Interspeech 2018
2-6 September 2018, Hyderabad

Improving Sparse Representations in Exemplar-Based Voice Conversion with a Phoneme-Selective Objective Function

Shaojin Ding, Guanlong Zhao, Christopher Liberatore, Ricardo Gutierrez-Osuna

Department of Computer Science and Engineering, Texas A&M University, USA

{shjd, gzhao, cliberatore, rgutier}@tamu.edu

Abstract

The acoustic quality of exemplar-based voice conversion (VC) degrades whenever the phoneme labels of the selected exemplars do not match the phonetic content of the frame being represented. To address this issue, we propose a Phoneme-Selective Objective Function (PSOF) that promotes a sparse representation of each speech frame with exemplars from a few phoneme classes. Namely, PSOF enforces group sparsity on the representation, where each group corresponds to a phoneme class. The sparse representation for exemplars within a phoneme class tends to activate or suppress simultaneously using the proposed objective function. We conducted two sets of experiments on the ARCTIC corpus to evaluate the proposed method. First, we evaluated the ability of PSOF to reduce phoneme mismatches. Then, we assessed its performance on a VC task and compared it against three baseline methods from previous studies. Results from objective measurements and subjective listening tests show that the proposed method effectively reduces phoneme mismatches and significantly improves VC acoustic quality while retaining the voice identity of the target speaker.

Index Terms: voice conversion, sparse representation, exemplar-based methods

1. Introduction

Voice conversion (VC) seeks to convert an utterance from a source speaker to make it sound as if a target speaker produced it. VC finds applications in many real-world scenarios such as personalized text-to-speech synthesis [1], speaker spoofing [2] and pronunciation training [3]. Several VC frameworks have been proposed; among them, statistical parametric methods based on Gaussian Mixture Models (GMM) [4, 5] are widely used and can achieve convincing performance. Recently, methods based on sparse representations have become another promising approach. This exemplar-based VC framework constructs dictionaries of source and target exemplars selected from a parallel training corpus. At runtime, a source spectrum is represented as a sparse non-negative combination of exemplars in the source dictionary, and the target spectrum is generated by multiplying the sparse representation with the target dictionary.

Exemplar-based VC methods have several advantages: they require much smaller training corpora [6], and they are more robust to noisy speech than GMMs [7]. However, exemplar-based methods lead to phoneme mismatches since the phoneme labels of the selected exemplars may not match the phonetic content of the frame being represented. These phoneme mismatches tend to be speaker-dependent, which reduces the similarity between the source and target sparse representations [8, 9] and introduces distortions in the converted speech. Moreover, this phoneme-mismatch problem becomes more severe as the size of the dictionary increases [8, 9].

In this paper, we address the phoneme-mismatch problem by improving the sparse representation. Namely, we jointly optimize the standard objective function (Mean-Square Error with L_1 constraint) combined with a Phoneme-Selective Objective Function (PSOF) based on the $L_{2,1}$ norm [10] . The $L_{2,1}$ norm enforces group sparsity, and therefore each speech frame tends to be represented using exemplars from a few phoneme classes. Based on PSOF, we propose a modified exemplar-based VC framework (VC-PSOF) that operates as shown in Figure 1 and Figure 2. Namely, during training we construct a phoneme-categorized exemplar dictionary from labeled speech data. Then, at runtime, we compute the sparse representation of the source spectrum by jointly optimizing the standard objective function and the PSOF. Experimental results show that our proposed method effectively reduces the phoneme mismatches and significantly improves VC acoustic quality while capturing the voice identity of the target speaker.

Relation to prior work. Several previous studies have examined the phoneme-mismatch issue. In prior work [11], we used a compact dictionary with a single exemplar per phoneme class. Although mismatches were effectively reduced, the converted speech was lacking in spectral details since no phoneme variations were considered in the exemplar set. Aihara et al. [8] and Berrak Sisman et al. [12] solved this mismatch problem by incorporating phoneme information into the exemplar dictionary. They categorized exemplars into sub-dictionaries according to their phoneme labels, and then selected different sub-dictionaries to represent the speech frames using a sub-dictionary selection procedure [8] or phoneme labels at runtime [12]. The sub-dictionary selection procedure requires extra computations, and acquiring phoneme labels at runtime is impractical. In contrast with their work, we use PSOF to *implicitly* encourage the sparse coding algorithm to represent a source speech frame using exemplars from as few phoneme classes as possible–at no extra computational cost. In practice, the PSOF-selected phoneme classes are similar to the ground-truth phoneme labels.

