Joint nonnegative matrix factorization for exemplar-based voice conversion

Zhizheng Wu[1], *Eng Siong Chng*[1], *Haizhou Li*[1,2]

[1]Nanyang Technological University, Singapore
[2]Institute for Infocomm Research, Singapore

Abstract

Recently, exemplar-based sparse representation methods have been proposed for voice conversion. These methods reconstruct a target spectrum through a weighted linear combination from a set of basis spectra, called exemplars. To include temporal constraint, multiple-frame exemplars are employed when estimating the linear combination weights, namely activations, by the nonnegative matrix factorization technique with a sparsity constraint. In practice, low-resolution mel-scale filter bank energies rather than high-resolution spectra are employed to estimate the activations in order to reduce computational cost and memory usages. However, the conversion performance degrades due to the loss of the spectral details in the low-resolution representations. In this study, we propose a joint nonnegative matrix factorization technique to estimate the activations using both the low- and high-resolution features simultaneously. In this way, we include temporal information by using multiple-frame low-resolution exemplars for computational efficiency and one-frame high-resolution exemplars to improve spectral details at the same time. The VOICES database was employed to assess the performance of the proposed method. The experiments confirmed the effectiveness of the proposed method over conventional nonnegative matrix factorization method in term of both objective spectral distortion and subjective evaluation.

Index Terms: Voice conversion, exemplar, sparse representation, nonnegative matrix factorization, joint nonnegative matrix factorization

1. Introduction

The objective of voice conversion is to change the paralinguistic information such as speaker individuality in the speech signal of one speaker (source) to match that of another speaker (target), while keeping the language content. This technique can be used for personalizing text-to-speech [1], speaking-aid [2, 3], spoofing attack [4] and other applications [5, 6]. One of the most important problems for voice conversion is how to build a robust conversion function. As the spectral attributes which relate to voice timbre contain significant speaker individuality information, the majority of the past work focused on the spectral feature mapping, which is also the focus of this study.

A large number of methods have been proposed to implement flexible spectral mapping functions. These methods implement linear conversion functions as well as nonlinear conversion function. Gaussian mixture model [7] and partial least squares regression [8] based methods are examples to implement linear mapping functions, assuming the source and target features have a linear relationship. Nonlinear conversion methods such as these implemented by neural network [9] and kernel partial least squares regression [10] assume that there is a nonlinear relationship between source and target features. These methods are generally efficient in converting the speaker identity. However, low-resolution spectral features such as Mel-cepstral coefficients (MCCs) and Line spectral frequencies (LSFs) are usually adopted to represent the high-resolution spectra, and the spectral details are lost during the dimensionality reduction process. In addition, these methods attempt to minimize the spectral distance between source and target features on the training data, and this optimization objective will lead to mapping functions that captures the average of the spectra. Thus, these methods usually generate over-smoothed speech that sounds unnatural.

Recently, an alternative nonparametric framework, namely *exemplar-based sparse representation*, have been proposed to model the high-resolution spectra directly for voice conversion [11, 12, 13]. This class of method assumes that a target spectrogram can be generated from a small set of basis target spectra, namely *exemplars*, through a weighted linear combination. Acoustically aligned source-target exemplars from the training data are stored in the coupled source-target dictionaries, and they are assumed to be able to share the same linear combination weights, also called *activations*, to approximate the source-target spectrograms. At run time, the activations for each source spectrogram/utterance are estimated from the source dictionary, and then applied to the target dictionary to generate the corresponding target spectrogram. In this way, the target spectrogram is generated from the real target speech exemplars rather than generated from model parameters. In order to include the temporal contextual constraint, multiple-frame exemplars are used in the source dictionary. If high-resolution features are employed in the source exemplars, the computational cost is considerably high when the window size of an exemplar is large. In the previous work [12], the low-resolution features, namely Mel-scale filter bank energies derived from the high-resolution spectra, were adopted in the source dictionary to reduce computational cost and memory usage. However, the conversion performance drops as the spectral details are lost to some degree.

