HHS Public Access

Author manuscript

IEEE/ACM Trans Audio Speech Lang Process. Author manuscript; available in PMC 2020 January 24.

Published in final edited form as:

IEEE/ACM Trans Audio Speech Lang Process. 2018 December ; 26(12): 2267–2276. doi:10.1109/TASLP.2018.2860682.

Structured Sparse Spectral Transforms and Structural Measures for Voice Conversion

Yunxin Zhao,
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211 USA

Mili Kuruvilla-Dugdale,
Department of Communication Science and Disorders, University of Missouri, Columbia, MO 65211 USA

Minguang Song
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211 USA

Abstract

We investigate a structured sparse spectral transform method for voice conversion (VC) to perform frequency warping and spectral shaping simultaneously on high-dimensional (D) STRAIGHT spectra. Learning a large transform matrix for high-D data often results in an overfit matrix with low sparsity, which leads to muffled speech in VC. We address this problem by using the frequency-warping characteristic of a source–target speaker pair to define a region of support (ROS) in a transform matrix, and further optimize it by nonnegative matrix factorization (NMF) to obtain structured sparse transform. We also investigate structural measures of spectral and temporal covariance and variance at different scales for assessing VC speech quality. Our experiments on ARCTIC dataset of 12 speaker pairs show that embedding the ROS in spectral transforms offers flexibility in tradeoffs between spectral distortion and structure preservation, and the structural measures provide quantitatively reasonable results on converted speech. Our subjective listening tests show that the proposed VC method achieves a mean opinion score of "very good" relative to natural speech, and in comparison with three other VC methods, it is the most preferred one in naturalness and in voice similarity to target speakers.

Keywords

Voice conversion; structured sparse spectral transform; NMF; frequency warping; objective measures

I. Introduction

VOICE conversion (VC) has been an active topic of research with many potential applications. In voice conversion, the spectral and prosodic features of a source speaker's

Corresponding author: Yunxin Zhao. zhaoy@missouri.edu;.

speech are modified to make the resynthesized speech sound like that from a target speaker. As spectral features carry significant personal voice characteristics, spectral conversion has been the focus of study by many researchers. Typical spectral conversion approaches include spectral mapping based on Gaussian mixture model (GMM) and hidden Markov models (HMM) [1]–[4], vector quantization [5], linear or locally linear spectral transformation [6], [7] and frequency warping [8], [9], exemplar-based sparse representations using NMF [10]–[12] with extensions to unit selection [13], [14], and nonlinear spectral transformation through neural networks [15]–[18]. In prosody conversion, a pitch contour is often linearly transformed so that the mean and standard deviation of the converted log F0 match that of the target speaker [1]. Time scale and pitch scale modifications for VC are also suggested in PSOLA analysis and synthesis [19]. Although much progress has been made over the years, the 2016 Voice Conversion Challenge [20], [21] has revealed that the speech quality of converted voices is still below that of natural speech by a large margin, and objective measures that closely correlate with human perception need to be developed to facilitate VC research.

One well recognized shortcoming in voice conversion is that the converted speech often sounds muffled. This is attributable to excessive smoothing on speech spectral and temporal structures when minimizing mean spectral distortion in VC. Several efforts have aimed to address this issue. The approach of global variance directly compensates for the reduced spectral contrast in VC by enhancing temporal variance in each feature dimension. Maximum-likelihood estimation of spectral parameter trajectories was employed to increase the global variance of converted speech [3], and constrained optimization was formulated to match the global variance of converted and target speech while minimizing mean spectral distortion [22]. The approach of frequency warping attempts to keep the spectral details of source speech while warping its frequency axis to match that of target speech. Based on low-D line-spectral-pair (LSP) GMM, piecewise linear frequency-warping functions were derived to match the formants of source and target speech [8] or to maximize spectral segment correlations of source and target speech [23]. Based on high-D STRAIGHT spectra, dynamic frequency warping was performed directly [9] or combined with exemplar-based sparse representation [24]. These methods of frequency warping all required a follow-up step of spectral energy correction, referred to as amplitude scaling or residue compensation. The approach of sparse learning attempts to discover structure through regularization, as commonly practiced in machine learning for high-D data, for example, sparse principal component analysis [25]. In exemplar–based voice conversion, parameter optimization was also regularized by a L-1 constraint [10]–[12].

Although it is desirable to assess speech quality by human listeners, subjective listening tests are time consuming and subject to variability due to listener experience and compliance. Meanwhile, current objective measures for VC put emphasis on spectral distortion but overlook the loss of spectral-temporal structure due to over smoothing. As such, the outcomes often do not correlate well with subjective judgement of speech quality, which becomes a barrier to VC research. In image processing, the traditional mean squared error (MSE) measure also does not correlate well with human perceived image quality, making automatic assessment of algorithms such as compression and enhancement a challenging task. The structural-similarity-based objective measure [26] that is focused on structural

degradation instead of MSE has been shown to correlate much better with human visual perception and has gained wide spread adoption. In speech enhancement and source separation, predicting the intelligibility of noisy and enhanced speech is an important issue. Several types of subband covariance measures (normalized by variance) have been proposed for this purpose [27], and the method STOI that computes subband correlation coefficients over a temporal integration interval of 384 ms has been shown successful [28].

In the current work, we investigate novel methods of spectral conversion and objective quality assessment to address the above issues in voice conversion. We propose to use structured sparse transforms for simultaneous frequency warping and spectral shaping on high-D STRAIGHT spectra. Converting high-D spectra directly preserves spectral details better than converting low-D spectral features, and accurate spectral envelop representation is important to speech timbre perception [29]. However, learning a large transform matrix for high-D spectra generally results in overfitting. As an overfit transform lacks the physical structure or sparsity underlying the conversion problem, irrelevant source spectral components may mix into a target spectral component to produce muffled speech in VC. In our approach, we reinforce the physical structure or sparsity property of spectral transforms by using source-target frequency-warping constraints to form region-of-support (ROS) in transform matrices and optimize the matrices by nonnegative matrix factorization (NMF) [30], making the elements within ROS fine tuned for spectral conversion and the elements outside ROS fixed as zeros, which is difficult to achieve by L-1 alone. To emphasize structural similarity between converted and target speech, we measure local and global spectral and temporal covariance; to emphasize structural contrasts in converted speech, we measure local and global spectral and temporal variances. The proposed methods are evaluated on ARCTIC dataset [31] with 12 speaker pairs, by using spectral distortion and the proposed structural measures, and by subjective listening tests on speech quality, voice similarity to target speakers, as well as naturalness.

The rest of this paper is organized as follows. In Section II, the proposed spectral transform method using different ROS structures is described. In Section III, the spectral and temporal structural measures are defined. In Section IV, experiments and results are detailed. In Section V, conclusions are drawn.

II. Structured Sparse Spectral Conversion

The task of converting a source magnitude spectrum $\mathbf{x} = [x_1 \cdots x_d]^T$ to a target magnitude spectrum $\mathbf{y} = [y_1 \cdots y_d]^T$ can be implemented by $\hat{\mathbf{y}} = \mathbf{W}\mathbf{x}$, with $\mathbf{W} = [w_{i,j}]_{d\times d}$ the transform matrix. If $\mathbf{W}$ is adequately sparse, then the transform may be interpreted as performing frequency warping and spectral shaping simultaneously, i.e.,

$$\hat{y}_i = \sum_{j:\varphi_\alpha(j)=i} w_{\varphi_\alpha(j),j} x_j \tag{1}$$

where $\varphi_\alpha(.)$ is the function that warps the j-th frequency of the source to the i-th frequency of the target, and $w_{\varphi_\alpha(j),j}$ scales the contribution of x_j to $\hat{y}_i$.

Given a set of source-target spectral pairs, **W** is commonly estimated by minimizing an average distortion between the target and the transformed spectra subject to an L-1 constraint. On the other hand, for magnitude spectra, a sparsity constraint in the form of a region of support (ROS) can be directly specified for **W**, where elements of **W** outside the ROS are all fixed to zeros. This direct embedding of a ROS in **W** is convenient when using the multiplicative parameter update algorithm (MPUA) of NMF [28]. In the iterative estimation procedure, MPUA keeps the ROS of the estimated **W** within the initialized ROS. Exploiting this property, we use source-target speakers' frequency warping paths to form different ROS's and obtain a family of spectral transforms to tradeoff spectral distortion and structure preservation.

