Interspeech 2018
2-6 September 2018, Hyderabad

Learning Structured Dictionaries for Exemplar-based Voice Conversion

Shaojin Ding, Christopher Liberatore, and Ricardo Gutierrez-Osuna

Department of Computer Science and Engineering, Texas A&M University, United States
{shjd, cliberatore, rgutier}@tamu.edu

Abstract

Incorporating phonetic information has been shown to improve the performance of exemplar-based voice conversion. A standard approach is to build a phonetically structured dictionary, where exemplars are categorized into sub-dictionaries according to their phoneme labels. However, acquiring phoneme labels can be expensive, and the phoneme labels can have inaccuracies. The latter problem becomes more salient when the speakers are non-native speakers. This paper presents an iterative dictionary-learning algorithm that avoids the need for phoneme labels, and instead learns the structured dictionaries in an unsupervised fashion. At each iteration, two steps are alternatively performed: cluster update and dictionary update. In the cluster update step, each training frame is assigned to a cluster whose sub-dictionary represents it with the lowest residual. In the dictionary update step, the sub-dictionary for a cluster is updated using all the speech frames in the cluster. We evaluate the proposed algorithm through objective and subjective experiments on a new corpus of non-native English speech. Compared to previous studies, the proposed algorithm improves the acoustic quality of voice-converted speech while retaining the target speaker's identity.

Index Terms: voice conversion, dictionary learning

1. Introduction

Voice conversion (VC) is a technique that converts utterances from a source speaker to sound as if a target speaker had produced them. VC finds applications in many real-world tasks such as pronunciation training [1], personalized text-to-speech synthesis [2], and speaker spoofing [3]. Different approaches to VC have been proposed, statistical parametric methods based on Gaussian Mixture Models (GMM) [4, 5] and Deep Neural Networks (DNN) [6-9] being widely used. For low-resource settings, however, methods based on sparse representations have been shown to be more effective. In these methods, exemplars from a source speaker and a target speaker are selected from a parallel training corpus. At runtime, a source spectrum is represented as a sparse non-negative combination of the source exemplars, and then the target spectrum is approximated by multiplying the source's sparse weight matrix with the target's exemplars. Exemplar-based methods require much smaller training corpora [10] and are more robust to noisy speech than GMMs [11]. As a result, exemplar-based methods can be very useful in applications where collecting a large corpus is impractical or the acoustic quality of speech is poor (e.g., pronunciation training [12, 13]).

Recent studies [14-17] have shown that the performance of exemplar-based VC can be improved by incorporating phonetic information. A standard approach is to build a phonetically structured dictionary, where exemplars are categorized into sub-dictionaries according to their phoneme labels. Phoneme labels are generally derived from either force alignment (FA) or automatic speech recognition (ASR). FA can produce accurate results but requires an orthographic transcription of the utterance, which can be expensive. When transcriptions are not available, phoneme labels can be obtained via ASR, but the process is error-prone. These problems are compounded in the case of non-native speech due to mispronunciations and disfluencies. In the end, using inaccurate phoneme labels can degrade rather than improve VC performance. Even if the phoneme labels are accurate, they are often too coarse to fully capture detailed phonetic information in speech (e.g., allophones).

To address this problem, we propose an iterative dictionary-learning algorithm with hard-decision rules (HDDL) that avoids the need for phoneme labels. The proposed algorithm is inspired by "hard-decision Expectation Maximization" algorithms [18-22] commonly used for learning models that depend on unobserved latent variables. Figure 1 summarizes the approach. The algorithm consists of two main steps: cluster update and dictionary update. A cluster is defined as a set of speech frames sharing acoustic similarities. In the cluster update step, each training speech frame is assigned to the cluster whose sub-dictionary can best represent the speech frame (i.e., with the lowest residual). In the dictionary update step, the assignment of clusters is fixed, and the sub-dictionary for a cluster is updated using all the speech frames in the cluster. Once the structured dictionaries are learned, we can use any exemplar-based method to perform VC. We conducted both objective and subjective experiments to evaluate the proposed algorithm and compared it against two baseline VC methods. Our results show that HDDL improves the acoustic quality and retains the target speaker's identity on non-native English speech with no extra computations at runtime. In our final analysis, we use the ground-truth phoneme labels and show that the learned structured dictionaries are phonetically meaningful.