To address this issue, in this study we propose a joint optimization technique to estimate the activation weights taking both the low-resolution and the high-resolution features into consideration simultaneously. With the low-resolution features, we can include the temporal constraint without significantly increasing the computational cost and the memory usage too much, whereas the spectral details can be taken into account by using the high-resolution spectra. In practice, the nonnegative matrix factorization with a sparsity constraint technique is employed to find the activation weights, and we hence call the proposed method as *joint nonnegative matrix factorization*, which minimizes the joint spectral distance of the low-resolution and high-resolution features simultaneously. Similar to the previous work, the same activation weights are applied to the coupled target dictionary to generate the target spectrogram.

2. Conventional nonnegative matrix factorization for voice conversion

Recently, to model high-resolution spectra for spectral details, exemplar-based sparse representation is proposed for voice conversion in [11, 12]. The basic idea of such exemplar-based methods is to represent a spectrum as a weighted linear combinations of a limited set of basis spectra, and can be formulated as

$$\mathbf{x}^{(\mathrm{DFT})} \approx \sum_{n=1}^{N} \mathbf{a}_n^{(\mathrm{DFT})} \cdot h_n = \mathbf{A}^{(\mathrm{DFT})}\mathbf{h}, \tag{1}$$

where $\mathbf{x}^{(\mathrm{DFT})} \in \mathcal{R}^{p\times 1}$ is the high-resolution spectrum, N is the total number of speech segments, called exemplars, $\mathbf{A}^{(\mathrm{DFT})} = [\mathbf{a}_1^{(\mathrm{DFT})}, \mathbf{a}_2^{(\mathrm{DFT})}, \cdots, \mathbf{a}_N^{(\mathrm{DFT})}] \in \mathcal{R}^{p\times N}$ is the dictionary consisting of exemplars extracted from the source training data, $\mathbf{a}_n^{(\mathrm{DFT})}$ is the n^{th} speech exemplar which has the same dimension as $\mathbf{x}^{(\mathrm{DFT})}$, $\mathbf{h} = [h_1, h_2, \cdots, h_N] \in \mathcal{R}^{N\times 1}$ is the vector consisting of nonnegative weights, also called activation vector and h_n is the activation weight of the n^{th} speech exemplars.

As each frame of a spectrum can be modeled independently using the same dictionary, we can therefore represent the spectrogram of a source utterance as

$$\mathbf{X}^{(\mathrm{DFT})} \approx \mathbf{A}^{(\mathrm{X})}\mathbf{H}, \tag{2}$$

where $\mathbf{X}^{(\mathrm{DFT})} \in \mathcal{R}^{p\times M}$ is the high-resolution source spectrogram, M is the number of frame in the source utterance and $\mathbf{H} \in \mathcal{R}^{N\times M}$ is the activation matrix, the column vector of which is the activation vector as presented in Eq. (1).

Similar to the source spectrogram, the target spectrogram can also be represented by the target dictionary with corresponding activation weights. However, at runtime, the target spectrogram is not available, and the target dictionary and the activations estimated from the source spectrogram are used to find the target spectrogram. It is assumed that the coupled source-target dictionaries with acoustically aligned exemplars can share the same activation weights. In this way, the target spectrogram can be generated by

$$\hat{\mathbf{Y}}^{(\mathrm{DFT})} = \mathbf{B}^{(\mathrm{DFT})}\mathbf{H}, \tag{3}$$

where $\hat{\mathbf{Y}}^{(\mathrm{DFT})} \in \mathcal{R}^{q\times M}$ is the generated target spectrogram, and $\mathbf{B}^{(\mathrm{DFT})} \in \mathcal{R}^{q\times N}$ is the dictionary consisting of the target speech exemplars extracted from the target training data. Note that each column vector of the target dictionary is acoustically aligned with that of the source dictionary from the same entry.