A. Spectral Transform Estimation

Given parallel training speech of a pair of source-target speakers, dynamic time warping (DTW) is applied to generate aligned magnitude spectral sequences:

$$\mathbf{A} = [\mathrm{a}_1 \cdots \mathrm{a}_K]$$
$$\mathbf{B} = [\mathrm{b}_1 \cdots \mathrm{b}_K],$$

where $\mathbf{a_i} \in R^d$ and $\mathbf{b_i} \in R^d$, $i = 1,\ldots, K$, are the aligned spectral pairs of the source and target, respectively. In order to reduce spectral dynamic range for NMF, a cubic root compression ($x^{1/3}$) is applied to spectral components prior to spectral conversion, and the converted spectral components are decompressed prior to waveform synthesis, similar to the approach of [10].

Using K-L divergence $\mathscr{D}_{KL}(\cdot \,\|\, \cdot)$ [30] with a L-1 regularization, the cost function of approximating **B** by **WA** becomes:

$$J = \mathscr{D}_{KL}(\mathbf{B} \,|\, |\, \mathbf{WA}) + \lambda \|\mathbf{W}\|_1 \tag{2}$$

The MPUA-based iterative solution for **W** [30] is

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\left(\frac{\mathrm{B}}{\mathrm{WA}}\right)\mathbf{A}^T}{\mathbf{1}_{d\times K}\mathbf{A}^T + \lambda \mathbf{1}_{d\times d}} \tag{3}$$

where "⊗" and "–" stand for element-wise multiplication and division, respectively, and $\mathbf{1}_{d\times K}$ and $\mathbf{1}_{d\times d}$ are all-ones matrices (elements all being 1's). It is easy to see that the element-wise multiplication ensures the zeros in $\mathbf{W}^{(r)}$ of the r-th iteration to stay as zeros in $\mathbf{W}^{(r+1)}$ of the (r+1)-th iteration. Therefore a ROS can be initialized in $\mathbf{W}^{(0)}$ and be inherited in the converged **W**.

In Fig. 1(a), operations of the proposed spectral transform method at the training and conversion stages are illustrated, and for comparison, in Fig. 1(b) operations of a generic NMF-based exemplar approach are illustrated. In (a), at the training stage the parallel training spectral matrices $\mathbf{D_A}$ and $\mathbf{D_B}$ are used to learn a transform matrix $\widehat{\mathbf{W}}$, and at the

conversion stage, $\widehat{\mathbf{W}}$ is directly multiplied with the source spectral matrix $\mathbf{X}$ to generate the converted spectral matrix $\widehat{\mathbf{Y}}$. In (b), during training the aligned parallel training data $\mathbf{D_A}$ and $\mathbf{D_B}$ are produced as over-complete dictionaries, and during conversion NMF is first performed on the source spectral matrix $\mathbf{X}$ to derive an activation matrix $\widehat{\mathbf{H}}$ by using the source dictionary $\mathbf{D_A}$, and $\widehat{\mathbf{H}}$ is next transferred to the target dictionary $\mathbf{D_B}$ to produce the converted spectral matrix $\widehat{\mathbf{Y}}$. The activation $\widehat{\mathbf{H}}$ can also be estimated with the K-L criterion [10], [11]. It is easy to see that at the conversion stage, the transform approach has a much lower complexity than the exemplar approach, as the latter requires NMF for each source utterance, while the former only requires a matrix multiplication.

The method described in Fig. 1(a) is further integrated with GMM to obtain data-dependent probabilistic transforms. To do so, mel-frequency cepstral coefficients (MFCC) features of source training speech are analyzed with the frame length and shift identical to those used in STRAIGHT spectral analysis. A size-M GMM is estimated from the MFCC features by the EM algorithm, and the source MFCC vectors are partitioned into M clusters based on the maximum posterior probability rule [32]. The clustering structure is transferred to the aligned source-target magnitude spectral matrices $\mathbf{A}$ and $\mathbf{B}$ to yield M pairs of matrices $\{(\mathbf{A_m}, \mathbf{B_m}), m = 1,\ldots, M\}$ for estimating mixture-specific transforms $\widehat{\mathbf{W}}_{\mathrm{m}}$, $m = 1,\ldots, M$. During conversion, the source MFCC features are used to compute the posterior probabilities $\gamma_{m,t}$ for each model component m and frame t, and the source magnitude spectral vectors $\mathbf{x_t}$ are transformed as

$$\hat{\mathbf{y}}_{\mathbf{t}} = \sum_{m=1}^{M} \gamma_{m,t} \widehat{\mathbf{W}}_{\mathrm{m}} \mathbf{x}_{\mathrm{t}} \tag{4}$$

where $\sum_{m=1}^{M} \gamma_{m,t} \widehat{\mathbf{W}}_{\mathrm{m}}$ defines the posterior-probability-weighted transform at time t.

B. Frequency Warping Constrained ROS Embedding

In MPUA (3), the transform $\mathbf{W}$ is commonly initialized as an all-ones matrix [30]. Fig. 2(a) illustrates the ROS of such a matrix, and it is referred to as WNC hereafter. In this case, the task of discovering a sparse structure in $\mathbf{W}$ is left to the constrained estimation. Since for high-D spectral features the matrix $\mathbf{W}$ is large, an L-1 regularization is insufficient for obtaining a well structured $\widehat{\mathbf{W}}$. As the result, each target spectral component might be a weighted mixture of many source spectral components, making the converted speech sound muffled.

To address this problem, we investigate using source-target speakers' frequency warping paths to form a ROS constraint in $\mathbf{W}^{(0)}$, and let MPUA (3) to refine only those elements within the ROS of $\mathbf{W}$, which is easier than estimating an entire $\mathbf{W}$ matrix. For convenience, we define frequency warping paths by the bilinear frequency warping function [33]:

$$\varphi_{\alpha}(\omega) = \omega + 2\tan^{-1}\left(\frac{(1-\alpha)\sin(\omega)}{1-(1-\alpha)\cos(\omega)}\right) \tag{5}$$

where ω is angular frequency and each α value specifies one warping path, with $\alpha < 1$ warps frequency from low to high as in male-to-female conversion, and $\alpha > 1$ warps frequency from high to low as in female-to-male conversion. We quantize the parameter α to a set of values within a feasible range $[\alpha_{\min}, \alpha_{\max}]$, and consider three ways of forming a region of support: relaxed, intermediate, and strict. It is worth noting that previous efforts on using bilinear frequency warping for VC mostly operate on cepstral features [34]–[35], where a transform matrix is a function of α, but the warping constraint would not form a ROS to define a sparse matrix.

1) Wide Warping Range Support (WWR): This region of support is shown as the dotted area in Fig. 2(b), which is enclosed by the two extreme frequency warping paths defined by $\alpha_{\min}$ and $\alpha_{\max}$. Within this ROS, the elements of $\mathbf{W}^{(0)}$ are initialized as 1's, and outside it the elements are initialized as 0's. This ROS makes $\mathbf{W}$ sparse, but it still leaves NMF with a large degree of freedom within the ROS to learn relations of frequency warping and spectral shaping from the source-target parallel speech.

2) Single Warping Path Support (WSP): This region of support is shown as the dotted curve in Fig. 2(c), corresponding to the best frequency warping path for a source-target speaker pair. To determine the best path, α values are enumerated over the range of $\alpha_{\min}$ to $\alpha_{\max}$. For each α value, an ROS based on the corresponding warping path is embedded in $\mathbf{W}_{\alpha}^{(0)}$ (the dependency of $\mathbf{W}^{(0)}$ on α is made explicit here to explain the selection procedure for α or warping path) and MPUA (3) is performed on the source-target speakers' training data to give an estimated $\widehat{\mathbf{W}}_{\alpha}$. Frobenius norm is used to measure the error: $E_{\alpha} = \|\mathbf{B} - \widehat{\mathbf{W}}_{\alpha}\mathbf{A}\|_F$, and the best warping path is selected by $\alpha^* = \arg\min E_{\alpha}$. Because $\mathbf{W}_{\alpha}^{(0)}$ is very sparse, the selection procedure would not incur heavy computations. However, for discretized frequency bins, using a single path may leave certain target frequency bins outside the ROS. To prevent this from happening, if a target frequency bin is not on the best warping path, then it is linked with the source frequency bins that are warped to the neighboring target bins, and the ROS is extended accordingly.

