector in $\mathbf{B}^{(\mathrm{DFT})}$ corresponds to a column vector in $\mathbf{A}^{(\mathrm{DFT})}$, and they are obtained from the aligned data in the way described in Eqs. (2) and (3). Thus, the converted spectrogram can be generated as

$$\hat{\mathbf{Y}}^{(\mathrm{DFT})} = \mathbf{B}^{(\mathrm{DFT})}\mathbf{H}, \tag{6}$$

where $\hat{\mathbf{Y}}^{(\mathrm{DFT})} \in \mathcal{R}^{F\times M}$ is the converted spectrogram, $\mathbf{B}^{(\mathrm{DFT})} \in \mathcal{R}^{F\times N}$ is the fixed target dictionary of exemplars from target training data, and $\mathbf{H}$ is as found in Eq. (5).

Due to the nonnegative nature of source spectrogram $\mathbf{X}^{(\mathrm{DFT})}$ and source dictionary $\mathbf{A}^{(\mathrm{DFT})}$, the nonnegative matrix factorization (NMF) technique [29], [30] is employed to estimate the activation matrix $\mathbf{H}$, which is found by minimizing

the objective function

$$\mathbf{H} = \arg\min_{\mathbf{H}\geq 0} d(\mathbf{X}^{(\mathrm{DFT})}, \mathbf{A}^{(\mathrm{DFT})}\mathbf{H}) + \lambda\|\mathbf{H}\|_1, \quad (7)$$

where λ is the sparsity penalty factor. Here, only the activation matrix $\mathbf{H}$ is estimated and the dictionary $\mathbf{A}^{(\mathrm{DFT})}$ is fixed. In practice, the generalised Kullback-Leibler (KL) divergence is used for $d(\mathbf{X}^{(\mathrm{DFT})}, \mathbf{A}^{(\mathrm{DFT})}\mathbf{H})$. Similar to [30], we minimize the objective function in Eq. (7) by iteratively applying the following multiplicative updating rule:

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{A}^{(\mathrm{DFT})^\top} \frac{\mathbf{X}^{(\mathrm{DFT})}}{\mathbf{A}^{(\mathrm{DFT})}\mathbf{H}}}{\mathbf{A}^{(\mathrm{DFT})^\top} + \lambda}, \quad (8)$$

where $\otimes$ represents element-wise multiplication and divisions are also element-wise. The convergence of Eq. (7) using this update rule is proven in [30]. In our study with real speech data, it is observed that this update rule converges robustly.

In the following, we make several modifications to the basic setup described above.

B. Spectrum compression

The relative range between high- and low-intensity observations is an important factor which affects the activation matrix estimation, as well as spectrogram generation. In the context of source separation, changing the dynamic range of spectrograms by exponentiating them has been found to affect the performance significantly [33]. In a similar way, we introduce a spectral compression parameter ρ into the computation of the activation matrix, as follows:

$$(\mathbf{X}^{(\mathrm{DFT})})^\rho \approx (\mathbf{A}^{(\mathrm{DFT})})^\rho \mathbf{H}, \quad (9)$$

$$\hat{\mathbf{Y}}^{(\mathrm{DFT})} = ((\mathbf{B}^{(\mathrm{DFT})})^\rho \mathbf{H})^{1/\rho}. \quad (10)$$

We conducted an analysis of the spectral shapes varying the compression factor from 0.2 to 1.0 as presented in Fig. 2. It shows that a smaller compression factor implies more emphasis on higher frequency bands that have low intensity. Note that the spectral compression will not change the nonnegative nature of a spectrogram.

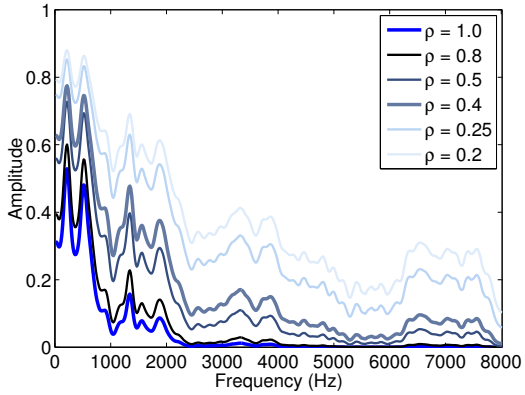

Fig. 2: Illustration of spectral shapes with corresponding compression factors (ρ).

Values $\rho < 1$ compress the spectrum, making its values closer to each other. The KL divergence is linear in terms of the scale of its arguments [40], and compression/expansion affects the relative weight of small/large intensity observations in the estimation. Note that ρ can be applied to the source spectrogram and dictionaries in advance before estimating the activation matrix. This allows us to use the same objective function and updating rule as in Eqs. (7) and (8), to find the activation matrix $\mathbf{H}$. When $\rho = 1$, Eqs. (9) and (10) are reduced to Eqs. (5) and (6), respectively.

C. Contextual information

So far, as shown in Eqs. (5) and (9), no contextual information is taken into consideration when estimating the activation matrix, in other words, each frame is modeled independently. It is easy to understand that contextual information is important in modeling a speech signal. To benefit from the context, we suggest using a speech segment that spans multiple consecutive frames as an exemplar in the source dictionary. The column vectors in each exemplar are stacked into a single vector to simplify the notation. In this way, an exemplar from the dictionary in Eqs. (4), (5) and (9) can be defined as

$$\mathbf{a}_n^{(\mathrm{DFT})} = \left[\mathbf{a}_{n,-q}^{(\mathrm{DFT})}; \cdots; \mathbf{a}_{n,-1}^{(\mathrm{DFT})}; \mathbf{a}_{n,0}^{(\mathrm{DFT})}; \mathbf{a}_{n,+1}^{(\mathrm{DFT})}; \cdots; \mathbf{a}_{n,+q}^{(\mathrm{DFT})}\right], \quad (11)$$

where $L = 2\times q+1$ is the *window size* of an exemplar, $\mathbf{a}_{n,0}^{(\mathrm{DFT})}$ is exactly the same as $\mathbf{a}_n^{(\mathrm{DFT})}$ in Eq. (5), $\mathbf{a}_{n,-q}^{(\mathrm{DFT})}$ and $\mathbf{a}_{n,+q}^{(\mathrm{DFT})}$ are the q^{th} frames preceding and following frame $\mathbf{a}_{n,0}^{(\mathrm{DFT})}$ in the original time sequence, respectively. Thus, the stacking vector $\mathbf{a}_n^{(\mathrm{DFT})} \in \mathcal{R}^{(L\times F)\times 1}$ is able to represent an exemplar spanning L frames.

D. Using low-resolution features for faster computation

As shown in Eqs. (5), (6), (9) and (10), the size of the activation matrix $\mathbf{H}$ is independent of the feature dimensionality of the source dictionary $\mathbf{A}^{(\mathrm{DFT})}$. On the other hand, the feature dimensionality of $\mathbf{A}^{(\mathrm{DFT})}$ will affect the computation and memory usage considerably, especially when a relatively large context is used. To overcome this, we propose a new implementation of NMF using low-resolution features in the source dictionary. This kind of coupled dictionaries has previously been applied e.g. to combine good time and frequencies resolutions [41], to expand the bandwidth of speech [42] and to do robust automatic speech recognition [30].

We consider an exemplar without contextual information first and define the low-resolution implementation as

$$\mathbf{W}(\mathbf{X}^{(\mathrm{DFT})})^\rho \approx \mathbf{W}(\mathbf{A}^{(\mathrm{DFT})})^\rho \mathbf{H}, \quad (12)$$

where $\mathbf{W} \in \mathcal{R}^{U\times F}$ is the matrix to perform dimensionality reduction, and U is the dimensionality of the low-resolution feature with $U \leq F$ constraint.