As only the source information is available at runtime, we need to solve Eq. (2) to find the activation weights. In practice, due to the nonnegative nature of the spectrogram, the nonnegative matrix factorization (NMF) technique [14, 15] is adopted to estimate the activation matrix $\mathbf{H}$, which is found by minimizing the following objective function:

$$\mathbf{H} = \arg\min_{\mathbf{H}\geq 0} d(\mathbf{X}^{(\mathrm{DFT})}, \mathbf{A}^{(\mathrm{DFT})}\mathbf{H}) + \lambda\|\mathbf{H}\|_1, \tag{4}$$

where λ is the parameter to control the sparsity of the activation matrix, and $d(\mathbf{X}^{(\mathrm{DFT})}, \mathbf{A}^{(\mathrm{DFT})}\mathbf{H})$ is the spectral distance between the reference and the reconstructed source spectrograms. In practice, the generalized Kullback-Leibler divergence (KLD) [16] is employed to implement the spectral distance, as the KLD is found to be efficient in the noise robust automatic speech recognition research [15].

By minimizing the objective function of Eq. (4), we can derive the following multiplicative updating rule:

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{A}^{(\mathrm{DFT})^\top}\frac{\mathbf{X}^{(\mathrm{DFT})}}{\mathbf{A}^{(\mathrm{DFT})}\mathbf{H}}}{\mathbf{A}^{(\mathrm{DFT})^\top} + \lambda}, \tag{5}$$

where $\otimes$ indicates element-wise multiplication and the divisions are also element-wise. In [15], it is proved that the multiplicative updating rule can be applied iteratively to minimize the objective function as presented in Eq. (4).

In order to include temporal information, in our previous work [12], a frame stacking approach was proposed. In this method, multiple consecutive frames are stacked as a supervector to represent the exemplars in the source dictionary. In order to reduce the computational cost, low-resolution features are employed instead of high-resolution spectra with slightly performance drop. In practice, Mel-scale filter bank energies were employed as the low-resolution features.

The dictionary construction is based on the frame alignment such as dynamic time warping (DTW). More details about the dictionary construction process can be found in [12].

3. Proposed joint nonnegative matrix factorization for voice conversion

The conventional nonnegative matrix factorization technique uses either the low-resolution or the high-resolution features to estimate the activation weights. The low-resolution features are flexible in capturing the temporal contextual information with low computational cost, while the high-resolution spectra contain more spectral details, but the computational cost and memory occupation will increase considerably when contextual information is included.

To benefit from the flexibility of the low-resolution features and spectral details of the high-resolution spectra, we propose a joint nonnegative matrix factorization(Joint-NMF) technique to model both the low- and the high-resolution features simultaneously. The Joint-NMF approach is briefly introduced in this section.

Suppose for each source utterances, the high-resolution spectrogram is represented using the high-resolution source dictionary as shown in Eq. (2). Simultaneously, the low-resolution spectrogram is generated from the low-resolution source dictionary as

$$\mathbf{X}^{(\mathrm{MEL})} \approx \mathbf{A}^{(\mathrm{MEL})}\mathbf{H}, \tag{6}$$

where $\mathbf{X}^{(\mathrm{MEL})}$ is the low-resolution spectrogram corresponding to $\mathbf{X}^{(\mathrm{DFT})}$, $\mathbf{A}^{(\mathrm{MEL})}$ is the low-resolution version of $\mathbf{A}^{(\mathrm{DFT})}$, and $\mathbf{H}$ is exactly the same as that in Eq. (2).

The proposed Joint-NMF with sparsity constraint estimates the activation matrix $\mathbf{H}$ from both Eq. (2) and Eq. (6) simultaneously. The activation matrix is found by minimizing the following objective function:

$$\begin{aligned}\mathbf{H} \quad = \quad & \arg\min_{\mathbf{H}\geq 0} \alpha \cdot d(\mathbf{X}^{(\mathrm{DFT})}, \mathbf{A}^{(\mathrm{DFT})}\mathbf{H}) \\ & + (1-\alpha)\cdot d(\mathbf{X}^{(\mathrm{MEL})}, \mathbf{A}^{(\mathrm{MEL})}\mathbf{H}) \\ & + \lambda\|\mathbf{H}\|_1,\end{aligned} \tag{7}$$

where α is the weighting factor to balance the low-resolution and high-resolution KLD. The objection function in Eq. (7) can be minimized by iteratively applying the following multiplica-

tive updating rule:

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{(1-\alpha)\mathbf{A}^{(\mathrm{MEL})\top}\frac{\mathbf{X}^{(\mathrm{MEL})}}{\mathbf{A}^{(\mathrm{MEL})}\mathbf{H}} + \alpha\mathbf{A}^{(\mathrm{DFT})\top}\frac{\mathbf{X}^{(\mathrm{DFT})}}{\mathbf{A}^{(\mathrm{DFT})}\mathbf{H}}}{(1-\alpha)\mathbf{A}^{(\mathrm{MEL})} + \alpha\mathbf{A}^{(\mathrm{DFT})} + \lambda}. \tag{8}$$

In the updating rule, if $\alpha = 0$, only the low-resolution features are used to estimate the activation weights; if $\alpha = 1$, only the high-resolution features are used; and any values between 0 and 1 are performing trade-off between low-resolution and high-resolution features.

4. Experiments

To examine the performance of the proposed method, we conducted experiments on VOICES database. Speech data from two male speakers and two female speakers was employed, and 10 utterances from each speaker were randomly selected as a training set and 20 utterances without overlapping with those in the training set were used as an evaluation set. We conducted four conversions including inter-gender and intra-gender pairs, and reported the average performance over the four pairs.

To extract features, the speech signals were down-sampled to 16 kHz. STRAIGHT [17] was employed to extract 513-order spectral envelope and fundamental frequency (F0). The spectral envelope was used as the high-resolution feature, while the 23-order Mel-scale filter energies from the spectral envelope were employed as the low-resolution feature. 24-order Mel-cepstral coefficients (MCC) were then extracted from the spectral envelope. Note that MCCs were only used for frame alignment and calculating the spectral distortion.

4.1. Reference methods and setups

In the previous work [12], it had been shown that both NMF-MEL and NMD-DFT achieved better performance in the listening test than the well-established ML-GMM system [7]. We hence use NMF-MEL and NMF-DFT as the reference baselines in this work. The reference methods and the setups are summarized as follows.

- *NMF-DFT*: This is the exemplar-based method using high-resolution spectra extracted by STRAIGHT in the source dictionaries. the conventional nonnegative matrix factorization (NMF) technique is adopted to find the activation weights, as discussed in Section 2.
- *NMF-MEL*: This is the exemplar-based method using the low-resolution Mel-scale filter bank energies in the source dictionaries. Similar to NMF-DFT, the conventional NMF technique is employed to estimate the activation weights.
- *Joint-NMF*: This is the exemplar-based method using the proposed joint nonnegative matrix factorization technique to estimate the activation weights. Both low-resolution and high-resolution features are used to find the activation weights.

In the experiment, spectral envelopes were converted by above conversion methods, while the source F0 was converted by shifting the mean and normalizing the covariance to those of the target.

In practice, for both NMF and joint-NMF algorithms, the updating rules Eqs. (5) and (8) were repeated for 500 iterations, and the activation matrix $\mathbf{H}$ was initialized to unity. The sparsity penalty parameter was set to 0.1 based on the experience of previous work [12].

4.2. Objective evaluations

To evaluate the proposed method objectively, we adopted the Mel-cepstral distortion (MCD), which is computed between the generated and the matched reference target Mel-cepstra, as the objective evaluation measure. In practice, Mel-cepstral coefficients (MCCs) were used for MCD calculation. The MCD value of a paired MCC features is calculated as $\mathrm{MCD[dB]} = \frac{10}{\log 10}\sqrt{2\sum_{i=1}^{24}(c_i - c_i^{\mathrm{conv}})}$, where c_i and c_i^{conv} are the i^{th} coefficients of the matched target and generated MCC features, respectively. As the involved methods did not generated MCCs directly, we computed MCCs from the generated spectrograms first and then calculated the MCD values. In this way, the MCD values are comparable with other methods in the literature to some degree. As the MCD value is based on independent feature pairs, we hence calculated the MCD values frame by frame over all the matched feature pairs in the evaluation set, and reported the mean of the MCD values. Noted that the lower MCD values, the smaller spectral distortions we can expect.