3) Narrow Warping Range Support (WNR): This region of support is shown as the dotted area in Fig. 2(d), which is enclosed by the frequency warping paths that are nearest to the best warping path (see IV.A for details). Relative to WWR, the WNR method reduces the ROS for a source-target speaker pair, preventing irrelevant source spectral components from mixing into a target spectral component. Relative to WSP, the slightly relaxed ROS by WNR allows NMF to learn a more flexible frequency warping function for a source-target speaker pair, since a bilinear function is just an approximation to the actual frequency warping relation.

C. Examplar-Based Probabilistic Spectral Conversion

Here we sketch out an exemplar-based probabilistic spectral conversion method that is used in the comparative experiments in Section IV. By taking the GMM-clustered magnitude spectral matrix pairs $\{(\mathbf{A_m}, \mathbf{B_m}), m = 1,\ldots, M\}$ as $\boldsymbol{M}$ pairs of source-target exemplar

dictionaries, a source spectral matrix $\mathbf{X}$ is factorized in M ways as $\mathbf{X} \approx \mathbf{A}_{\mathrm{m}}\widehat{\mathbf{H}}_{\mathrm{m}}$, $m = 1,\ldots, M$. Transferring the activations $\widehat{\mathbf{H}}_{\mathrm{m}}$ to the target dictionaries $\mathbf{B_m}$ yields a converted spectral matrix as

$$\widehat{\mathbf{Y}} = \sum_{m=1}^{M} \mathbf{B}_{\mathrm{m}}\widehat{\mathbf{H}}_{\mathrm{m}}\mathbf{P}_{\mathrm{m}} \tag{6}$$

where $\mathbf{P_m} = \mathrm{diag}(\gamma_{m,1}, \cdots, \gamma_{m,N})$ and the $\gamma_{m,t}$'s are the posterior probabilities (Section II.A). The clustered exemplar dictionaries are shorter than a global dictionary and hence the activation matrices are smaller, which reduces computation in NMF. In [11], phoneme categorized dictionaries were predefined for voice conversion, and for each frame, one dictionary was selected and used. In contrast, the multiple dictionaries here are clustered by GMM, and the conversion combines the component dictionaries and activations probabilistically. The exemplar method as sketched here is only generic, enhancement such as using contextual frames [10]–[12] would perform better for VC but is beyond the scope of the current work.

III. Spectral and Temporal Structural Measures

The spectral and temporal structural measures include local and global covariance between converted and target speech and variance of converted speech in mel filters and frames, respectively. The spectral matrices of the target and converted speech are denoted by $\mathbf{Y} = [y_{i,j}]$ and $\widehat{\mathbf{Y}} = \left[\hat{y}_{i,j}\right]$, respectively.

A. Spectral Measures

In Fig. 3, the computation of a local covariance centered at a mel filter k_o (= 25) and at time t (= 0.6s) is illustrated. A window of $2B + 1$ mel filters is applied to the t-th columns of $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ to compute the local covariance

$$\sigma^{F}_{\hat{y}y}(k_o, t) = \frac{1}{2B+1} \sum_{k=k_o-B}^{k_o+B} \left(\hat{y}_{k,t} - \mu^{F,\hat{y}}_{k_o,t}\right)\left(y_{k,t} - \mu^{F,y}_{k_o,t}\right) \tag{7}$$

where $\mu^{F,y}_{k_o,t} = \frac{1}{2B+1}\sum_{k=k_o-B}^{k_o+B} y_{k,t}$ is the local spectral mean of $\mathbf{Y}$, and $\mu^{F,\hat{y}}_{k_o,t}$ is defined similarly for $\widehat{\mathbf{Y}}$.

Since the local scales of $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ affect the value of $\sigma^{F}_{\hat{y}y}(k_o, t)$, it is desired to normalize the spectral values within each window by their local means, which is equivalent to normalizing $\sigma^{F}_{\hat{y}y}(k_o, t)$ by the local means of $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$. This is preferable to using correlation coefficient that normalizes co-variance by standard deviations as in STOI, since converted speech in VC often suffers from reduced variance, and normalizing by standard deviations would artificially boost the normalized covariance values for the low-variance VC methods. Based on the notion that larger variance may suggest better speech quality and large covariance

may suggest higher similarity, we measure the two statistics separately to keep both of them in perspective.

Taking into account of the local scale normalization, the local spectral covariance at the utterance level (Cov_{LF}) is averaged on $\sigma^{F}_{\hat{y}y}(k_o, t)$ over (k_o, t) as

$$Cov_{LF} = \frac{1}{T(L-2B)} \sum_{t=1}^{T} \sum_{k_o=B+1}^{L-B} \frac{\sigma^{F}_{\hat{y}y}(k_O, t)}{\mu^{F,y}_{k_o,t} \mu^{F,\hat{y}}_{k_o,t}} \tag{8}$$

where L and T denote the number of mel filters and frames, respectively. The local spectral variances of the target and converted speech are simply $\sigma^{F}_{yy}(k_o, t)$ and $\sigma^{F}_{\hat{y}\hat{y}}(k_o, t)$. The utterance-level average local spectral variance (Var_{LF}) is defined similarly as in (8) with the covariance replaced by the respective variance terms. Additionally, global spectral covariance Cov_{GF} and variance Var_{GF} are obtained by extending the spectral window to cover all L mel filters.

B. Temporal Measures

Again in Fig. 3, the computation of a local covariance centered at a time frame t_o (= 1s) and at a fixed mel filter k (= 20) is illustrated. A window of $2Q + 1$ frames is applied to the k-th rows of $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ to compute the local covariance

$$\sigma^{T}_{\hat{y}y}(k, t_o) = \frac{1}{2Q+1} \sum_{t=t_o-Q}^{t_o+Q} \left(\hat{y}_{k,t} - \mu^{T,\hat{y}}_{k,t_o}\right)\left(y_{k,t} - \mu^{T,y}_{k,t_o}\right) \tag{9}$$

where $\mu^{T,y}_{k,t_o} = \frac{1}{2Q+1} \sum_{t=t_o-Q}^{t_o+Q} y_{k,t}$ is the local mean of $\mathbf{Y}$ and $\mu^{T,\hat{y}}_{k,t_o}$ is defined similarly for $\widehat{\mathbf{Y}}$. Taking into account of local scale normalization as in spectral covariance, the utterance-level average temporal local covariance (Cov_{LT}) is defined as

$$Cov_{LT} = \frac{1}{(T-2Q)L} \sum_{t_o=Q+1}^{T-Q} \sum_{k=1}^{L} \frac{\sigma^{T}_{\hat{y}y}(k, t_O)}{\mu^{T,y}_{k,t_o} \mu^{T,\hat{y}}_{k,t_o}} \tag{10}$$

The temporal local variances are $\sigma^{T}_{yy}(k, t_o)$ and $\sigma^{T}_{\hat{y}\hat{y}}(k, t_o)$ for the reference and converted speech, respectively. The utterance-level average local spectral variance (Var_{LT}) is defined similarly as in (10) with the covariance replaced by the respective variance terms. Again, global temporal covariance Cov_{GT} and variance Var_{GT} are defined by extending the temporal window to cover all frames of each utterance.

IV. Experiments and Results

We used the ARCTIC dataset [31] of two male (bdl, rms) and two female (clb, slt) speakers (12 source-target speaker pairs) in evaluations. The first 20 sentences a0001 through a0020 were for system training (the training data size matched that of [23], [24]), and the next 20

sentences a0021 through a0040 for conversion testing. TANDEM-STRAIGHT toolbox [36], [37] was used to generate magnitude spectra and pitch parameters and to synthesize converted speech. speech sampling rate was 16 kHz, and speech analysis used 1024-point DFT with the frame shift of 4 ms. HTK toolkit [38] was used to compute 39 MFCC parameters (13 MFCCs, 13Δ's and 13ΔΔ's) with the window size and frame rate identical to sTRAIGHT analysis. To construct the exemplar dictionaries, $\mathbf{D_A}$ and $\mathbf{D_B}$, DTW was first applied to the MFCC feature sequences of source and target training speech, and the alignments were then transferred to the STRAIGHT spectral sequences, where aligned frames with mismatched voicing features (voiced vs. unvoiced) were removed. In pitch conversion, the mean and standard deviation of the source speaker's logF0 contour were linearly transformed to match the statistics of the target speaker [1].