In practice, the low-resolution used here corresponds to the Mel-scale, so that each column of $\mathbf{W}$ is the triangular magnitude response of a filter, and thus, for simplicity, we use $\mathbf{X}^{(\mathrm{MEL})} \in \mathcal{R}^{U\times M}$ for $\mathbf{W}(\mathbf{X}^{(\mathrm{DFT})})^\rho$, and denote $\mathbf{W}(\mathbf{A}^{(\mathrm{DFT})})^\rho = \mathbf{A}^{(\mathrm{MEL})} \in \mathcal{R}^{U\times N}$. In this way, Eq. (12) becomes:

$$\mathbf{X}^{(\mathrm{MEL})} \approx \mathbf{A}^{(\mathrm{MEL})}\mathbf{H}. \quad (13)$$

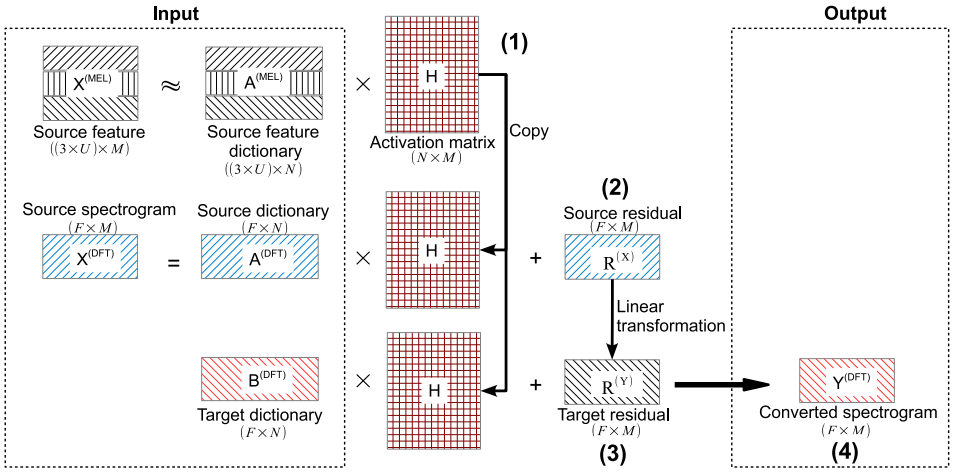

Fig. 3: Illustration of the exemplar-based sparse representation with residual compensation framework, which consists of four processes: (1) estimating the activation matrix $\mathbf{H}$ using low-resolution features; (2) calculating the source residuals $\mathbf{R}^{(\mathrm{X})}$ following Eq. (14); (3) mapping source residuals to target residuals following Eq. (16); and (4) generating converted spectrograms $\hat{\mathbf{Y}}^{(\mathrm{DFT})}$ following Eq. (17). In the diagram, U and F are the feature dimensionalities of low-resolution and high-resolution features, respectively, $3 \times U$ means an exemplar spanning three frames, M is the number of frames in a source utterance, and N is the number of exemplars in a dictionary. As discussed in Section III-F, U is set to 50, and F is set to 513.

Note that Eq. (13) is similar to Eq. (5), this allows us to use the same estimation method. To benefit from multiple-frame exemplars, we follow the same stacking as in Eq. (11) to establish a context-dependent source dictionary by simply replacing $(\mathbf{a}_{n,+q}^{(\mathrm{DFT})})^{\rho}$ with $\mathbf{a}_{n,+q}^{(\mathrm{MEL})} = \mathbf{W}(\mathbf{a}_{n,+q}^{(\mathrm{DFT})})^{\rho}$.

The advantage of using low-resolution features rather than high-resolution features to estimate the activation matrix is that the computational complexity can be reduced greatly. Even though low-resolution features are used to estimate the activation matrix as Eq. (13), the activations are applied on the high-resolution dictionary $\mathbf{B}^{(\mathrm{DFT})}$ to generate the converted spectrogram as Eq. (10). Note that the low-resolution source dictionary $\mathbf{A}^{(\mathrm{MEL})}$ is acoustically aligned with the high-resolution dictionary $\mathbf{B}^{(\mathrm{DFT})}$, we hence assume that $\mathbf{B}^{(\mathrm{DFT})}$ can also share the activation matrix $\mathbf{H}$ with $\mathbf{A}^{(\mathrm{MEL})}$.

E. Compensating model residual

The modeling error between the observed source spectrogram $\mathbf{X}^{(\mathrm{DFT})}$ and the modeled spectrogram $((\mathbf{A}^{(\mathrm{DFT})})^{\rho}\mathbf{H})^{1/\rho}$ is also called residual. We propose a residual compensation technique to enhance the spectral mapping performance.

To perform residual compensation, during offline training, we use a development set in the following four steps to calculate source and target spectrogram residuals:

(a) Estimate the activation matrix $\mathbf{H}$ using low-resolution features as that in Eq. (13).

(b) Apply $\mathbf{H}$ to the source and target spectral dictionaries to reconstruct source and target spectrograms, respectively.

(c) Calculate source residuals $\mathbf{R}^{(\mathrm{X})}$ by subtracting the magnitude of the modeled spectrograms with the corresponding reference source spectrograms as

$$\mathbf{R}^{(\mathrm{X})} = \log(\mathbf{X}^{(\mathrm{DFT})}) - \log(((\mathbf{A}^{(\mathrm{DFT})})^{\rho}\mathbf{H})^{1/\rho}). \quad (14)$$

(d) Calculate target residuals $\mathbf{R}^{(\mathrm{Y})}$ by subtracting the magnitude of the converted spectrograms with the corresponding reference target spectrograms as

$$\mathbf{R}^{(\mathrm{Y})} = \log(\mathbf{Y}^{(\mathrm{DFT})}) - \log(((\mathbf{B}^{(\mathrm{DFT})})^{\rho}\mathbf{H})^{1/\rho}). \quad (15)$$

We obtain the source-target residual pairs by using the corresponding reference source-target frame alignment information. With the paired source-target residuals, a mapping can be established as

$$\mathbf{R}^{(\mathrm{Y})} \approx \mathcal{F}(\mathbf{R}^{(\mathrm{X})}). \quad (16)$$

In practice, the mapping is implemented by partial least squares (PLS) regression, which is able to handle high-resolution features. The details of PLS regression can be found in [16].

The runtime conversion process is illustrated in Fig. 3. Similar to that in training, the activation matrix $\mathbf{H}$ is first estimated using low-resolution features as that in Eq. (13). Next, $\mathbf{H}$ is applied to the source and target spectral dictionaries to generate reconstructed source and converted spectrograms, respectively. Then, source residuals are calculated by subtracting the modeled spectrograms with reference source spectrograms in log-scale as that in Eq. (14). After that, the source residuals are mapped to target, which are added to the converted spectrograms on a logarithmic amplitude scale. Finally, the residual compensated spectrograms will be reverted back to a linear amplitude scale. The whole process is formulated as

$$\hat{\mathbf{Y}}^{(\mathrm{DFT})} = \exp(\log(((\mathbf{B}^{(\mathrm{DFT})})^{\rho}\mathbf{H})^{1/\rho}) + \mathcal{F}(\mathbf{R}^{(\mathrm{X})})). \quad (17)$$

We perform residual compensation on a logarithmic amplitude scale to guarantee the nonnegative nature of a spectrogram.

F. Dictionary Construction

As shown in Fig. 3, two kinds of dictionaries are involved in this work. Before constructing dictionaries, the feature representations used in this work are presented as follows:

- *High-resolution Magnitude spectra* consist of a sequence of 513-dimensional spectral envelopes extracted by STRAIGHT [43], and the envelopes are passed to STRAIGHT to reconstruct an audible speech signal at runtime. The frequency resolution of the high-resolution spectra is similar to that from the discrete Fourier transform (DFT). To this end, we use label DFT to denote the high-resolution spectra used in the source spectrogram $\mathbf{X}^{(\mathrm{DFT})}$, the target spectrogram $\mathbf{Y}^{(\mathrm{DFT})}$ or $\hat{\mathbf{Y}}^{(\mathrm{DFT})}$, the source spectral dictionary $\mathbf{A}^{(\mathrm{DFT})}$ and the target spectral dictionary $\mathbf{B}^{(\mathrm{DFT})}$.
- *Low-resolution mel-scale filter-bank energies (MELs)* are obtained by passing the high-resolution magnitude spectrogram to 50 mel-scale filter-banks, where the lower and upper frequencies are set to be 133.33 Hz and 6,855.5 Hz, respectively. In this paper, MELs are used as low-resolution features in $\mathbf{X}^{(\mathrm{MEL})}$ and in the source feature dictionary $\mathbf{A}^{(\mathrm{MEL})}$, which are used to estimate the activation matrix.
- *Mel-cepstral coefficients* (MCCs) are obtained by applying the mel-cepstral analysis technique [44] on a magnitude spectrogram and keeping 24 dimensions as features. During synthesis, MCCs are reverted back to a magnitude spectrogram, which is then passed to STRAIGHT to reconstruct an audible speech signal.

Given a parallel dataset between source and target speakers, we take the following steps to extract paired exemplars:

(a) Extract high-resolution magnitude spectrograms (spectral envelopes) from both source and target speech signals using STRAIGHT.
(b) Apply mel-cepstral analysis technique [44] to the magnitude spectrogram to compute MCCs.
(c) Apply 50 Mel-scale filter-banks to the source spectrograms to compute 50-dimensional MELs.
(d) Perform dynamic time warping (DTW) to align the source and target MCCs to obtain frame-by-frame source-target alignment.
(e) Apply the frame alignment information to the source and target magnitude spectrograms to obtain the high-resolution source $\mathbf{a}_n^{(\mathrm{DFT})}$ and target exemplars $\mathbf{b}_n^{(\mathrm{DFT})}$, respectively.
(f) Apply the same frame alignment information to the source MELs to obtain the low-resolution source dictionaries $\mathbf{a}_n^{(\mathrm{MEL})}$.