4.2.1. Effect of multiple-frame exemplars

We first examined the effect of using multiple-frame exemplars for both the low- and high-resolution features. A multiple-frame exemplar spans several consecutive frames, and the number of frames is called the window size of an exemplar in this work. Figure 1 presents the spectral distortion results of both the NMF-DFT and the NMF-MEL methods as a function of the window size of an exemplar. For the NMF-DFT method, only the window size lower than 9 was adopted because of the heavy computational cost for a longer window. It is observed that as the window size increases, the spectral distortions decrease from 5.66 dB and 5.77 dB to 5.44 dB and 5.54 dB of the NMF-DFT and the NMF-MEL methods, respectively. When the window size is 7 or 9, the NMF-DFT achieves the lowest distortion, that is 5.44 dB, while the NMF-MEL obtains the lowest distortion, that is 5.54 dB, when the window size is 9. This observations confirm the effectiveness of the multiple-frame exemplars.

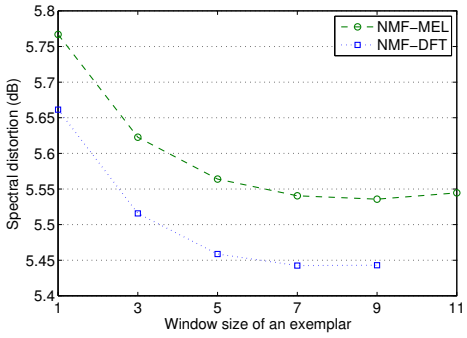

Figure 1: Spectral distortion of NMF-DFT and NMF-MEL as a function of the window size of an exemplar.

When the same window size is adopted, the NMF-DFT always achieves lower spectral distortion than the NMF-MEL. We note that the NMF-DFT method uses high-resolution spectra with spectral details, while the NMF-MEL method does not. It implies that the spectral details are important in estimating more accurate activations. We also note that in the NMF-DFT method, the dimensionality of the spectra is 513, while in the

NMF-MEL method, the feature dimensionality is 23. Thus, with the same window size, the memory usage of the NMF-DFT method is about 22 times higher than the NMF-MEL method. If the window size of the NMF-DFT is 1, and that of the NMF-MEL is 9, the dimensionality of the exemplars in the NMF-MEL method is about 2.5 times lower than the NMF-DFT methods.

4.2.2. *Effect of joint nonnegative matrix factorization*

We then checked the effect of the proposed joint nonnegative matrix factorization (Joint-NMF). In this method, 9-frame exemplars were used for the low-resolution features, while single-frame exemplars were employed for the high-resolution spectra. The reason is for computational efficiency, as multiple-frame high-resolution exemplars require much heavier computation. The spectral distortions as a function the weighting factor α are presented in Figure 2. It is observed that as the weighting factor α increases from 0, the spectral distortion decreases and reaches the minimum, that is 5.43 dB, when α equals to 0.3. After that, the spectral distortion increases along with the value of α. We note that the Joint-NMF method achieves lower spectral distortion than both the NMF-DFT and the NMF-MEL methods with any window size in the exemplars. It confirms the effectiveness of the proposed Joint-NMF over the conventional NMF-DFT and NMF-MEL methods.

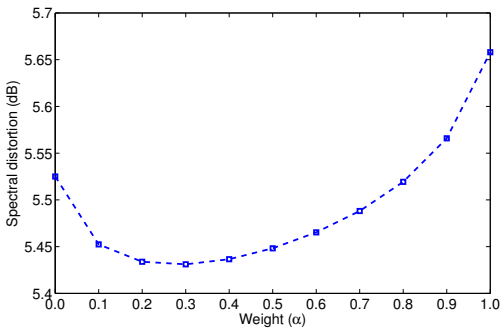

Figure 2: The spectral distortion results as a function of the weighting factor α

4.3. Subjective evaluations

We conducted listening tests to assess the performance in terms of speech quality and speaker individuality of the proposed method in comparison with the NMF-DFT and the NMF-MEL methods. We adopted the Amazon Mechanical Turk (AMT), a kind of crowdsourcing platform, to conduct the listening tests. This platform had been used for speech quality assessment in [18, 19, 20]. In the evaluation set, there were 4 conversion pairs, each of which had 20 utterances, as a result there were 80 (4×20) generated utterances. Note that ten listeners were involved in all the listening tests.