In addition to the proposed structured sparse spectral transform method with different ROS constraints, we conducted comparative experiments on the generic exemplar method discussed in Section II.C, referred to as HEX, the classical GMM method [1], and the global variance constrained GMM, referred to as CGMM [22]. The size of GMM was set to 8 for all of the VC methods. The hyper-parameters used for regularizing NMF in the spectral transform and exemplar methods were empirically chosen, where the four transform methods all used $\lambda = 0.2$, and the exemplar method used $\lambda = 1.0$ on the basis that an activation matrix may be quite large for a long utterance. Further details of L-1 regularization are discussed in Section IV.B.4. Among the three comparison VC methods, HEX is closely related to the proposed spectral transform method in using NMF and STRAIGHT vocoder, GMM is widely used in statistical VC, CGMM has a global variance enhancement to GMM, and the software code of the last two methods are publicly available [22]. Although we could not directly compare with many other VC methods due to our resource limitations, the outcomes of the current study allow indirect comparisons through results available in the literature such as the 2016 Voice Conversion Challenge [20], [21].

In objective evaluation, the conventional root-mean-square error (RMSE) of mel-filter bank log spectra and the spectral and temporal measures defined in Section III were used, where we compared the proposed spectral transform method using four types of ROS with the VC methods of HEX, GMM, and CGMM. In subjective evaluation, listening tests were conducted on mean opinion score (MOS), voice similarity to target speaker, and preference between voice conversion methods in terms of naturalness and voice similarity to target speaker. To make the load of the listening tests manageable, we only compared the spectral transform method WNR with the methods of HEX, GMM, and CGMM. The choice on WNR was based on an informal listening assessment and a consideration on the distortion-structure tradeoff. Relative to WNR, WNC and WWR both sounded somewhat muffled, consistent with their lower co-variance and variance values. On the other hand, though WSP had slightly higher covariance and variance values than WNR, its distortion was also larger, and WNR sounded slightly more natural than WSP occasionally.

A. Spectral Transformation

The four different ROS's in $\mathbf{W}^{(0)}$ (Fig. 2) were evaluated, where the same $\mathbf{W}^{(0)}$ was used to initialize every mixture component in a GMM for MPUA of NMF. The bilinear warping

parameter α was enumerated within [0.8 1.2] with a step size of 0.04. For the WSP method, given a source-target speaker pair, the best frequency-warping path parameterized by α_m was determined for the mixture-specific spectral matrix pair ($\mathbf{A_m}$, $\mathbf{B_m}$). The mode $\alpha^* = \text{mode}\{\alpha_m, m = 1,\ldots M\}$ (ties were broken randomly) then defined the best warping path and ROS for the speaker pair. For the WNR method, a ROS was formed in one of two ways. If two α_m values were found immediately next to α^* from each side, then the frequency warping paths parameterized by the two values were used to define the ROS; otherwise if only one α_m was found immediately next to α^*, then the frequency warping paths parametrized by the α_m and α^* itself were used to form the ROS. These were the two scenarios used in our experiments but other rules may certainly be introduced for WNR.

To gain insights on the estimated transform matrices, the selected α_m's and α^* (bold) for each speaker pair are detailed in Table I for WNR. It is observed that for within-gender voice conversion, the selected α was close to unity, indicating little or no warping, whereas for between-gender voice conversion, the selected α deviated from unity, with $\alpha < 1$ for male-to-female conversion and $\alpha > 1$ for female-to-male conversion, conforming to the needs for low-to-high and high-to-low frequency warping in these two directions. It is worth noting that the determined α_m's are meaningful for mixture components with mainly voiced spectra but not so for unvoiced spectra, and the α_m's in the latter mostly became outliers. With the mixture size $M = 8$, there were normally two mixture components being mainly unvoiced. Therefore, for each mixture component we computed a frequency-of-occurrence-based voicing probability p: if $p < 0.5$, then the component would not be considered in selecting α_m and α^*. On the other hand, the derived ROS was used in estimating the transforms for all mixture components. Furthermore, spectral energy at very high frequencies was too weak for reliable transform estimation. We empirically limited the spectral transforms to be below 7.55 KHz, where beyond the frequency the source spectral components were directly transferred to the target without conversion.

To examine the spectral shaping effect of the transform $\mathbf{W}$, we define a log-sum term for each target frequency index i, $(\log \mathbf{W})_i \triangleq \log(\sum_{\varphi_{\alpha(j)=i}} w_{\varphi_\alpha(j),j})$, to describe the scaling effect on source spectral energy. This term ties closely to the spectral transformation defined in (1). Fig. 4(a) shows the log $\mathbf{W}$ curve for converting the voice of speaker bdl to speaker slt, and Fig. 4(b) shows the curve of converting the voice of slt to bdl. The spectral-shaping effects of $\mathbf{W}$ are confirmed by the large dynamic ranges in both curves. The approximate peak-valley reversal between the two curves is appealing as they correspond to opposite conversion directions of the same two speakers. In Appendix, the simultaneous frequency warping and spectral shaping effect of the transform is further illustrated by 3D plots of the estimated $\mathbf{W}$'s for the four cases of ROS.

B. Objective Evaluations

The objective measure of root-mean-square error (RMSE) on mel-filter bank log spectra is defined as

$$E_{dB} = \sqrt{\frac{1}{L}\sum_{l=1}^{L}(10\log_{10}y_l - 10\log_{10}\hat{y}_l)^2} \quad (11)$$

where y_l and $\hat{y}_l$ are the target and converted spectral components of the l–th mel filter of two aligned frames. There were $L = 50$ triangular mel filters [39], covering the range of 133.33 Hz to 6855.5 Hz. The errors were averaged over all aligned spectral pairs.

The local spectral covariance and variance were computed over a wide range of window sizes. Because the resulting values exhibited consistent patterns over the large range, to save space we only report results for window size $(2B + 1)$ of 7, 11 and 15 mel filters. The temporal window sizes $(2Q + 1)$ were 35, 55, and 75 frames, corresponding to 200 ms, 280 ms, and 360 ms, respectively, based on the notion that auditory system's temporal integration time was within a few hundred miliseconds [40]. The windows were rectangle without tapering. In addition, global spectral and temporal covariance and variance were computed.

1) Spectral Distortion: In Table II, the RMSE results on the mel-filter spectra are provided for the seven VC cases. It is seen that within the spectral transform group, imposing the frequency-warping-based ROS increased RMSE, with the trend that the stronger the constraint, the larger the error. Similarly, within the GMM-based group, CGMM had a much larger error than GMM. The exemplar-based VC method had the lowest error among the seven methods.

2) Spectral Covariance and Variance: In Table III, the results of spectral covariance and variance are provided, and the variances of target speech (TAR) are also included for reference. Within the spectral transform group, imposing the ROS constraints in $\mathbf{W}^{(0)}$ improved spectral covariance and variance. one trend was that the stronger the constraints, the larger the covariance and variance, with the exception that WNR had a slightly larger global covariance than WSP. The other trend was that covariance and variance values increased with the size of the window. In comparison with the spectral transform methods, the exemplar method, HEX, had weaker covariance and variance. Within the GMM group, CGMM had much larger covariance and variance than GMM did. The local and global variances of the converted speech were below those of the target speech in general, suggesting reduced spectral contrasts in converted speech. The transform method of WSP gave the largest local covariance and variance values, and WNR was close to WSP. on the other hand, CGMM gave the largest global covariance and variance, and its global variance exceeded that of the target speech by 4% (the local covariance and variance of CGMM remained below those of WSP until the window size approached that of the global window).

3) Temporal Covariance and Variance: In Table IV, the results of local and global temporal covariance and variance are detailed. It is observed that WSP again gave the largest local covariance and variance, the covariance values given by WNR were comparable to those of WSP. While CGMM maintained high variance values, its covariance values were below those of WNR. The exemplar method gave slightly lower covariance and variance

values than WNC. The effect of imposing ROS in $\mathbf{W}^{(0)}$ on temporal structures was similar to that on spectral structures (Table III), and so was the effect of window size. At the global scale, WSP gave the highest covariance, while CGMM again gave the largest variance. Temporal variances of the converted speech remained below those of the target speech at the studied time scales.