In the above six steps, we produce the paired exemplars from the training dataset. A simple way to construct dictionaries is by putting all these paired exemplars as column vectors in the corresponding dictionaries such as $\mathbf{A}^{(\mathrm{DFT})}$, $\mathbf{B}^{(\mathrm{DFT})}$ and $\mathbf{A}^{(\mathrm{MEL})}$, and keeping them unchanged for runtime conversion. In the experiments, we examine the performance of dictionaries using a subset of the paired exemplars, for example, randomly selecting exemplar pairs and storing as column vectors in corresponding dictionaries.

IV. Experiments

We conducted experiments using the VOICES[1] database [45] to assess the performance of the proposed exemplar-based sparse representation with residual compensation method. Speech data from two male speakers (jal and jcs) and two female speakers (sas and leb) was used. Voice conversion was conducted for all the 12 speaker pairs including 4 intra-gender and 8 inter-gender conversions. In each pair, 10 utterances were randomly selected as a training set, 10 utterances as a development set, and 20 utterances as an evaluation set. There was no overlapping across the three sets.

A. Reference methods and setups

To validate our proposals, we use a large set of state-of-the-art methods as the reference baselines, that include the well established ML-GMM method and several variations of the partial least squares (PLS) regression based methods. Note that PLS-based methods also depend on GMM in order to implement local transformations, but use more advance techniques to obtain the mapping function. In addition, we implement several nonnegative matrix factorization (NMF) based methods within the exemplar-based sparse representation framework to show the intermediate methods towards the proposed method. They are summarized as follows:

- *ML-GMM*: The joint-density Gaussian mixture model (JD-GMM) method with dynamic feature constraint proposed by Toda et al. [12] is a well-established baseline method. Note that in this work, 24-dimensional MCCs were used to represent the spectral envelope. Cross-diagonal covariance was adopted in the JD-GMM.
- *ML-GMM-GV*: The ML-GMM method with global variance enhancement in [12]. We used the same configuration as ML-GMM and used a postprocessing technique as presented in [46] to perform the GV implementation.
- *DKPLS*: The dynamic kernel partial least squares (DKPLS) regression method has been shown to be effective for implementing a nonlinear conversion function [21]. We used the same configuration as that in [21].
- *DKPLS-DFT*: The dynamic kernel partial least squares (DKPLS) regression method was also applied to high-resolution spectrograms. Comparing with DKPLS, this implementation is to examine the flexibility of DKPLS in face of high-dimensional features.
- *DPLS-DFT*: The partial least squares (PLS) regression method [16] was applied to high-resolution spectrograms, and three consecutive spectra were stacked as source features to include dynamic information for predicting a target spectrum. Comparing with DKPLS-DFT, this method is to evaluate the performance of basic PLS without kernel transformation.
- *NMF-DFT*: This is the basic exemplar-based sparse representation method implemented by nonnegative matrix factorization (NMF). High-resolution magnitude spectra were employed in the source and target dictionaries. It

[1] http://catalog.ldc.upenn.edu/LDC2006S01

used Eq. (5) to estimate the activation matrix and Eq. (6) to generate the converted spectrograms.

- *NMF-DFT-SC*: This is the NMF-DFT method with spectral compression, as presented in Eq. (9) and (10). With reference to NMF-DFT, this method is to show the effect of spectral compression.
- *NMF-MEL-SC*: This is the NMF with spectral compression method using low-resolution Mel-scale filter-bank energies (MELs) in the source dictionary. It employed Eq. (13) to estimate the activation matrix, and Eq. (10) to produce the converted spectrograms. With reference to NMF-DFT-SC, this method is to show the effect of feature dimensionality reduction in the source dictionary.
- ***NMF-MEL-SC-RC (Proposed)***: This is the complete exemplar-based sparse representation with residual compensation method as presented in Section III-E and Fig. 3. With reference to NMF-MEL-SC, we show the effect of residual compensation.

Table I summarizes the voice conversion methods with involved feature representations, and the equations for activation matrix estimation and spectrogram generation. The spectral mapping was performed using above methods, while F0 was converted by a simple linear conversion, normalizing the mean and variance of the source speech to equalize that of the target. In this work, we only deal with the magnitude spectra, while adopting minimum-phase for all the methods when reconstructing the speech signals. In practice, the STRAIGHT vocoder was employed. For a fair comparison, we shared across all the methods the same frame alignment obtained from frame-by-frame DTW.

TABLE I: Summary of the implemented methods and their formulations.

Method	Spectral feature	Activation estimation	Spectrogram generation
ML-GMM	MCCs	n/a	Eq. (39) in [12]
ML-GMM-GV	MCCs	n/a	Eq. (16) in [46]
DKPLS	MCCs	n/a	Eq. (6) in [21]
DKPLS-DFT	DFTs	n/a	Eq. (6) in [21]
DPLS-DFT	DFTs	n/a	Eq. (10) in [16]
NMF-DFT	DFTs	Eq. (5)	Eq. (6)
NMF-DFT-SC	DFTs	Eq. (9)	Eq. (10)
NMF-MEL-SC	DFTs, MELs	Eq. (13)	Eq. (10)
NMF-MEL-SC-RC	DFTs, MELs	Eq. (13)	Eq. (17)

In the first iteration of Eq. (8), $\mathbf{H}$ was initialized to unity, and the update rule was repeated for 500 iterations. The sparsity penalty factor λ was empirically set to 0.1 which is selected according to the performance on the development set.

We conducted both objective and subjective evaluations to assess the performance of the reference methods discussed above. In both evaluations, we first assessed the performance of NMF-based methods to show the effect of the incremental modifications to the basic exemplar-based sparse representation, namely, NMF-DFT. After that, we compared our proposed NMF-MEL-SC-RC method with the baselines, namely, ML-GMM, DKPLS, DKPLS-DFT and DPLS-DFT methods.

B. Objective evaluations

Mel-cepstral distortion (MCD) calculated between the converted and corresponding reference target Mel-cepstra was employed as an objective evaluation measure. The MCD was calculated as

$$\text{MCD[dB]} = \frac{10}{\log 10}\sqrt{2\sum_{i=1}^{24}(c_i - c_i^{\text{conv}})^2}, \qquad (18)$$

where c_i and c_i^{conv} are the i^{th} coefficients of the reference target and converted MCCs, respectively. For the DKPLS-DFT, PLS-DFT, DPLS-DFT and exemplar-based sparse representation methods, MCCs were computed from the converted spectrograms to calculate the MCD, so that the objective measure was comparable across all the methods. The MCD was calculated frame-by-frame over all the paired frames in the evaluation set, and the average MCD value was reported. A lower MCD value indicates smaller spectral distortion.

1) Effect of dictionary construction: We examined the effect of dictionary construction using NMF-DFT method which represents a basic sparse representation method. In the training set, there were approximately 5500 exemplars from each speaker pair. If all the exemplars were used, the computation and memory usage would be considerably high. Thus, we varied the number of exemplars from 500 to 5500 and observed the conversion performance. Note that the exemplars were randomly selected from the training set, and a smaller set of exemplars was always a subset of a larger set.

Fig. 4 presents the spectral distortions as a function of the number of exemplars, and the results on the development and evaluation sets are presented for comparison. It is observed that the spectral distortion decreases as the number of exemplars increases. With 3000 exemplars, we achieve almost the same performance as that of 5500 exemplars in terms of spectral distortion . The results on the development set are consistent with those on the evaluation set. Note that 3000 exemplars yield about 2 times faster computation and about half memory usage comparing with 5500 exemplars. We hence use $N = 3000$ exemplars in the following experiments.

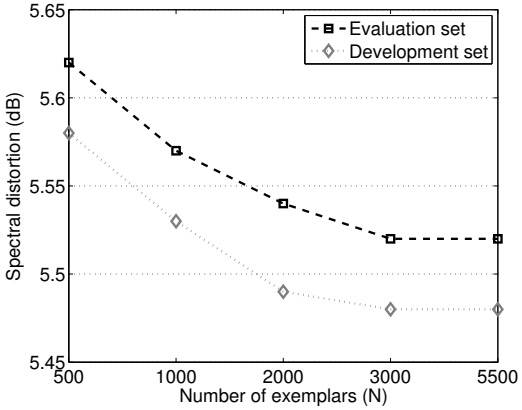

Fig. 4: Spectral distortion as a function of the number of exemplars N in a dictionary.