2. Literature review

The most common approaches for transforming spectral features in VC are based on statistical parametric models. Within these models, GMMs and neural networks have been explored the most. GMM-based methods [4, 5] model the joint distribution of source and target short-time spectra using a GMM. Then, the converted spectral features are estimated by mapping source spectral features to the target space. In neural-network-based methods [13, 14], various architectures such as restricted Boltzmann machines and stacked autoencoders have been used for transforming spectral features directly. Other

10.21437/Interspeech.2018-1272

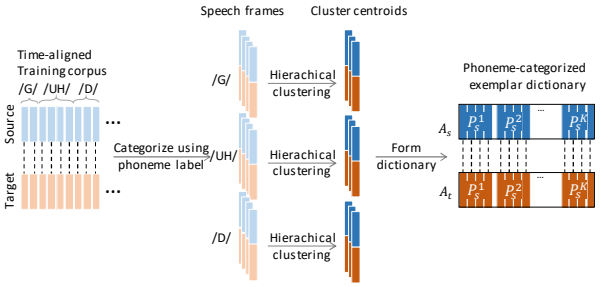

Figure 1: *Training phase of VC-PSOF.*

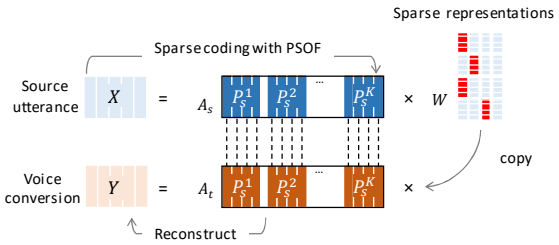

Figure 2: *Testing phase of VC-PSOF.*

statistical models including partial least squares [15] and HMMs [16] have also shown success in VC.

Alternatively, nonparametric exemplar-based methods have been increasingly popular in recent years. Takashima et al. [7] first applied exemplar-based sparse representation to solve the VC problem in noisy environments. Wu et al. [6] improved the original sparse representation by using both high-resolution and low-resolution features to capture spectral details and enforce temporal continuity. Aihara et al. [8, 9] and Berrak Sisman et al. [12] incorporated phoneme information to solve the phoneme-mismatch problem. Liberatore et al. [11, 17] addressed the same issue by constructing compact exemplar dictionaries with a single centroid per phoneme.

3. Conventional framework for exemplar-based VC

In conventional exemplar-based VC frameworks, a source exemplar dictionary $A_s \in \mathbb{R}^{D\times N}$ and a time-aligned target exemplar dictionary $A_t \in \mathbb{R}^{D\times N}$ are first selected from the source and target speakers, where N is the number of exemplars and each exemplar is a D-dimensional spectral feature vector. Then a source utterance $X \in \mathbb{R}^{D\times T}$ with T frames can be represented as:

$$X \cong A_s W \tag{1}$$

where $W \in \mathbb{R}^{N\times T}$ is a sparse non-negative weight matrix (i.e. a sparse representation). Given X and A_s, W can be approximated by minimizing the objective function in eq. (2) through sparse coding,

$$\underset{W}{\operatorname{argmin}}\, d(X, A_s W) + \lambda \|W\|_1, \qquad s.t.\ W \geq 0 \tag{2}$$

where $d(\cdot)$ is a distance metric, typically the KL-divergence or the Euclidean distance, and the L_1 norm term is often included to enforce sparsity in W. To generate a target utterance $\hat{Y} \in \mathbb{R}^{D\times T}$, we multiply the sparse representation of the source utterance with the target exemplar dictionary as:

$$\hat{Y} = A_t W \tag{3}$$

4. Promoting phoneme selectivity in exemplar-based VC

As described before, the standard objective function in eq. (2) can lead to phoneme mismatches [8, 9]. To address this issue, we propose a Phoneme-Selective Objective Function (PSOF) based on the $L_{2,1}$ norm [10].

4.1. Phoneme-selective objective function

We define phoneme selectivity as the property where each spectrum frame is represented with exemplars from as few phoneme classes as possible. In other words, given a source spectrum and a source exemplar dictionary, the estimated weight matrix should be group sparse: for each column, only the weights within a few phoneme classes should be activated. In practice, the most common mathematical tool to enforce group sparsity is the $L_{2,1}$ norm [10]. Therefore, we define a Phoneme-Selective Objective Function (PSOF) $\Psi(W)$ as,

$$\Psi(W) = \sum_{j=1}^{T} \sum_{k=1}^{K} \sqrt{\sum_{i=1, i\in g_k}^{N} w_{ij}^2} \tag{4}$$

where w_{ij} denotes the (i,j)-th element of the weight matrix W, K denotes the number of phoneme groups, g_k represents the k-th phoneme group in the dictionary, and T and N are as defined in Section 3. By minimizing PSOF, we force the weights within a phoneme class to be activated or suppressed at the same time, and therefore we achieve group sparsity in the weight matrix.