Overall, the evaluation results of the spectral transform VC method indicated that the looser the constraint, the lower the distortion and the weaker the structural preservation, and vice versa. The ROS of WNR provided flexibility in tradeoffs between distortion and structure preservation. The spectral and temporal structural measures of covariance and variance provided informative patterns on converted speech that were not captured by global variance alone.

4) Effects of L-1 Regularization: We also studied the effect of the regularization hyper-parameter λ in (3) on spectral-transform-based VC within the range of $0.0 \leq \lambda \leq 10.0$. We found that spectral distortion increased with λ, consistent with the notion that sparsity regularization increases estimation bias. Unlike ROS embedding, increasing λ had insignificant effect on spectral and temporal covariance and variance. For example, relative to $\lambda = 0$, the largest increase in global spectral and temporal covariance was 0.87% and 0.5%, respectively, and in global variance the increase was 1.8% and 2.6%, respectively (obtained by $\lambda = 2$, beyond which both covariance and variance values decreased). As a contrast, relative to WNC, the global spectral and temporal covariance improvement by WSP was 9.99% and 14.68%, and in variance the increase was 31.74% and 60.69%, respectively. This suggests that L-1 regularization alone was insufficient for preserving speech spectral and temporal structures in voice conversion. Detailed results of this study are omitted here due to space limitation.

C. Subjective Evaluations

The listeners recruited were students of University of Missouri. Each student volunteer was screened for hearing loss and was included only if he or she passed hearing screening at 500 Hz, 1 k, 2 k, 4 k, and 8 kHz using a 20 dB HL pure tone. The number of listeners varied in different listening tests, which were constrained by the hearing screening outcomes and students' availabilities.

1) Speech Quality Assessment: In this test, the 5-scale Mean Opinion Score (MOS) (1 = bad, 2 = fair, 3 = good, 4 = very good, 5 = excellent) was used to assess speech quality of the four voice conversion methods: HEX, WNR, CGMM, and GMM. In addition, the perceived quality of natural speech (NAT) and speech processed by the STRAIGHT vocoder (VOC) (without VC) were assessed. There were 42 listeners, each listened to a sound track of 56 audio samples, with 48 samples from the combination of 4 VC methods and 12 source-target speaker pairs (4 × 12), 4 samples from the 4 ARCTIC speakers, and 4 samples from the vocoded speech of the 4 ARCTIC speakers. The sentence utterances were randomly taken from the test set. The mean and standard deviation (stdev) of the MOS scores are given in Table V.

In Table V, the MOS score of the STRAIGHT vocoded speech (VOC) was slightly inferior to that of the natural speech (NAT), while the converted speech all had lower MOS scores than that of VOC. Among the 4 VC methods, the proposed WNR had the highest score of 3.74. As the natural speech MOS was 4.47, we also rescaled all MOS scores relative to NAT score = 5, and the scores are shown in the last row of the table. It is seen that relative to MOS = 5 for natural speech, the MOS rating of WNR was “very good.” Although the score of WNR was still below that of VOC, the latter directly used the prosodic features of original speech to its own advantage. The general impressions within the 4 VC methods were that speech of WNR sounded clear and natural, speech of HEX and GMM sounded muffled, speech of CGMM sounded better than GMM, but it was unnaturally undulating at times, which agreed with its large global variances (Tables III and IV). CGMM and GMM also sounded slightly distorted at times. Based on the paired Student-t test [41], the mean differences in MOS (first row) for WNR vs. HEX, WNR vs. GMM, and WNR vs. CGMM were all statistically significant at the level of $p \leq 0.005$.

2) Similarity Assessment: In this test, a 4-scale score (1 = different and absolutely sure, 2 = different and sure, 3 = same but not sure, 4 = same and absolutely sure) was used to assess the voice similarity concerning an audio sample generated by one of the 4 VC methods and an audio sample of a target speaker. For sanity check, similarity between target speech samples was also assessed. Each listener listened to a sound track of 50 pairs of audio samples. Within a pair, the first audio sample always came from a target speaker, and the second audio sample was from one of the four VC methods for the corresponding target speaker. In each sound track, there were 48 audio sample pairs, covering the combination of 4 VC methods and 12 speaker pairs evenly, and additionally, there were 2 pairs of target sentences. The sentences were randomly taken from the test set, and within each audio sample pair, the two sentence texts were different.

In order to check the effect of vocoder on voice similarity assessment, we partitioned this test into two subtests. In the first subtest, the target speech was the original natural speech in ARCTIC. In the second subtest, the target speech was the ARCTIC speech analyzed and synthesized by the STRAIGHT vocoder. There were 18 listeners in the first subtest, and 13 listeners in the second subtest. The mean and standard deviation of the similarity scores are given in Tables VI and VII for the two subtests.

It is observed that among the 4 VC methods, the WNR method received the highest similarity score, but the score was more than 1 point below the similarity between the target samples. On the other hand, taking the vocoded target speech as the reference increased the scores of WNR and HEX (both VC methods used this vocoder in analysis-synthesis), and decreased the scores of GMM and CGMM (neither of the two VC methods used this vocoder in analysis-synthesis).

Paired Student-t test was used to evaluate the statistical significance of the mean similarity score difference between WNR and each of the three comparison VC methods. In Table VI, where the target was natural speech, the mean score differences of WNR vs. HEX and WNR vs. CGMM were both statistically insignificant at the level of $p = 0.01$, but the difference between WNR and GMM was statistically significant at $p = 0.01$. In Table VII, where the

target was vocoded speech, the mean score difference between WNR and HEX was statistically insignificant at the level of $p = 0.01$, but the difference between WNR vs. GMM as well as WNR vs. CGMM was both statistically significant as the level of $p \leq 0.005$.

3) Preference Assessment on Similarity: In this test, the WNR method was directly compared with each of the 3 other VC methods in term of similarity to target voice. Each listener listened to a sound track consisting of 36 triplets of audio samples. Each triplet had one sample from a target speaker, one converted by WNR, and one converted by one of HEX, GMM, or CGMM. The 36 triplets covered the combinations of the 3 comparison VC methods and 12 speaker pairs evenly. In each triplet, the target sample always came first which was then followed by two voice-converted speech samples, where the order of WNR and the comparison VC method was randomized to avoid order effects. For each triplet, a listener selected the preferred VC method based on better voice similarity to the target, with "equal" for indistinguishability between the two. Like in the similarity assessment test 2), this preference test was partitioned into two subtests, where the first subtest used natural speech of ARCTIC as the target and the second subtest used STRAIGHT vocoded speech as the target. There were 16 listeners in the first subtest and 20 listeners in the second subtest. Listeners' selection counts are summarized in term of percentage of preference in Tables VIII and IX for each pair of VC methods under comparison.

It is seen that the proposed WNR method was preferred more than each of the other three VC methods. When using vocoded target speech as reference, the preference on WNR was largely increased against HEX, and so was the preference on WNR against GMM. The Wilcoxon signed-rank test [41] was performed regarding statistical significances of the differences in preference count for the two subtests (Tables VIII and IX) and the three cases of WNR vs. HEX, WNR vs. GMM, and WNR vs. CGMM. In every case the difference was statistically significant at the level of $p \leq 0.005$.

4) Preference Assessment on Naturalness: In this test, the WNR method was directly compared with each of the other 3 VC methods in naturalness. There were 34 listeners, each listened to a sound track consisting of 36 pairs of audio samples. Each pair had one sample by WNR, and one sample by one of HEX, GMM, or CGMM, where the order of WNR and the comparison VC method was randomized to avoid order effects. The 36 pairs of audio samples covered combinations of the 3 comparison VC methods and the 12 speaker pairs evenly. For each audio sample pair, a listener selected the preferred VC method having better naturalness, with "equal" for indistinguishability. Listeners' selection counts are summarized as percentage of preference for each of the three pairs of VC methods in Table X.

It is seen that the WNR method was more preferred than the other three VC methods by large margins. Again, the Wilcoxon signed-rank test was performed to assess significance in the difference of preference counts for WNR vs. HEX, WNR vs. GMM, and WNR vs. CGMM. In every case, the difference was statistically significant at the level $p \leq 0.005$.