2) Analysis of the NMF-based methods: By setting the number of exemplars in the dictionary to be $N = 3000$, we assessed the performance of the incremental modifications to the basic exemplar-based sparse representation setup. The spectral distortions of the NMF-based methods are presented in Table II, and discussed in details in this Section. Without voice conversion, the spectral distortions between reference source and target MCCs are 7.77 dB and 7.91 dB on the development and evaluation sets, respectively.

TABLE II: Comparison of spectral distortions of the NMF-based methods

Conversion method	Window size	Spectral distortion (dB)	
		Development	Evaluation
No conversion	n/a	7.77	7.91
NMF-DFT	1	5.48	5.52
NMF-DFT-SC	1	5.05	5.13
NMF-MEL-SC	1	5.08	5.18
NMF-DFT	7	5.25	5.31
NMF-DFT-SC	7	4.91	4.99
NMF-MEL-SC	9	4.93	5.03
NMF-MEL-SC-RC	9	n/a	4.92

Firstly, we examined the effectiveness of spectral compression by comparing the NMF-DFT and NMF-DFT-SC methods. Correspond to Fig. 2, spectral distortions as a function of the varied compression factors on both development and evaluation sets are presented in Fig. 5. We observed consistent behavior in both development and evaluation sets. A compression factor of 0.4 gives the lowest spectral distortions of 5.05 dB and 5.13 dB on the development and evaluation sets, respectively. We hence choose $\rho = 0.4$ as the compression factor to represent NMF-DFT-SC method as shown in Table II in the reminding experiments. When a compression factor is set to 1.0, NMF-DFT-SC is reduced to the NMF-DFT method.

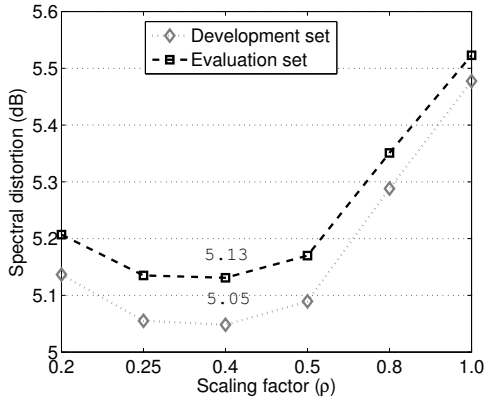

Fig. 5: Spectral distortion as a function of spectral compression factors (ρ).

Secondly, we examined the effect of feature dimensionality reduction in the source dictionary by comparing the NMF-DFT-SC and NMF-MEL-SC methods. As shown in Table II, NMF-DFT-SC method which uses high-resolution source dictionary to estimate the activation matrix produces a spectral distortion of 5.13 dB on the evaluation set. On the other hand, NMF-MEL-SC gives a slightly higher spectral distortion of 5.18 dB on the same set after performing dimensionality reduction.

We also examined the computational costs between the NMF-DFT-SC and the NMF-MEL-SC methods. The computing time was computed over all the testing data, and the average performance for generating one second of speech was reported. The computational costs to generate one second of speech as a function of the window sizes in an exemplar are presented in Fig. 6. We note that the computing time was calculated only for the 500 iterations' multiplicative updates. The MATLAB [2] codes were run on graphics processing unit (GPU), called GeForce GTX TITAN [3]. In general, the NMF-MEL-SC method is about 7 times faster than the NMF-DFT-SC method.

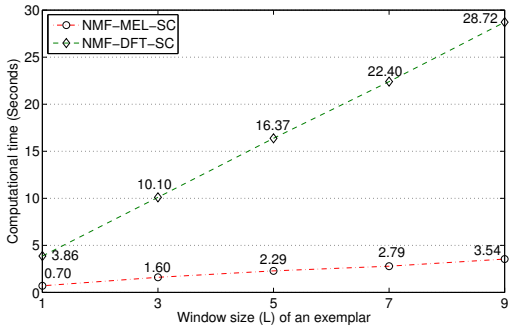

Fig. 6: Computing time to generate one second of speech as a function of the window size (L) of an exemplar.

Thirdly, we examined the effect of using multiple-frame exemplars to assess whether an exemplar spanning multiple consecutive frames is useful. Fig. 7 presents the spectral distortion results as a function of the window size L of an exemplar on both development and evaluation sets. For NMF-MEL-SC method, it is observed that as the window size increases, the spectral distortions consistently decrease and reach their minimum at 9 on both development and evaluation sets. While for NMF-DFT-SC method, spectral distortions have similar trends, but they reach their minimum at 7 and 9 on the development and evaluation sets, respectively. Due the heavy computations, we did not test the performance of NMF-DFT-SC when the window sizes were larger than 9. When the window size equals to 9, the source dictionary size of NMF-DFT-SC is $3000 \times (513 \times 9) = 3000 \times 4617$, while that of NMF-MEL-SC is $3000 \times (50 \times 9) = 3000 \times 450$.

We analyzed the activation weight for NMF-MEL-SC by setting the window size of an exemplar to 9. Fig. 8 presents an example of activation weights calculated for a single observation. In the example, there are only 13 exemplars that have weights greater than 0.01, and only two of them have weights greater than 0.05. For further analysis, we calculated the average top weights over one utterance. Given an utterance, the activation matrix was calculated; then the activations corresponding to each source frame were sorted

[2]http://www.mathworks.com/products/matlab/

[3]http://www.geforce.com/hardware/desktop-gpus/geforce-gtx-titan

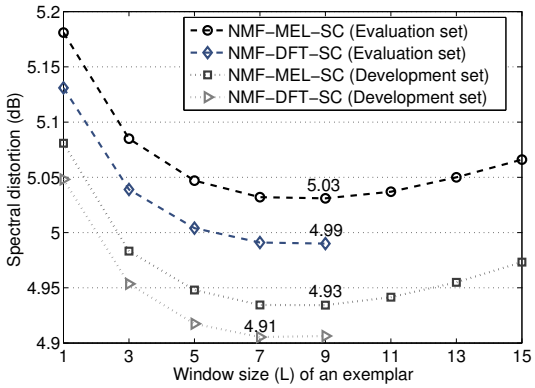

Fig. 7: Spectral distortion as a function of the window size (L) of an exemplar.

in a descending order; after that, the activations weights were averaged over all the source frames in the utterance. The top 100 averaged weights are presented in Fig. 9. It shows that the top 30 or 50 exemplars contribute more to the generated target spectrogram, while the other 2950 exemplars have weights that almost equal to zero. It implies only 1 % or even fewer exemplars are activated in generating each target spectrum, and confirms that effectiveness of the sparsity constraint.

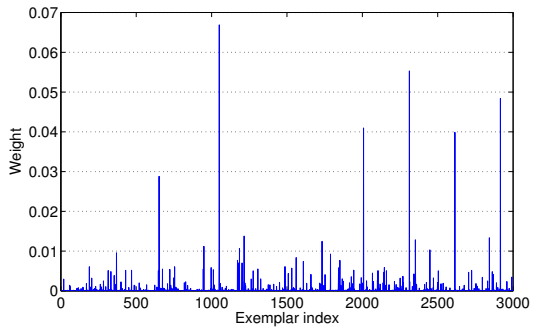

Fig. 8: Illustration of the activation weights associated to each exemplar to generate a target spectrum.

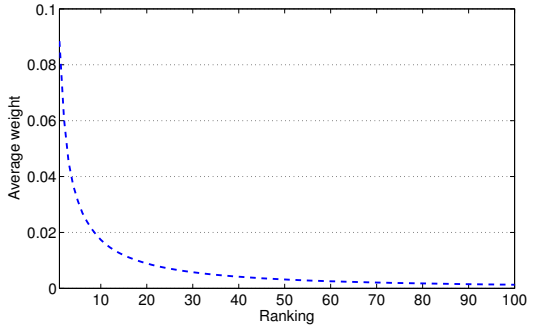

Fig. 9: Illustration of top 100 activation weights out of the 3000 exemplars, which are averaged over an utterance.

Lastly, we assessed the effectiveness of residual compensation. We first applied the NMF-MEL-SC method using 9-frame exemplars on the development set to produce the residuals. Then, a mapping between the source and target residuals was established using dynamic partial least squares (DPLS) regression. 20 latent factors were adopted in the DPLS regression based on the performance of DPLS-DFT which was tuned on the development set. After that, at runtime conversion, the same NMF-MEL-SC method was applied to each source utterance to predict a converted spectrogram and at the same time generate a source residual. The pre-trained DPLS mapping function was applied to the source residual to predict the target residual, which was compensated to the converted spectrogram.