To encourage phoneme selectivity in the sparse coding algorithm, we jointly minimize PSOF as,

$$\underset{W}{\operatorname{argmin}}\, d(X, A_s W) + \lambda \|W\|_1 + \beta \Psi(W), \qquad s.t.\ W \geq 0 \tag{5}$$

where β is a penalty term for the proposed PSOF.

Since eq. (5) is convex, sparse coding algorithms such as Non-negative Matrix Factorization (NMF) [18] and Fast Iterative Shrinkage-Thresholding (FISTA) [19] can still be used to optimize it.

4.2. Phoneme-categorized exemplar dictionary

Enforcing group sparsity requires exemplars to be categorized into phoneme groups. To achieve this, we construct phoneme-categorized dictionaries similarly as in [8, 12]. Given K phonemes, source and target dictionaries A_s and A_t are further divided into K sub-dictionaries. For each sub-dictionary, we select a number of speech frames according to their phoneme labels, and then use hierarchical clustering [20] to find M cluster centroids, which then become M exemplars. Formally, A_s and A_t can be expressed as,

$$A_s = [P_s^1, P_s^2, \ldots, P_s^K] \tag{6}$$

$$A_t = [P_t^1, P_t^2, \ldots, P_t^K] \tag{7}$$

where $P_s^i \in \mathbb{R}^{D\times M}$ and $P_t^i \in \mathbb{R}^{D\times M}$ denote the source and target sub-dictionaries of the i-th phoneme, respectively. In practice, the phoneme labels of speech data are acquired from either force alignment or ASR. Figure 1 shows the process of constructing a phoneme-categorized exemplar dictionary.

4.3. Voice conversion based on PSOF

In summary, the workflow of the VC framework based on PSOF is as follows. During training, we construct source and target phoneme-categorized exemplar dictionaries from a labeled and time-aligned parallel speech corpus. During testing,

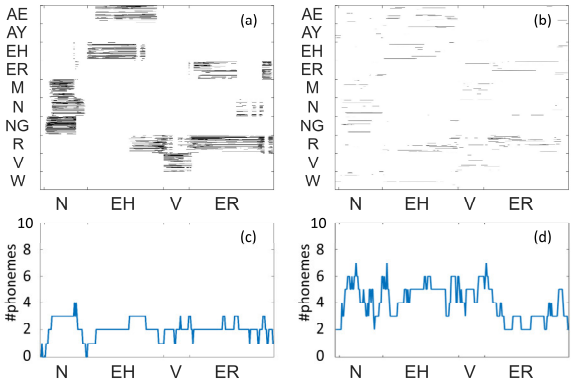

Figure 3: *Visualization of sparse representations ($M = 100$) for the word 'never': (a) VC-PSOF, (b) Baseline. The x-axis values are the transcriptions of the word, and the y-axis values are the phonemes of exemplars in phoneme-categorized sub-dictionary. (c) and (d) are the number of phoneme classes that were used in the sparse representations.*

we compute the sparse representation of a source utterance by optimizing the objective function in eq. (5). Then, we combine the source sparse representation with the target exemplars and generate the converted target utterance. Figure 1 and Figure 2 illustrate the overall process.

5. Experimental setup

To evaluate the proposed VC framework, we conducted experiments on four speakers from the ARCTIC [21] corpus: two male speakers (BDL, RMS) and two female speakers (SLT, CLB). Parallel utterances from these four speakers and the corresponding phonetic transcriptions were used in the experiments. For each speaker, we selected three separate sets: a training set with 20 utterances, a development set with 20 utterances, and a testing set with 50 utterances[1]. Our choice of using a small training set was motivated by applications where collecting a large corpus is impractical (e.g., pronunciation training [22, 23]); these applications are where exemplar-based methods are most beneficial. Four VC pairs were considered for the experiments: BDL to RMS (m-m), RMS to SLT (m-f), SLT to CLB (f-f), and CLB to BDL (f-m). In what follows, all the results are averaged over these four VC pairs.