V. Discussion and Conclusion

We have developed a structured sparse spectral transform method for voice conversion that performs frequency warping and spectral shaping simultaneously on high-D STRAIGHT spectra. We have also designed objective measures for assessing VC speech quality in terms of spectral and temporal structural similarity to target and structural contrasts. Our experimental evaluations have demonstrated that the frequency-warping-based ROS facilitates efficient learning of structured sparse transform matrices and provides flexibility in tradeoffs between spectral distortion and speech structure preservation. Our spectral transform approach further has a low computation complexity at the conversion stage, where for each speech utterance it performs direct matrix multiplications instead of iterative NMF as required by the exemplar approach.

The objective measures of covariance and variance have revealed a clear relation between the strictness of ROS constraints and the level of structure preservation. That the subjective MOS rating of the structured sparse transform VC method is “very good” suggests the relevance of local spectral and temporal structures to speech quality. Furthermore, local covariance is suggestive of voice similarity to target, as the structured transform VC method that has high values of local covariance also has high similarity scores. Spectral distortion also affects perceived speech quality and voice similarity, as the examplar method of HEX with lower spectral distortions than CGMM is rated higher in MOS and voice similarity, even though the structure scores of HEX are lower than CGMM. Moreover, the studied objective measures are all spectral-feature-based, while prosodic features that are also important to speech perception have not been utilized. In this regard, the undulating impression of CGMM might have attributed to its lower MOS and voice similarity, even though its covariance and variance values are competitive.

Further progress on VC speech quality and voice similarity calls for improvements not only on spectral conversion, but also on prosody conversion and speech vocoder. As humans would synthesize multiple relevant speech features in subjective evaluations, it is desirable to combine different objective measures in VC speech quality prediction. Similarly, it is important to consider different factors of speech quality and voice similarity in learning models for VC. Our current work suggests the merit of integrating physical structure constraints with mathematical optimization. In a broader view, multiple objective learning that is becoming common in deep learning provides a promising framework for such a task as well.

Acknowledgment

The authors would like to thank Prof. H. Kawahara for his help with the Tandem-Straight toolkit. They would also like to acknowledge Dr. J. H. Benisty and Prof. D. Malah for making their code of CGMM and GMM available in the public domain.

The work of M. Kuruvilla-Dugdale was supported in part by the National Institutes of Health (R15 DC016383). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Heiga Zen.

Appendix

In Fig. 5, the estimated transforms **W** are shown for using the ROC's of WNC, WWR, WSP, and WNR in a male-to-female conversion (rms-to-slt). It is observed that WNC yielded a noisy **W** with low sparsity; WSP yielded a **W** that was constrained by one warping path and with the highest sparsity; the **W**'s from WWR and WNR were between the two extremes and the sparsity given by WNR was higher than that by WWR.

References

[1]. Stylianous Y, Cappe O, and Moulines E, "Continuous probabilistic transform for voice conversion," IEEE Trans. Speech Audio Process., vol. 6, no. 2, pp. 131–142, 3 1998.

[2]. Kain A and Macon MW, "Spectral voice conversion for text-to-speech synthesis," in Proc. Int. Conf. Acoust., Speech, Signal Process, 1998, vol. 1, pp. 285–288.

[3]. Toda T, Black A, and Tokuda K, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectories," IEEE Trans. Audio, Speech, Lang. Process, vol. 25, no. 8, pp. 2222–2235, 11 2007.

[4]. Zen H, Nankaku Y, and Tokuda K, "Continuous stochastic feature mapping based on trajectory HMMs," IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 2, pp. 417–30, 2 2011.

[5]. Abe M, Nakamura S, Shikano K, and Kuwabara H, "Voice conversion through vector quantization," in Proc. Int. Conf. Acoust., Speech, Signal Process, 1988, pp. 655–658.

[6]. Popa V, Silen H, Nurminen J, and Gabbouj M, "Local linear transformation for voice conversion," in Proc. Int. Conf. Acoust., Speech, Signal Process, 2012, pp. 4517–4520.

[7]. Helander E, Virtanen T, Nurminen J, and Gabbouj M, "Voice conversion using partial least square regression," IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 5, pp. 912–921, 7 2010.

[8]. Erro D, Moreno A, and Bonafonte A, "Voice conversion based on weighted frequency warping," IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 5, pp. 922–931, 7 2010.

[9]. Toda T, Saruwatari H, and Shikano K, "Voice conversion algorithm based on Gaussian micture model with dynamic frequency warping of STRAIGHT spectrum," in Proc. Int. Conf. Acoust., Speech, Signal Process, 2001, pp. 841–844.

[10]. Wu Z, Virtanen T, Chng ES, and Li H, "Examplar based sparse representation with residue compensation for voice conversion," IEEE Trans. Audio, Speech, Lang. Process, vol. 22, no. 10, pp. 1506–1521, 10 2014.

[11]. Aihara R, Nakamura T, Takiguchi T, and Ariki Y, "Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary," in Proc. Int. Conf. Acoust., Speech, Signal Process, 2014, pp. 7894–7898.

[12]. Ming H, Huang D, Xie L, Zhang S, Dong M, and Li H, "Examplar-based sparse representation on timbre and prosody for voice conversion," in Proc. Int. Conf. Acoust., Speech, Signal Process, 2016, pp. 5175–5179.

[13]. Wu Z, Tirtanen T, Kinnunen T, Chng ES, and Li H, "Examplar-based unit selection for voice conversion utilizing temporal information," in Proc. Interspeech, 2013, pp. 3057–3061.

[14]. Jin Z, Finkelstein A, Diverdi S, Lu J, and Mysore GJ, "CUTE: A concatenative method for voice conversion using exemplar based unit selection," in Proc. Int. Conf. Acoust., Speech, Signal Process, 2016, pp. 5660–5664.

[15]. Desai S, Raghavendra EV, Tegnanarayana B, Black AW, and Prahallad K, "Voice conversion using artificial neural networks," in Proc. Int. Conf. Acoust., Speech, Signal Process, 2009, pp. 3893–3896.

[16]. Nakashika T, Takashima R, Takiguchi T, and Ariki Y, "Voice conversion in high-order eigen space using deep belief nets," in Proc. Interspeech, 2013, pp. 369–372.

[17]. Xie F-L, Qian Y, Fan Y, Soong FK, and Li H, "Sequence error minimization error training of neural network for voice conversion," in Proc. Interspeech, 2014, pp. 2283–2287.

[18]. Sun L, Li K, Wang H, Kang S, and Meng H, “Phonetic posteriorgrams for many-to-one voice conversion without parallel data training,” in Proc. Int. Conf. Multimedia Expo, 2016, pp. 1–6.

[19]. Valbret H, Moulines E, and Tubach JP, “Voice transformation using PSOLA technique,” in Proc. Int. Conf. Acoust., Speech, Signal Process, 1992, pp. 145–148.

[20]. Toda T et al., “The voice conversion challenge 2016,” in Proc. Interspeech, 2016, pp. 1632–1636.

[21]. Wester M, Wu Z, and Yamagishi J, “Analysis of the voice conversion challeng 2016 evaluation results,” in Proc. Interspeech, 2016, pp. 1637–1641.

[22]. Benisty H and Malah D, “Voice conversion using GMM with enhanced global variance,” in Proc. Interspeech, 2011, pp. 669–672.

[23]. Tian X, Wu Z, Lee SW, and Chng ES, “Correlation-based frequency warping for voice conversion,” in Proc. 9th Int. Symp. Chin. Spoken Lang. Process, 2014, pp. 211–215.

[24]. Tian X, Lee SW, Wu Z, Chng ES, and Li H, “An exemplar-based approach to frequency warping for voice conversion,” IEEE Trans. Audio, Speech, Lang. Process, vol. 25, no. 10, pp. 1863–1876, 10 2017.

[25]. Zou H, Hastie T, and Tibshirani R, “Sparse principal component analysis,” J. Comput. Graph. Statist, vol. 15, no. 2, pp. 265–286, 2006.

[26]. Wang Z, Bovik AC, Sheikh HR, and Simoncelli EP, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process, vol. 13, no. 4, pp. 600–612, 4 2004. [PubMed: 15376593]

[27]. Goldsworthy RL and Greenburg JE, “Analysis of speech-based speech transmission index methods with implications for nonlinear operations,” J. Acoust. Soc. Amer, vol. 116, no. 6, pp. 3679–3689, 2004. [PubMed: 15658718]

[28]. Taal CH, Hendriks RC, Heusdens R, and Jensen J, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” IEEE Trans. Audio, Speech, Lang. Process, vol. 19, no. 7, pp. 2125–2136, 9 2011.