We applied our method on the evaluation set. As shown in Table II, it is observed that residual compensation is able to reduce the spectral distortion from 5.03 dB for NMF-MEL-SC to 4.92 dB for NMF-MEL-SC-RC. It implies the effectiveness of residual compensation.

3) Overall performance assessment: We compared the proposed NMF-MEL-SC-RC method with the state-of-the-art baselines, namely, ML-GMM method and PLS-based methods. Table III presents the spectral distortions on the development and evaluation sets. Note that only $N = 3000$ exemplars rather than all the exemplars are adopted in the dictionary for the NMF-MEL-SC-RC method, while ML-GMM and PLS-based methods use all the frames $N \approx 5500$ in the training set. In the ML-GMM and ML-GMM-GV methods, the number of Gaussian components was set to 32 based on the MCD results on the development set. It is observed that on the evaluation set, the ML-GMM and ML-GMM-GV methods achieve spectral distortions of 5.19 dB and 5.71 dB, respectively, and the DKPLS method which performs nonlinear mapping gives a spectral distortion of 4.95 dB. Two variants of DKPLS, namely, DKPLS-DFT and DPLS-DFT, are applied on the high-resolution features and produce spectral distortions of 5.13 dB and 5.26 dB, respectively. On the same set, our NMF-MEL-SC-RC method achieves 4.92 dB, which is lower than both the ML-GMM method and the three PLS-based methods.

TABLE III: Comparison of spectral distortions of the proposed and baseline methods

Conversion method	Window size	Spectral distortion (dB)	
		Development	Evaluation
No conversion	n/a	7.77	7.91
ML-GMM	3	5.09	5.19
ML-GMM-GV	3	5.64	5.71
DKPLS	3	4.88	4.95
DKPLS-DFT	3	5.07	5.13
DPLS-DFT	3	5.18	5.26
NMF-MEL-SC-RC	9	n/a	4.92

We then examined the flexibility of the proposed NMF-MEL-SC-RC comparing with the DKPLS method. The number of exemplars in the dictionary or the number of frames as training was varied from 500 to 3000. For a fair comparison, the development set for training the residual mapping was also varied accordingly. The spectral distortions on the evaluation set as a function of the number of exemplars/frames is presented in Fig. 10. We observe that the NMF-MEL-SC-RC method has a similar behavior to the NMF-DFT method as shown in Fig. 4, and that the effect of the DKPLS method is consistent with [21]. It is worth noting that NMF-MEL-SC-RC is more stable than the DKPLS method even when the

training data is limited. Note that the exemplars/frames were randomly selected from the training set and high-resolution features for NMF-MEL-SC-RC were paired with MCCs used in DKPLS.

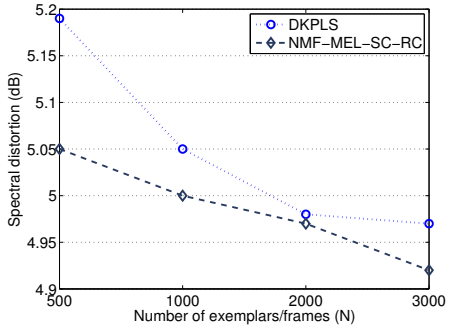

Fig. 10: Spectral distortions on the evaluation set as a function of the number of exemplars/frames N as training.

In general, our NMF-MEL-SC-RC method robustly achieves lower spectral distortions with varying the training data, and also works well on high-resolution features. PLS-based methods give lower spectral distortions on low-resolution features, but the performance drops considerably when applied to high-resolution features.

C. Subjective evaluations

We conducted listening tests to compare the performance between our proposed NMF-MEL-SC-RC and the baseline methods in terms of speech quality and speaker individuality. Amazon Mechanical Turk (AMT)[4], a crowd sourcing platform, was used in each listening test. The same platform has been earlier used in subjective evaluations, e.g. in [47], [48], [49]. In the evaluation set, there were 12 conversion pairs and each pair had 20 utterances. Hence, in total there were 240 (12×20) converted utterances. In the listening test, 20 utterances were randomly selected from the 240 utterances for each listener (or worker as called in AMT) to avoid bias on utterances. Moreover, 3 golden standard pairs were randomly mixed with the 20 real testing utterances to prevent cheating as advised in [49]. In each test, ten paid listeners were involved.

1) Analysis of the NMF-based methods: Four preference listening tests were conducted to assess the effect of the incremental modifications to the basic sparse representation setup. We focused on speech quality in this section.

Firstly, we performed an AB preference evaluation between NMF-DFT and NMF-DFT-SC methods regarding the effect of spectral compression. In an AB preference test, speech samples converted by two methods, namely NMF-DFT and NMF-DFT-SC, were presented to listeners in a random order, and the listeners were asked to choose the one that sounded more natural. Fig. 11 presents the preference scores. It shows that the preference scores are consistent with the spectral distortions and NMF-DFT-SC achieves significantly better speech quality than NMF-DFT without spectral compression. In general, both objective and subjective evaluations confirm the effectiveness of spectral compression.

[4]https://www.mturk.com

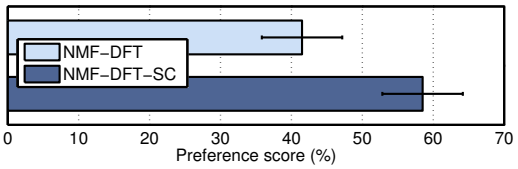

Fig. 11: Preference test results of speech quality with 95% confidence intervals for NMF-DFT and NMF-DFT-SC methods.

Secondly, a three-way preference test was conducted to examine the effect of the feature dimensionality reduction in the source dictionary. Different from the AB preference test, the three-way test had three options. Two speech samples generated from NMF-DFT-SC and NMF-MEL-SC methods were randomly presented to each listener, and then the listeners were asked to decide which sample is more natural. If they were not able to perceive the difference between two samples, they were asked to choose the option claiming no preference. The preference test results are presented in Fig. 12. It is observed that around 60 % sample pairs have the same speech quality, while NMF-DFT-SC achieving around 25 % preferences is slightly better than NMF-MEL-SC of around 15 %, but the difference is not statistically significant. Again, note that NMF-MEL-SC was around 7 times faster than NMF-DFT-SC and that there was 10 times memory reduction of NMF-MEL-SC comparing with NMF-DFT-SC. We hence conclude that NMF-MEL-SC with slightly performance drop is computationally more efficient than NMF-DFT-SC.

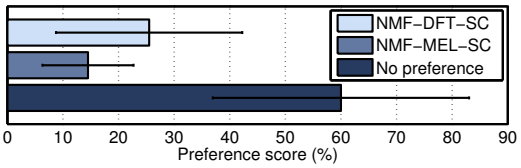

Fig. 12: Preference test results of speech quality with 95 % confidence intervals for NMF-DFT-SC and NMF-MEL-SC methods.

Thirdly, we conducted an AB preference test to assess the effectiveness of using multiple-frame exemplars in the source dictionary. NMF-MEL-SC using a single frame as an exemplar was compared with that using a nine-frame speech segment as an exemplar. Fig. 13 presents the subjective evaluation results. It shows that NMF-MEL-SC using nine-frame exemplars is significantly better than that using single-frame exemplars, and confirms the effectiveness of using multiple-frame exemplars. The subjective results are consistent with the objective spectral distortions.

Lastly, we conducted an AB preference test to examine the effectiveness of residual compensation by comparing NMF-MEL-SC and NMF-MEL-SC-RC. Fig. 14 presents the preference scores. It is observed that NMF-MEL-SC-RC achieves a significantly higher preference score than that of the

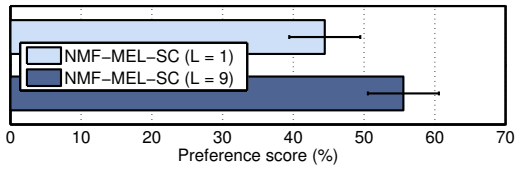

Fig. 13: Preference test results of speech quality with 95 % confidence intervals for NMF-MEL-SC with and without multiple-frame exemplars.

NMF-MEL-SC method. The preference results are consistent with the objective evaluations. In general, NMF-MEL-SC-RC achieves a lower spectral distortion and better than the NMF-MEL-SC method.

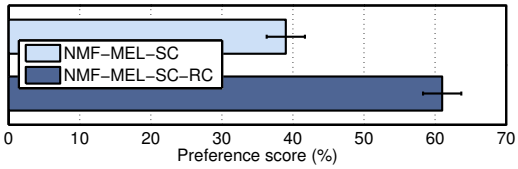

Fig. 14: Preference test results of speech quality with 95 % confidence intervals for NMF-MEL-SC and NMF-MEL-SC-RC methods.