For each utterance, we used STRAIGHT [24] to extract a 1,025-dimensional spectral envelope, fundamental frequency (F0) and aperiodicity. We compressed the STRAIGHT spectrum using 24 MFCCs (25 Mel-filterbanks, 25 coefficients, removing MFCC_0, which is energy). To construct the phoneme-categorized exemplar dictionary, we assigned each frame of MFCCs a phoneme label based on the ARCTIC transcription. As ARCTIC includes 41 phonemes, we set $K = 41$. No contextual dynamic features were used. Source and target utterances were time-aligned by dynamic time warping [25].

We used the SPAMS sparse coding toolbox [26, 27] to solve for eq. (5). As in prior work [11], we used the Euclidean Distance (i.e., Frobenius norm) $d(X, A_s W) = \|X - A_s W\|_F^2$ as the distance metric. Based on preliminary experiments, we set λ and β to 0.001 and 0.05, respectively. In addition, we normalized the source F0 to match the target space using log-scale mean and variance normalization [5]. Finally, we

[1] Utterances for each set were selected using a maximum entropy criterion to ensure good phonetic balance.

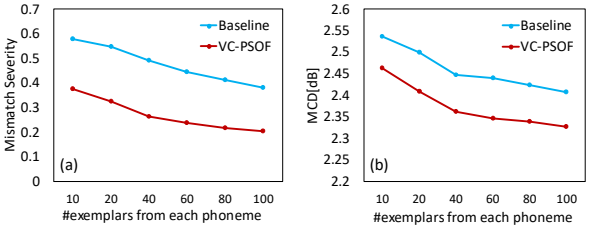

Figure 4: *(a) Mismatching severity ratio. (b) Average MCD of baseline and VC-PSOF.*

estimated the converted spectral envelope using converted MFCCs, and we synthesized the converted speech with transformed spectral envelope, normalized F0 and source aperiodicity.

6. Experiments

We set up two sets of experiments. In the first set, we evaluated the effectiveness of PSOF in reducing phoneme mismatches. In the second set, we compared the VC performance of the proposed method with three baseline methods.

6.1. Effectiveness of phoneme-selective objective function

To evaluate the effectiveness of the PSOF in reducing phoneme mismatches, we compared the VC-PSOF framework against a baseline system. The baseline system was the same as our approach for constructing the exemplar dictionary, but it optimized the objective function in eq. (2). We first visualize the sparse representation of both systems (Figure 3). In a second experiment, we evaluated the severity of phoneme mismatches. Lastly, we examine the Mel-Cepstral Distortion (MCD) [28] to determine if the reduction of phoneme mismatches would improve VC performance. In each experiment, we tested two systems with various numbers of exemplars in dictionaries. For each sub-dictionary, we set the number of exemplars to $M = 10$, 20, 40, 60, 80 and 100.

Visualization. Figure 3 shows the sparse representations of the word "never" from BDL (only showing the sub-dictionaries that were activated. In the baseline system (Figure 3 (b)), a speech frame can be represented by exemplars from arbitrary phoneme labels. In contrast, by jointly optimizing PSOF (Figure 3 (a)), the sparse coding objective is biased to use exemplars from the closest phonemes to represent a speech frame, reducing the number of phoneme mismatches. Additionally, as shown in Figure 3 (c) and Figure 3 (d), VC-PSOF usually represents a speech frame using fewer phoneme classes (~2), while the baseline system tends to represent a frame using ~4-6 phonemes. For example, speech frames for /EH/ are represented by exemplars from four phonemes (/AE/, /EH/, /R/, and /W/) in the baseline system, but only by two phonemes (/EH/ and /AH/) when using PSOF. In cases where PSOF is unable to represent frames using exactly one phoneme class, it tends to choose very similar additional phonemes (e.g., /EH/ and /AH/ are both central vowels).

Mismatch Severity. Given that phoneme mismatches reduce the similarity between source and target weights [8, 9], we use the dissimilarity between the source and target sparse representations of time-aligned parallel utterances to measure

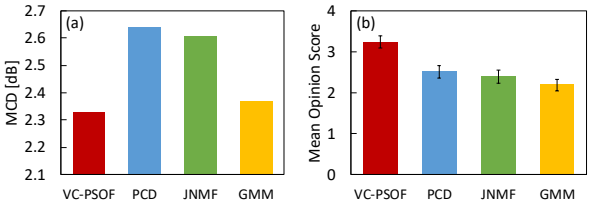

Figure 5: *(a) Average MCD of four systems. (b) Acoustic quality results with 95% confidence intervals.*

the mismatch severity, defined as,

$$mismatch\ severity = \frac{1}{T}\|W_s - W_t\|_F \qquad (8)$$

where $W_s \in R^{N\times T}$ and $W_t \in R^{N\times T}$ are the source and target sparse representations computed from time-aligned parallel utterances, and T is the number of frames. Results are shown in Figure 4 (a). The mismatch severity of VC-PSOF is always lower than that of the baseline, which indicates that VC-PSOF always helps to reduce the phoneme mismatches.