[29]. Villavicencio F, Robel A, and Rodet X, “Applying improved spectral modeling for high quality voice conversion,” in Proc. Int. Conf. Acoust., Speech, Signal Process, 2009, pp. 4285–4288.

[30]. Lee DD and Seung HS, “Algorithm for nonnegative matrix factorization,” in Proc. 13th Int. Conf. Neural Inf. Process. Syst., 4 2001, pp. 556–562.

[31]. Kominek J and Black AW, “CMU ARCTIC database for speech synthesis,” Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-LTI-03-177, 2003.

[32]. Rabiner L and Juang B-H, Fundamental of Speech Recognition. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[33]. Oppenheim AV and Johnson DH, “Discrete representation of signals,” Proc. IEEE, vol. 60, no. 6, pp. 681–691, 6 1972.

[34]. Erro D, Navas E, and Hernaez I, “Parameter voice conversion based on bilinear frequency warping plus amplitude scaling,” IEEE Trans. Audio, Speech, Lang. Process, vol. 21, no. 3, pp. 556–566, 3 2013.

[35]. Shah NJ and Patil HA, “Novel amplitude scaling method for bilinear frequency warping-based voice conversion,” in Proc. Int. Conf. Acoust., Speech, Signal Process, 2017, pp. 5520–5524.

[36]. Kawahara H. Development of exploratory research tools based on TANDEM-STRAIGHT; Proc. Asia Pac. Signal Inf. Process. Assoc. Annu. Summit Conf.; 2009. 111–120.

[37]. Kawahara H. TANDEM STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation; Proc. Int. Conf. Acoust., Speech, Signal Process; 2008. 3933–3935.

[38]. HTK toolkit. 2016 [Online]. Available: http://htk.eng.cam.ac.uk/

[39]. Huang X, Acero A, and Hon H, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Englewood Cliffs, NJ, USA: Prentice-Hall, 2001, pp. 314–315.

[40]. van den Brink G, “Detection of tone pulse of various durations in noise of various bandwidth,” J. Acoust. Soc. Amer, vol. 36, no. 6, pp. 1206–1211, 1964.

[41]. Richard R, Concepts and applications of inferential statistics, 3 2011 [Online]. Available: http://vassarstats.net/textbook/

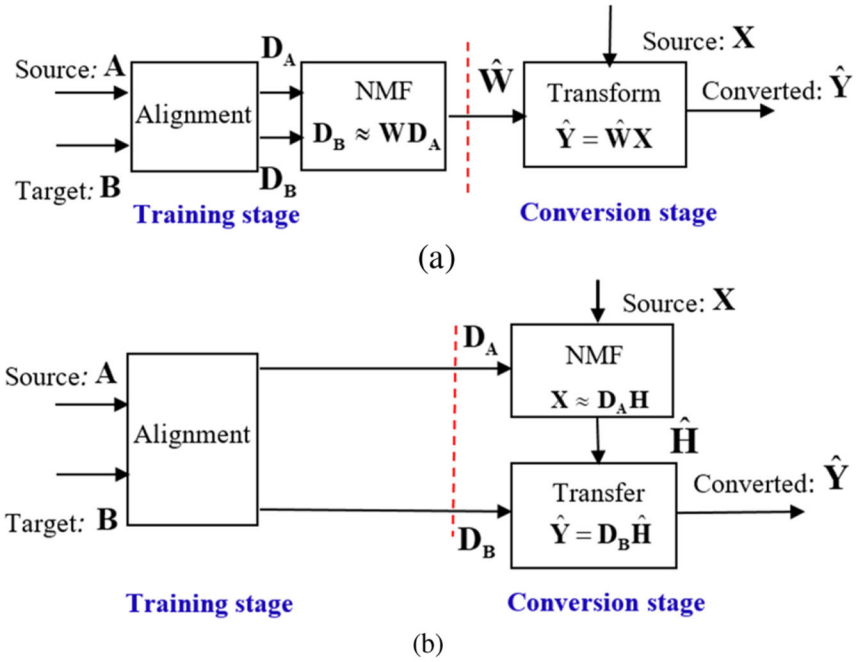

Fig. 1.

Comparison between NMF-based voice conversion approaches using (a) proposed spectral transformation and (b) exemplar activation transfer.

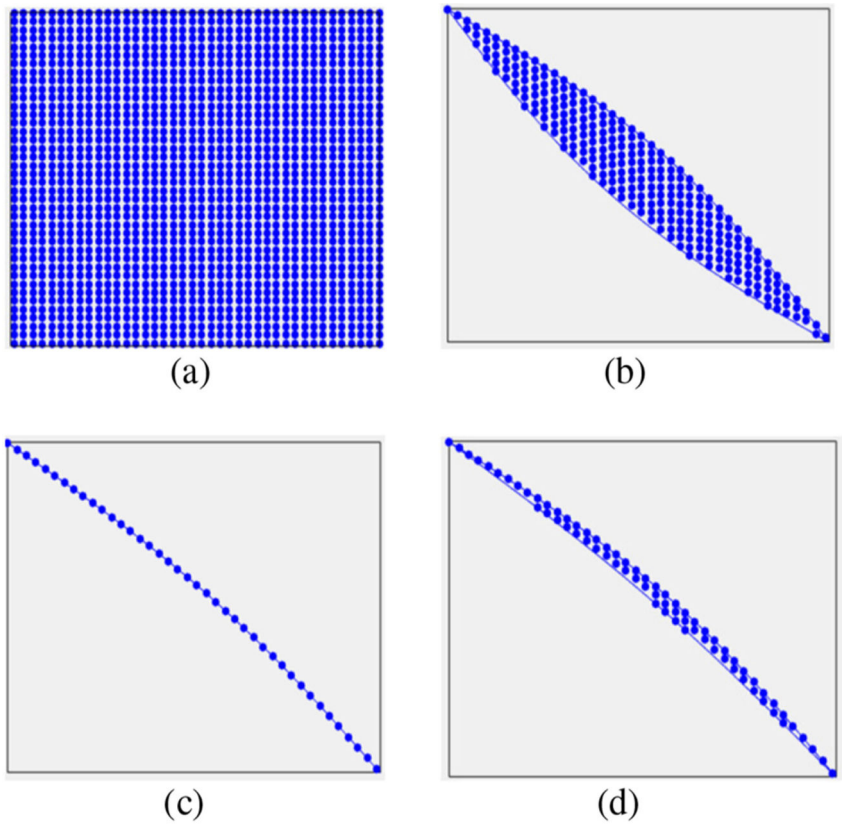

Fig. 2.

Comparison of regions-of-support in transform matrix initialization, where the dotted regions are initialized by ones, and the blank regions are zeros. (a) ROS covers the entire matrix (WNC) (b) ROS is constrained by the largest possible extent of frequency warping (WWR) (c) ROS is constrained by the best frequency warping path of a source-target speaker pair (WSP) (d) ROS is constrained by the best frequency warping path and its neighboring paths of a source-target speaker pair (WNR).

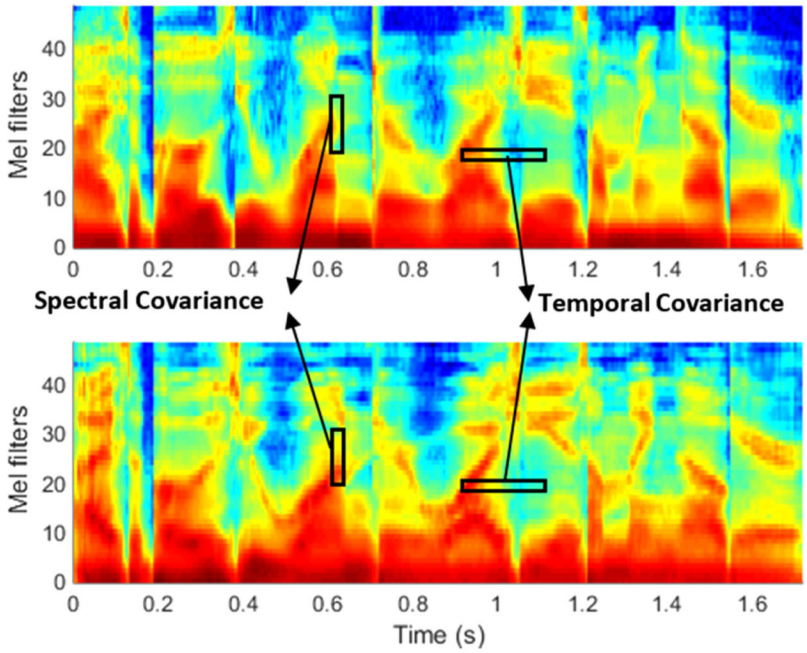

Fig. 3.
Illustration on local spectral and temporal covariance computation.