2) Overall speech quality assessment: We conducted subjective evaluations to assess the speech quality between the proposed and baseline methods.

Firstly, we conducted AB preference tests to assess the speech quality between the proposed NMF-MEL-SC-RC method and PLS-based methods, where PLS-based methods used the whole training set $N \approx 5500$, while our method used $N = 3000$ exemplars. Fig. 15 presents the preference test results. In Fig. 15a, it is observed that the proposed NMF-MEL-SC-RC achieves similar performance to DKPLS method, in the sense that each method's preference score falls into the other method's confidence intervals. Figs. 15b and 15c show that our method is significantly better than both DKPLS-DFT and DPLS-DFT methods. As the three PLS-based methods were compared with the same NMF-MEL-SC-RC method, the preference scores imply that even though DKPLS works well with low-resolution MCC features, the performances of DKPLS as well as DPLS are degraded considerably in face of high-dimensional features. We conclude that PLS-based methods are not as flexible as our NMF-MEL-SC-RC method in handling high-dimensional features. The preference scores are consistent with the spectral distortions as shown in Table III.

Secondly, we conducted an AB preference test to assess the flexibility of the NMF-MEL-SC-RC and DKPLS methods. Here, both methods used 500 exemplars/frames as training. Fig. 16 presents the preference scores with 95 % confidence intervals. It is observed that the proposed NMF-MEL-SC-RC method is slightly better than the DKPLS method when limited training data are available, but the difference is not statistically significant. Even though previous result with 5500/3000 exemplars shows that the two methods achieve almost the

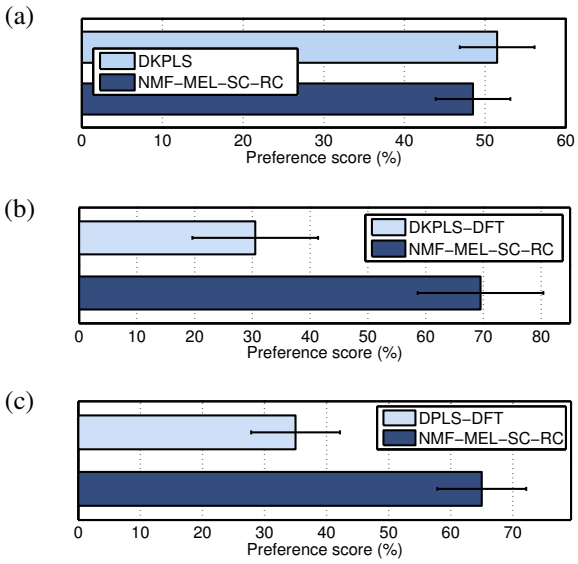

Fig. 15: Preference scores of speech quality with 95% confidence intervals for PLS-based methods with $N \approx 5500$ training frames and our proposed NMF-MEL-SC-RC method with $N = 3000$ exemplars in the dictionary. From top to bottom, (a), DKPLS vs. NMF-MEL-SC-RC; (b), DKPLS-DFT vs. NMF-MEL-SC-RC; and (c), DPLS-DFT vs. NMF-MEL-SC-RC.

same performance, the result using 500 exemplars shows the superiority of the proposed NMF-MEL-SC-RC. It confirms the flexibility of the proposed NMF-MEL-SC-RC, and it is consistent with the spectral distortions as presented in Fig. 10.

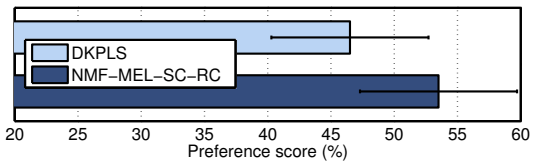

Fig. 16: Preference test results of speech quality with 95 % confidence intervals for DKPLS and NMF-MEL-SC-RC methods. Both methods used 500 exemplars/frames as training.

Next, we conducted a mean opinion score (MOS) test to compare the NMF-MEL-SC-RC and ML-GMM methods, as the ML-GMM method is a well-established baseline. The opinion score was set to a five-point scale: 1 = bad, 2 = poor, 3 = fair, 4 = good and 5 = excellent. 20 speech pairs were randomly selected from the 240 pairs, and each pair consisted of two conversion samples from the ML-GMM and NMF-MEL-SC-RC methods. The language content of the paired conversion samples was exactly the same. 3 pairs with natural speech were mixed with the 20 real testing pairs as golden standard pairs to exclude cheaters. The average MOS results with 95% confidence intervals are presented in Fig. 17. ML-GMM has a MOS of 2.49, while NMF-MEL-SC-RC achieves 3.15. It clearly shows that the proposed NMF-MEL-SC-RC method significantly outperforms the baseline ML-

GMM method.

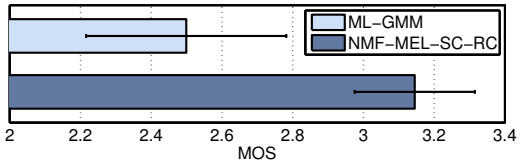

Fig. 17: Mean opinion scores with 95% confidence intervals for ML-GMM and NMF-MEL-SC-RC methods.

Finally, we conducted another MOS test to compare the NMF-MEL-SC-RC and ML-GMM-GV methods. It is generally believed that the ML-GMM-GV method is able to avoid the over-smoothing problem in conventional statistical parametric methods, and produces natural converted speech. We followed the same procedures as that for comparing the NMF-MEL-SC-RC and ML-GMM methods. Fig. 18 presents the average MOS results with 95 % confidence intervals. It is observed that ML-GMM-GV achieves a MOS of 3.07, while NMF-MEL-SC-RC gives 2.95. Although the average MOS of ML-GMM-GV is slightly higher than NME-MEL-SC-RC, the ML-GMM-GV method does not outperform our proposed NMF-MEL-SC-RC method significantly, in the sense that the MOS of each method falls into the other method's confidence intervals. We note that the scores in Figs. 17 and 18 cannot be compared directly, as the two tests were conducted independently and the listeners may not be exactly the same for both tests.

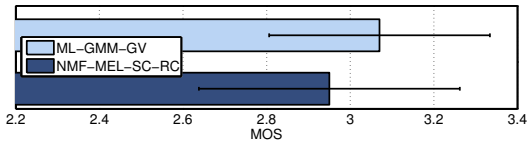

Fig. 18: Mean opinion scores with 95% confidence intervals for ML-GMM-GV and NMF-MEL-SC-RC methods.

3) Identity evaluations: We conducted subjective evaluations to assess the speaker similarity/individuality performance. From our previous experience [28], speech quality would affect the speaker individuality evaluation results in a preference test, as the listeners usually paid more attention on speech quality and preferred to choose samples that sound more natural. To this end, we conducted an XAB test for each method independently for fair comparison. In the test, 20 pairs were presented to the listeners, and each pair consisted of three speech samples: X = a converted sample, A = a source sample and B = a target sample. The language content of paired A and B was the same while that of X was different to make sure the listeners focus on the spectral attributes other than prosodic patterns. During the test, the converted sample was first presented as a reference. Then, source and target samples were presented in a random order. The listeners were asked to decide whether sample A or B sounded closer to X in terms of speaker individuality. The identification rate, which is the percentage of converted samples identified as target, was reported. Similar to [12], [21], all the inter-gender conversion pairs were identified correctly in an initial listening test. We hence only reported the results of intra-gender conversion, which was a more challenging task than the inter-gender task.

Fig. 19 presents the identification rate results. The results suggest that the NMF-based methods achieve slightly higher identification rates than the baseline ML-GMM method, and similar identification rates to the ML-GMM-GV method. In particular, NMF-MEL-SC-RC achieves an identification rate of 79.50 % while ML-GMM and ML-GMM-GV reaches 73.50 % and 80.00 %, and DKPLS attains 77.00 %. It is also observed that DPLS-DFT gives the lowest identification rate of 67.00 %. We note that the ML-GMM-GV achieves the smallest variance of the identification rates from all the listeners.

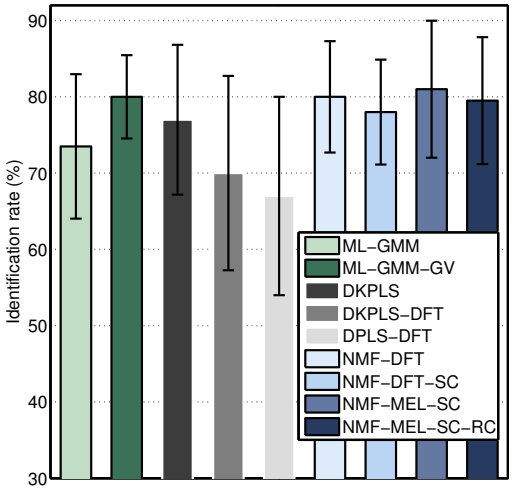

Fig. 19: Comparison of speaker identification rates of different voice conversion approaches.