We first conducted a preference listening test for speech quality. In the test, each subject listened to 20 utterances, which were randomly selected from the 80 utterances in the evaluation set. In addition, we randomly mixed three golden standard speech pairs with the 20 testing utterances to exclude cheating (randomly choose preference) as advised in [20].

During the preference listening tests, the converted speech sample by the NMF-DFT/NMF-MEL and the Joint-NMF methods were first presented to the listeners in a random order. Then, the listeners were asked to decide which one sounded more natural.

Figure 3 presents the preference results with 95 % for (a) NMF-DFT vs Joint-NMF and (b) NMF-MEL vs Joint-NMF. It is observed that the proposed Joint-NMF outperforms both the NMF-DFT and the NMF-MEL methods.

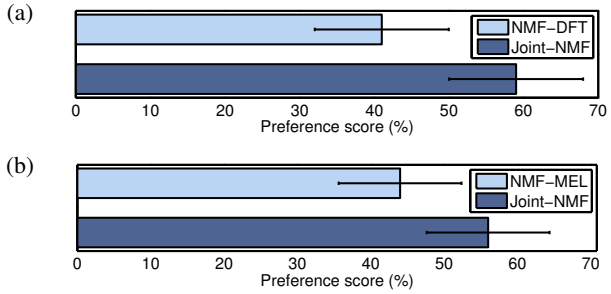

Figure 3: Preference results of speech quality with 95% confidence intervals for the baseline methods and our proposed Joint-NMF method: (a) NMF-DFT vs. Joint-NMF; and (b) NME-MEL vs. Joint-NMF.

We then evaluated the speaker individuality through listening tests. Speaker identification listening tests were conducted. In the test, Each listener first listened to a reference speech, which is the sample produced by one conversion methods, and then listened to samples A and B, corresponding to source and target speech in a random order. To make the listeners focus on the speech quality only, we used the same language content for A and B, while different language content for the reference speech. We performed the listening test for each method independent to avoid bias on speech quality. The identification rates are presented in Figure 4. It is observed that the three methods achieve similar performance in the sense that each method's identification rate is within the 95 % confidence interval of the other methods.

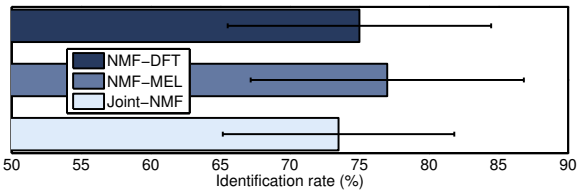

Figure 4: Speaker individuality identification rate with 95 % confidence intervals for the NMF-DFT, NMF-MEL and Joint-NMF methods.

5. Conclusions

In this paper, we proposed a joint nonnegative matrix factorization technique for exemplar-based voice conversion. In this technique, both multiple-frame low-resolution exemplars and single-frame high-resolution exemplars are adopted to estimate the activations. As such, the activations will benefit the temporal constraint from low-resolution features and spectral details from high-resolution features with reduced computational cost and memory usage. In practice, multiple-frame high-resolution features can also be included in this framework, however, the computational cost is very high. We will leave this in the future work.

6. References

[1] 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.

[2] 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.

[3] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[4] Z. Wu and H. Li, "Voice conversion and spoofing attack on speaker verification systems," in *Proc. Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2013.

[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] 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.

[7] 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.

[8] 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.

[9] S. Desai, A. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[10] 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.

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

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

[13] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, 2014.

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

[15] 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.

[16] A. Cichocki, R. Zdunek, and S.-i. Amari, "Csiszars divergences for non-negative matrix factorization: Family of new algorithms," in *Independent Component Analysis and Blind Signal Separation*. Springer, 2006, pp. 32–39.

[17] 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.

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

[19] 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.

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