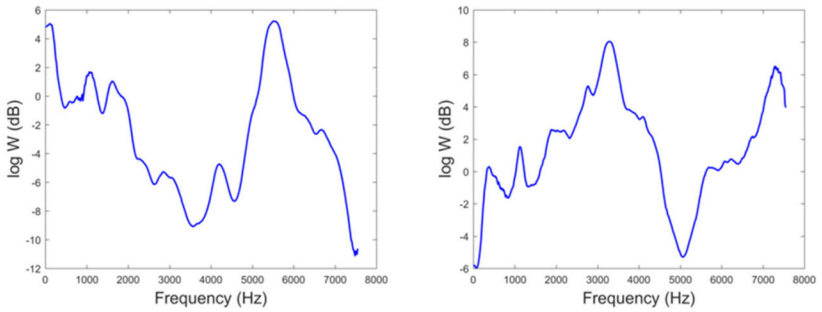

Fig. 4.

Spectral shaping by transformation matrix **W**(WNR). (a) male->female (bdl->slt) and (b) female->male (slt->bdl).

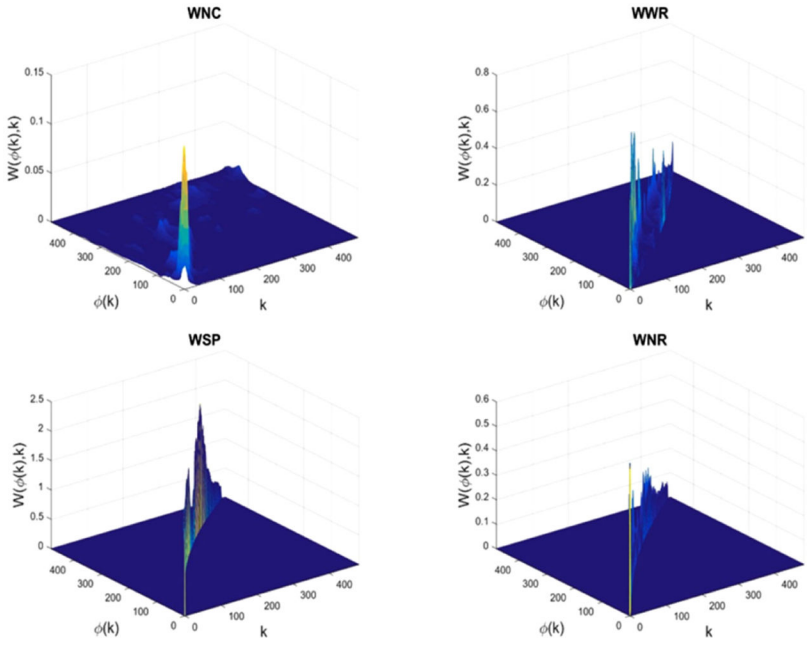

Fig. 5.
3D plots of estimated **W** for one pair of speakers.

TABLE I

Selected Frequnecy Warping Parameters α_m's and α^* (Bold) for WNR

m->m	bdl->rms	rms->bdl		
	1.00,1.04	0.92,**0.96**		
m->f	bdl->slt	rms->slt	bdl->clb	**rms->clb**
	0.84,**0.88**	0.80,**0.84**	0.88,**0.92**	0.84,**0.88**
f->m	clb->rms	slt->rms	clb->bdl	**slt->bdl**
	1.08,**1.12**	**1.12**,1.16	1.04,**1.08**,1.12	**1.08**,1.12
f->f	clb->slt	slt->clb		
	0.96,1.00	**1.00**,1.04		

TABLE II

RMSE of Log Spectra

	HEX	WNC	WWR	WNR	WSP	CGMM	GMM
MEAN	**7.713**	7.815	8.471	8.743	8.882	8.913	8.611
STD	0.713	0.760	0.725	0.716	0.709	0.723	0.730

TABLE III

Spectral Covariance and Variance of Converted and Target Speech

Mel filt.:	**7**		**11**		**15**		**GLOBAL**	
	COV	**VAR**	**COV**	**VAR**	**COV**	**VAR**	**COV**	**VAR**
HEX	0.057	0.054	0.104	0.098	0.144	0.134	0.675	0.635
WNC	0.061	0.060	0.111	0.109	0.154	0.150	0.691	0.668
WWR	0.065	0.094	0.121	0.164	0.169	0.223	0.758	0.847
WNR	0.074	0.133	0.134	0.212	0.185	0.276	0.762	0.875
WSP	**0.075**	**0.148**	**0.135**	**0.231**	**0.187**	**0.298**	0.760	0.880
CGMM	0.069	0.109	0.126	0.187	0.174	0.249	**0.785**	**0.962**
GMM	0.055	0.069	0.103	0.122	0.145	0.168	0.734	0.824
TAR	NA	0.150	NA	0.234	NA	0.300	NA	0.925

TABLE IV

Temporal Covariance and Variance of Converted and Target Speech

Time:	**200 ms**		**280 ms**		**360 ms**		**GLOBAL**	
	COV	**VAR**	**COV**	**VAR**	**COV**	**VAR**	**COV**	**VAR**
HEX	0.172	0.194	0.217	0.239	0.248	0.269	0.389	0.404
WNC	0.173	0.197	0.221	0.246	0.253	0.280	0.402	0.435
WWR	0.198	0.295	0.250	0.362	0.285	0.407	0.444	0.615
WNR	0.202	0.321	0.257	0.396	0.294	0.447	0.460	0.677
WSP	**0.203**	**0.331**	**0.258**	**0.408**	**0.295**	**0.461**	**0.461**	0.699
CGMM	0.197	0.328	0.250	0.405	0.287	0.460	0.450	**0.701**
GMM	0.175	0.260	0.220	0.317	0.250	0.356	0.383	0.531
TAR	NA	0.340	NA	0.424	NA	0.481	NA	0.742

Author Manuscript

Author Manuscript

Author Manuscript

TABLE V

Mean Opinion Score

	NAT	VOC	HEX	WNR	GMM	CGMM
mean	4.47	4.36	2.48	**3.74**	2.28	2.33
stdev	0.65	0.55	0.52	0.53	0.52	0.52
rescaled	5.00	4.88	2.77	4.18	2.55	2.61

TABLE VI

Similarity-to-Target (Natural) Score

	HEX	WNR	GMM	CGMM	TAR
mean	2.45	**2.52**	2.11	2.37	3.76
stdev	0.46	0.40	0.50	0.56	0.52

Author Manuscript

Author Manuscript

Author Manuscript

TABLE VII

Similarity-to-Target (Vocode) Score

	HEX	WNR	GMM	CGMM	TAR
mean	2.46	**2.62**	2.03	2.08	3.88
stdev	0.47	0.42	0.34	0.28	0.30

TABLE VIII

Preference (%) in Similarity-to-Target (Natural)

X vs WNR cases:	X=HEX	X=GMM	X=CGMM
X is preferred	18.28	14.58	11.22
WNR is preferred	**39.25**	**64.06**	**63.78**
X and WNR are equal	42.47	21.36	25.00

TABLE IX

Preference (%) in Similarity-to-Target (Vocode)

X vs WNR cases:	X=HEX	X=GMM	X=CGMM
X is preferred	22.54	13.39	13.50
WNR is preferred	**46.31**	**70.29**	**63.71**
X and WNR are equal	31.15	16.32	22.78

TABLE X

Preference (%) in Naturalness

X vs WNR cases:	X=HEX	X=GMM	X=CGMM
X is preferred	3.69	4.22	2.41
WNR is preferred	**69.21**	**84.66**	**82.17**
X and WNR are equal	27.09	10.92	15.42