In general, both the speech quality and identity tests confirm the effectiveness and flexibility of our proposed NMF-MEL-SC-RC method over a large set of state-of-the-art baseline methods. The subjective results are consistent with the objective spectral distortions.

V. Discussion

The experimental results have confirmed the effectiveness of the proposed exemplar-based sparse representation with residual compensation framework. Even though it is a non-parametric framework, there are some fundamental relationships between the proposed NMF-MEL-SC-RC method and the state-of-the-art methods as discussed in this section.

A. Relationship with vector quantization and frame selection methods

The vector quantization (VQ) and frame selection are two related methods that use original feature vectors to generate the target speech. In the VQ method, a codebook is established as a subset of the source-target feature pairs. At runtime conversion, the codebook is employed to find a target feature whose paired source feature is close to the given source feature. A similar method named frame selection [50], also

called unit selection [51], establishes a source-target correspondence during offline training. At runtime conversion, the correspondence is applied to select a target feature vector given a source feature vector.

The advantage of using the VQ and frame selection is that the selected target feature keeps the spectral details as original one. However, the hard clustering nature of VQ results in the incoherence phenomenon across frames. This phenomenon is also usually observed in the frame selection method [52]. On the other hand, in the proposed exemplar-based sparse representation method, exemplars in the paired source-target dictionaries can be treated as entries of a codebook in the VQ method, or source-target correspondence in the frame selection method. The proposed exemplar-based sparse representation with residual compensation method differs from the VQ and frame selection methods by presenting an observation as a linear combination of exemplars, whereas the VQ and frame selection methods calculate the selection costs for exemplars/frames independently from each other.

In addition, the frame selection method usually requires a significant amount of training data, while our method works with even 500 exemplars, that is 2.5 seconds of speech. Similar to the work in [50], our method can also be combined with the frame selection method, in the way that the converted speech by our method is used as the reference target speech for frame selection.

B. Relationship with JD-GMM methods

In JD-GMM based methods, the joint mean vectors $\boldsymbol{\mu}_k^{(z)}$ and covariance matrices $\boldsymbol{\Sigma}_k^{(z)}$ are employed to establish a conversion function to predict a target feature $\hat{\mathbf{y}}$ given a source feature $\mathbf{x}$:

$$\hat{\mathbf{y}} = \sum_{k=1}^{K} p_k(\mathbf{x})(\boldsymbol{\mu}_k^{(y)} + \boldsymbol{\Sigma}_k^{(yx)}(\boldsymbol{\Sigma}_k^{(xx)})^{-1}(\mathbf{x} - \boldsymbol{\mu}_k^{(x)})), \quad (19)$$

$$\boldsymbol{\mu}_k^{(z)} = [\boldsymbol{\mu}_k^{(x)}; \boldsymbol{\mu}_k^{(y)}] = \sum_{n=1}^{N} \mathbf{z}_n \cdot \gamma_{n,k}, \quad (20)$$

$$\boldsymbol{\Sigma}_k^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_k^{(xx)} & \boldsymbol{\Sigma}_k^{(xy)} \\ \boldsymbol{\Sigma}_k^{(yx)} & \boldsymbol{\Sigma}_k^{(yy)} \end{bmatrix} = \sum_{n=1}^{N} \gamma_{n,k} \cdot (\mathbf{z}_n - \boldsymbol{\mu}_k^{(z)})(\mathbf{z}_n - \boldsymbol{\mu}_k^{(z)})^\top, \quad (21)$$

where $\gamma_{n,k}$ is the occupation probability of the n^{th} frame belonging to the k^{th} Gaussian component [2], [12], and $p_k(\mathbf{x})$ is the posterior probability of the source feature $\mathbf{x}$ generated from the k^{th} Gaussian component.

On the other hand, as presented in Eq. (17), if the residual compensation is performed on a linear amplitude scale and the compression factor ρ is set to 1.0, the predicted target feature $\mathbf{y}^{(\text{DFT})}$ can be presented as

$$\hat{\mathbf{y}}^{(\text{DFT})} = \mathbf{B}^{(\text{DFT})}\mathbf{h} + \mathcal{F}(\mathbf{r}^{(\text{X})}), \quad (22)$$

where $\mathbf{r}^{(\text{X})}$ is one column of $\mathbf{R}^{(\text{X})}$.

Comparing Eqs. (19), (22) and (20), it is observed that both $\mathbf{B}^{(\text{DFT})}\mathbf{h} = \sum_{n=1}^{N} \mathbf{b}_n^{(\text{DFT})} \cdot h_n$ and $\boldsymbol{\mu}_k^{(\text{y})} = \sum_{n=1}^{N} \mathbf{y}_n \cdot \gamma_{n,k}$ do conversion as a linear combination of either $\mathbf{b}_n^{(\text{DFT})}$ or $\mathbf{y}_n$. Note that $\mathbf{b}_n^{(\text{DFT})}$ is a high-resolution spectrum, and $\mathbf{y}_n$ is the corresponding low-resolution MCC feature. The activation vector $\mathbf{h}$ is constrained to be sparse, while $\gamma_k = [\gamma_{1,k}, \gamma_{2,k}, \cdots, \gamma_{n,k}, \cdots, \gamma_{N,k}]$ does not have such a constraint. Thus, γ_k interpolates training samples to generate the mean vectors, in some sense, it tries to use as many as possible training samples to represent an unseen sample; while the sparse representation method uses a minimum number of samples to describe the unseen sample. For example, if a testing sample is included in the training, sparse representation similar to the VQ method is able to find the exact sample, while GMM-based approaches interpolate the whole training sample space. In this way, our proposed exemplar-based sparse representation with residual compensation method is able to avoid the over-smoothing effect introduced by the statistical average.

Moreover, as discussed in [53], the entries in $\mathbf{h}$ are conditionally dependent on each other given the dictionary and observation. However, the entries in γ_k are dependent only through the scalar normalization constant, otherwise they are independent from each other. Thus, $\mathbf{h}$ is able to benefit from the dependencies of exemplars for regression, while γ_k cannot.

C. Relationship with PLS

In partial least squares regression, given parallel data $\mathbf{X} \in \mathcal{R}^{d \times N}$ and $\mathbf{Y} \in \mathcal{R}^{d \times N}$, we have such decompositions:

$$\mathbf{X} \approx \mathbf{OP}, \quad (23)$$

$$\mathbf{Y} \approx \mathbf{VQ}, \quad (24)$$

where $\mathbf{P} \in \mathcal{R}^{p \times N}$ and $\mathbf{Q} \in \mathcal{R}^{p \times N}$ are projections or factor matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and $\mathbf{O} \in \mathcal{R}^{d \times p}$ and $\mathbf{V} \in \mathcal{R}^{d \times p}$ are loading matrices corresponding to $\mathbf{P}$ and $\mathbf{Q}$, respectively.

As discussed in [54], a transformation matrix $\mathbf{T}$ can be estimated from Eq. (23) and (24), and applied to a given source feature $\mathbf{x}$ for predicting a target feature $\hat{\mathbf{y}}$ during runtime as

$$\hat{\mathbf{y}} = \mathbf{Tx} + \mathbf{T}(\boldsymbol{\mu}^{(\text{Y})} - \boldsymbol{\mu}^{(\text{X})}) \quad (25)$$

$$= \mathbf{YP}^\top(\mathbf{QXX}^\top\mathbf{P}^\top)^{-1}\mathbf{QXx} + \mathbf{T}(\boldsymbol{\mu}^{(\text{Y})} - \boldsymbol{\mu}^{(\text{X})}), \quad (26)$$

where $\boldsymbol{\mu}^{(\text{X})}$ and $\boldsymbol{\mu}^{(\text{Y})}$ are the mean vectors of source and target training data, respectively.

Eq. (25) has a considerable similarity to Eq. (22), if we treat $\mathbf{P}^\top(\mathbf{QXX}^\top\mathbf{P}^\top)^{-1}\mathbf{QXx}$ as the activation vector $\mathbf{h}$, and $\mathbf{Y}$ as the target dictionary. Our method is hence similar and comparable to DPLS-DFT, which also operates on high-resolution features. The difference between our method and DKPLS/DKPLS-DFT is that DKPLS methods perform the spectral mapping in kernel space which introduces nonlinearity, while our method does not.