Mel-Cepstral Distortion (MCD). We also measured the MCD of the converted speech and the time-aligned target speech. Results are shown in Figure 4 (b). VC-PSOF always achieves lower MCD than the baseline, regardless of the number of exemplars. The average improvement in MCD of VC-PSOF with respect to the baseline is 0.1. This result indicates that the reduction of phoneme mismatches helps to improve VC performance.

6.2. Performance on voice conversion

We evaluated the VC-PSOF performance on a VC task through both objective and subjective experiments. In these experiments, we set the number of exemplars for each phoneme to 100 (4,100 in total). We also compared our methods against three baseline methods from previous studies: Phoneme-Categorized Dictionary (PCD) [8], Joint Non-negative Matrix Factorization (JNMF) [6], and GMM [4]. PCD shares the notion of phoneme-categorized exemplar dictionary as our proposed work, but at run-time it uses a selection procedure to explicitly select which sub-dictionaries to use and then computes the sparse representation on the selected sub-dictionaries using eq. (2). JNMF is another exemplar-based VC method. During training, it randomly selects exemplars from parallel utterances for use as an exemplar dictionary; at runtime, it computes the sparse representation using eq. (2) on this dictionary. We used the same number of exemplars in PCD and JNMF as VC-PSOF to guarantee a fair comparison. The GMM based method in [4] is one of the most widely used parametric models for VC. We did not use MLPG [5] in our GMM conversion, as it does not converge well under such a small training set. We set the number of mixtures in GMM to 32, as suggested in [5].

6.2.1. *Objective evaluation*

We evaluated the four systems objectively by computing the MCD of the converted speech and the time-aligned target speech. Figure 5 (a) summarizes the results. VC-PSOF outperforms PCD and JNMF significantly, and also achieves marginally better performance than GMM.

6.2.2. *Subjective evaluation*

We conducted listening tests on Amazon Mechanical Turk to evaluate the four systems subjectively. Following previous studies [29, 30], we measured the acoustic quality with a 5-point Mean Opinion Score (MOS) test and the speaker identity with a Voice Similarity Score (VSS) test ranging from -7 (definitely different speakers) to +7 (definitely the same speaker).

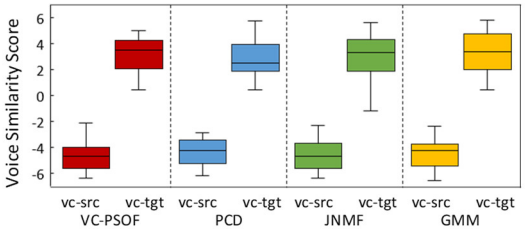

Figure 6: *Speaker identity results (vc-src: VSS between VC and source speaker. vc-tgt: VSS between VC and target speaker).*

Mean Opinion Score. Twenty-five participants rated 80 utterances from four VC systems: 20 utterances per system, 5 utterances per speaker pair. Figure 5 (b) shows the MOS results with 95% confidence intervals. With a limited number of training utterances, the three exemplar-based methods outperform GMM, in agreement with prior studies showing that exemplar-based methods require less training data to achieve reasonable performance [6]. VC-PSOF obtained a 3.23 MOS, which is higher than PCD, JNMF, and GMM with statistical significance ($p \ll 0.001$ in all cases).

Voice Similarity Score. Twenty-two participants rated 160 utterance pairs: 40 pairs (20 VC-source and 20 VC-target pairs) for each system and 10 pairs (5 VC-source and 5 VC-target pairs) for each speaker pair. For each utterance pair, participants were required to decide whether the two utterances were from the same speaker, and then rate their confidence in the decision on a 7-point scale. Following [29], VSS is computed by collapsing the above two fields into a 14-point scale. As shown in Figure 6, participants were "quite confident" that (1) VC-PSOF utterances and the source utterances were from different speakers (VSS: -5.09); and that (2) VC-PSOF utterances and the target utterances were from the same speaker (VSS: 3.08). In addition, we found no statistically significant differences in VSS on the four systems (VC-source VSS, $p \gg 0.05$; VC-target VSS, $p \gg 0.05$). Thus, these results indicate that VC-PSOF improves the acoustic quality of the converted speech significantly (as reflected in the MOS test) without sacrificing the identity of the converted speech.

7. Conclusion and future work

In this paper, we proposed a Phoneme-Selective Objective Function based on the $L_{2,1}$ norm. By jointly optimizing PSOF, we reduced phoneme mismatches in exemplar-based voice conversion and improved the acoustic quality of the voice conversions. We conducted two sets of experiments to validate the PSOF. Objective and subjective test results showed that the proposed method effectively reduced phoneme mismatches and significantly improved VC acoustic quality while capturing the voice identity of the target speaker.

In the current VC-PSOF method, the phoneme-categorized exemplar dictionaries are selected from labeled training data. Future work will focus on using dictionary learning to learn the phoneme-categorized exemplar dictionaries from the training data without using phoneme labels.

8. Acknowledgements

This work was supported by NSF awards 1619212 and 1623750.

9. References

[1] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 1998, pp. 285-288.

[2] Z. Wu and H. Li, "Voice conversion and spoofing attack on speaker verification systems," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, 2013, pp. 1-9.

[3] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech communication,* vol. 51, pp. 920-932, 2009.

[4] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on speech and audio processing,* vol. 6, pp. 131-142, 1998.

[5] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, pp. 2222-2235, 2007.

[6] Z. Wu, E. S. Chng, and H. Li, "Exemplar-based voice conversion using joint nonnegative matrix factorization," *Multimedia Tools and Applications,* vol. 74, pp. 9943-9958, 2015.

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

[8] R. Aihara, T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7894-7898.

[9] R. Aihara, T. Takiguchi, and Y. Ariki, "Parallel Dictionary Learning for Voice Conversion Using Discriminative Graph-Embedded Non-Negative Matrix Factorization," in *INTERSPEECH*, 2016, pp. 292-296.

[10] C. Ding, D. Zhou, X. He, and H. Zha, "R 1-PCA: rotational invariant L 1-norm principal component analysis for robust subspace factorization," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 281-288.

[11] C. Liberatore, S. Aryal, Z. Wang, S. Polsley, and R. Gutierrez-Osuna, "SABR: Sparse, Anchor-Based Representation of the Speech Signal," in *INTERSPEECH*, 2015.

[12] B. Sisman, H. Li, and K. C. Tan, "Sparse Representation of Phonetic Features for Voice Conversion with and without Parallel Data," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017.

[13] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),* vol. 22, pp. 1859-1872, 2014.

[14] S. H. Mohammadi and A. Kain, "A Voice Conversion Mapping Function Based on a Stacked Joint-Autoencoder," in *INTERSPEECH*, 2016, pp. 1647-1651.

[15] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 18, pp. 912-921, 2010.

[16] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996, pp. 389-392.

[17] C. Liberatore, G. Zhao, and R. Gutierrez-Osuna, "Voice Conversion through Residual Warping in a Sparse, Anchor-Based Representation of Speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[18] S. Sra and I. S. Dhillon, "Generalized nonnegative matrix approximations with Bregman divergences," in *Advances in neural information processing systems*, 2006, pp. 283-290.

[19] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences,* vol. 2, pp. 183-202, 2009.

[20] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika,* vol. 32, pp. 241-254, 1967.

[21] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[22] S. Aryal, D. Felps, and R. Gutierrez-Osuna, "Foreign accent conversion through voice morphing," in *INTERSPEECH*, 2013, pp. 3077-3081.

[23] S. Aryal and R. Gutierrez-Osuna, "Can voice conversion be used to reduce non-native accents?," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7879-7883.

[24] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology,* vol. 27, pp. 349-353, 2006.

[25] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *KDD workshop*, 1994, pp. 359-370.

[26] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research,* vol. 11, pp. 19-60, 2010.

[27] R. Jenatton, J. Mairal, G. Obozinski, and F. R. Bach, "Proximal Methods for Sparse Hierarchical Dictionary Learning," in *ICML*, 2010, pp. 487-494.

[28] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, 1993, pp. 125-128.

[29] D. Felps and R. Gutierrez-Osuna, "Developing objective measures of foreign-accent conversion," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 18, pp. 1030-1040, 2010.

[30] G. Zhao and R. Gutierrez-Osuna, "Exemplar selection methods in voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5525-5529.