There are two advantages of our method over PLS-based methods. First, the flexibility allows our method to be robust in handling high-resolution features and in face of limited training samples. The experimental results show that in face of high-resolution features and limited training samples, the performance of PLS-based methods is degraded considerably, while our method is still reliable in such scenario. Second, the experimental results also confirmed the scalability of our

method, in the sense that the performance of our method can be boosted by simply appending new exemplars to the dictionary without much tuning. However, for PLS-based methods, re-optimization is required in the face of new training data.

D. Computational complexity and memory footprint

In comparison to the reference methods, major drawbacks of our method are the computational complexity and memory footprint. From the perspective of computational complexity, to generate one second of target speech, the ML-GMM and DKPLS methods take 0.41 and 0.06 seconds, respectively, on a 2.3 GHz Intel i7 core when implemented in MATLAB. However, our method costs 19.02 seconds. The computational cost of our method is about 45 times higher than the ML-GMM method and about 295 times higher than the DKPLS method.

For the memory footprint, at runtime, the ML-GMM method only needs to store $32 + 2 \times 32 \times (48 + (48 + 24)) = 7712$ parameters, supposing 32 Gaussian mixtures, 24 dimensional MCCs, and static+delta coefficients. The DKPLS method only requires to store $(200 \times 3) * 24 = 14400$ parameters. However, our method needs to store the source and target dictionaries. The size of the target dictionary is 513×3000, and that of the source dictionary is $9 \times 50 \times 3000$. In total, the dictionaries' size is $513 \times 3000 + 9 \times 50 \times 3000 = 2889000$. Hence, the memory occupation of our method is about 375 times higher than the ML-GMM method, and is about 200 times higher than the DKPLS method.

VI. Conclusions

We proposed an exemplar-based sparse representation framework as an alternative nonparametric framework for voice conversion. The flexibility of this framework allows us to easily adapt it to new training data, and makes it more robust in handling high-resolution spectra directly to maintain spectral details for better speech quality. In addition, the use of coupled dictionaries avoids to estimate the correlation/covariance matrix which is required in conventional statistical methods and is problematically estimated when source-target feature pairs have a low correlation [35]. The experimental results confirmed the effectiveness of the proposed exemplar-based sparse representation with residual compensation method, which achieves a spectral distortion of 4.92 dB, a MOS of 3.15 and a speaker identification rate of 79.50 %, outperforming the baseline ML-GMM method which gives a spectral distortion of 5.19 dB, a MOS of 2.49 and a speaker identification rate of 73.50 %. Moreover, our method is also more flexible than PLS-based methods, and comparable with the ML-GMM-GV method.

Our main findings are:

- Sparse representation is able to produce relatively high quality speech. It allows us to model high-resolution features directly for spectral details, and the activation vector for regression is constrained to be extremely sparse, in this way, the over-smoothing effect can be avoided.
- Spectral compression is helpful. A compression factor is able to control the spectrum intensity and affect the estimation of activations as well as the spectrogram generation.
- Multiple-frame exemplars which are able to describe the time sequence structure of speech are beneficial to reduce spectral distortion and to produce better speech quality.
- Residual compensation works well to reduce spectral distortion and to enhance speech quality. The sparse representation modeling capacity can be boosted by compensating source model residuals to the converted spectrograms.

As an alternative framework, our method is complementary with the statistical parametric methods, which can be employed to perform residual compensation.

Currently, parallel data, which is not always available, is required to construct source-target dictionaries. It would be interesting to find a method to relax such a constraint. Moreover, the computation of exemplar-based sparse representation is considerably higher than the ML-GMM method. It is possible to reduce computational time by applying low-resolution features to estimate the activation matrix and by using a small set of exemplars in the dictionaries. We will continue those directions in near future as a follow-up work.

VII. Acknowledgement

The authors would like to thank Prof. H. Kawahara from Wakayama University, Japan, for sharing the MATLAB code of the STRAIGHT toolkit, and all the anonymous reviewers to for their efforts and valuable comments to improve our manuscript.

References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998.

[3] K. Park and H. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000.

[4] A. Mouchtaris, J. Van der Spiegel, P. Mueller, and P. Tsakalides, "A spectral conversion approach to single-channel speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1180–1193, 2007.

[5] M. Afify, X. Cui, and Y. Gao, "Stereo-based stochastic mapping for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[6] X. Cui, M. Afify, and B. Zhou, "Stereo-based stochastic mapping with context using probabilistic pca for noise robust automatic speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[7] T. Toda, A. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[8] T. Toda, K. Nakamura, H. Sekimoto, and K. Shikano, "Voice conversion for various types of body transmitted speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[9] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1988.

[10] K. Shikano, S. Nakamura, and M. Abe, "Speaker adaptation and voice conversion by codebook mapping," in *Proc. IEEE Int. Sympoisum on Circuits and Systems*, 1991.

[11] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[12] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[13] H. Zen, Y. Nankaku, and K. Tokuda, "Continuous stochastic feature mapping based on trajectory HMMs," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 2, pp. 417–430, 2011.

[14] H. Ye and S. Young, "Perceptually weighted linear transformations for voice conversion." in *Proc. Interspeech*, 2003.

[15] ——, "Quality-enhanced voice morphing using maximum likelihood transformations," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1301–1312, 2006.

[16] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[17] V. Popa, H. Silen, J. Nurminen, and M. Gabbouj, "Local linear transformation for voice conversion," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[18] M. Narendranath, H. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech communication*, vol. 16, no. 2, pp. 207–216, 1995.

[19] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[20] P. Song, Y. Bao, L. Zhao, and C. Zou, "Voice conversion using support vector regression," *Electronics letters*, vol. 47, no. 18, pp. 1045–1046, 2011.

[21] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 806–817, 2012.

[22] L.-H. Chen, Z.-H. Ling, Y. Song, and L.-R. Dai, "Joint spectral distribution modeling using restricted Boltzmann machines for voice conversion," in *Proc. Interspeech*, 2013.

[23] D. Sundermann and H. Ney, "VTLN-based voice conversion," in *Proc. the 3rd IEEE International Symposium on Signal Processing and Information Technology*, 2003.

[24] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003.

[25] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[26] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[27] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Trans. Audio, Speech and Language Processing*, vol. 21, no. 3, pp. 556–566, 2013.

[28] 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 (SSW8)*, 2013.

[29] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556–562, 2001.

[30] J. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[31] A. Hurmalainen, J. Gemmeke, and T. Virtanen, "Non-negative matrix deconvolution in noise robust speech recognition," in *ICASSP*, 2011.

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

[33] B. King, C. Févotte, and P. Smaragdis, "Optimal cost function and magnitude power for nmf-based speech separation and music interpolation," in *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2012.

[34] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 944–953, 2010.

[35] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, "Voice conversion with smoothed GMM and MAP adaptation," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, 2003.

[36] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992.

[37] F. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1984.

[38] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1208–1230, 2009.

[39] J. Dines, J. Yamagishi, and S. King, "Measuring the gap between HMM-based ASR and TTS," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1046–1058, 2010.

[40] T. Virtanen, "Unsupervised learning methods for source separation in monaural music signals," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. Springer US, 2006, pp. 267–296.

[41] J. Nam, G. J. Mysore, J. Ganseman, K. Lee, and J. S. Abel, "A super-resolution spectrogram using coupled plca." in *Proc. Interspeech*, 2010.

[42] D. Bansal, B. Raj, and P. Smaragdis, "Bandwidth expansion of narrowband speech using non-negative matrix factorization." in *Proc. Interspeech*, 2005.

[43] 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*, vol. 27, no. 3, pp. 187–207, 1999.

[44] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation." in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 1994.

[45] A. B. Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School of Science & Engineering at Oregon Health & Science University, 2001.

[46] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion." in *Proc. Interspeech*, 2012.

[47] M. K. Wolters, K. B. Isaac, and S. Renals, "Evaluating speech synthesis intelligibility using amazon mechanical turk," in *7th ISCA Speech Synthesis Workshop (SSW7)*.

[48] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, "Crowdmos: An approach for crowdsourcing mean opinion score studies," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[49] S. Buchholz and J. Latorre, "Crowdsourcing preference tests, and how to detect cheating." in *Proc. Interspeech*, 2011.

[50] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou, "Towards a voice conversion system based on frame selection," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[51] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[52] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

[53] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[54] J.-Y. Pan and J.-S. Zhang, "Large margin based nonnegative matrix factorization and partial least squares regression for face recognition," *Pattern Recognition Letters*, vol. 32, no. 14, pp. 1822–1835, 2011.