2. Literature review

Statistical parametric models, such as GMMs and DNNs, are among the most common VC methods. GMM-based methods [4, 5] learn the joint distribution of source and target short-time spectra, then estimate the target spectral features through least-squares regression. In contrast, DNN-based methods map the source spectral features directly into the target space through various network structures such as restricted Boltzmann machines [6], stacked auto-encoders [7], and variational auto-encoders [8]. Other statistical models such as partial least squares [23] and HMMs [24] have also shown success in VC tasks.

Alternatively, nonparametric exemplar-based methods have become increasingly popular in recent years. Takashima et al. [11] first applied exemplar-based sparse representation to perform VC in noisy environments. Methods have been proposed to improve the sparse representation in both exemplar

10.21437/Interspeech.2018-1295

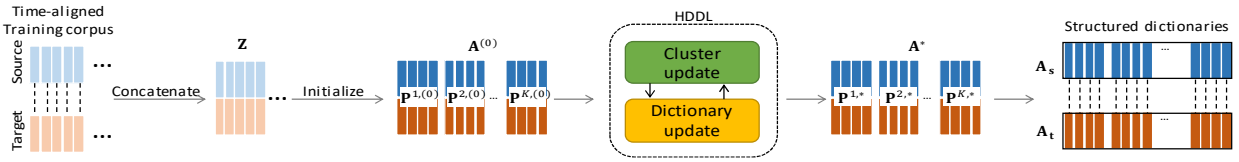

Figure 1: *The workflow of the HDDL.*

selection and matrix factorization methods [10, 25].

Including phonetic information has been shown to improve exemplar-based VC methods [14, 16, 17]. Aihara et al. [14] first used phoneme labels to construct a phonetically structured dictionary. Berrak Sisman et al. [16] used a similar strategy to build the structured dictionary in training, but they also use phoneme labels at runtime. Finally, Liberatore et al. [17] used the centroid for each phoneme as exemplars and constructed a compact dictionary.

Dictionary-learning techniques have been used to learn more effective exemplar dictionaries. Fu et al. [26] proposed a joint dictionary learning algorithm that directly learns exemplar dictionaries from time-aligned utterances. Aihara et al. [15] learned the exemplar dictionaries through a parallel dictionary learning algorithm with a graph-embedded discriminative constraint estimated from phoneme-labeled training data.

Relation to prior work. Our proposed method differs from prior studies in several respects. First, our method learns the dictionaries directly from the data, without using any supervision signals (phoneme labels [14-16], etc.). Second, in contrast with conventional exemplar-based methods [10, 11], HDDL is not restricted to build dictionaries from training frames; instead, the learned dictionaries reflect the distribution of the data, improving the VC performance. Finally, our learned dictionaries are with sub-dictionary structures, which is different from [15, 26].

3. Voice conversion framework

We first describe the conventional exemplar-based VC framework. During training, a source exemplar dictionary $\mathbf{A_s} \in \mathbb{R}^{D\times N}$ and a time-aligned target exemplar dictionary $\mathbf{A_t} \in \mathbb{R}^{D\times N}$ are learned, where N is the number of exemplars, and each exemplar is a D-dim spectral feature vector. At runtime, an L-frame source utterance $\mathbf{X} \in \mathbb{R}^{D\times L}$ is represented as,

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

where $\mathbf{W} \in \mathbb{R}^{N\times L}$ is a sparse non-negative weight matrix (i.e., a sparse representation). Given $\mathbf{X}$ and $\mathbf{A_s}$, $\mathbf{W}$ can be approximated via sparse coding:

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

where $d(\cdot)$ is a distance metric, typically the KL-divergence or the Euclidean distance, $\Psi(\mathbf{W})$ is a regularization term, most commonly based on the L_1 norm to enforce sparsity, and λ is the sparsity penalty. Given $\mathbf{A_t}$ and $\mathbf{W}$, a target utterance $\hat{\mathbf{Y}} \in \mathbb{R}^{D\times L}$ can be generated as:

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

3.1. Incorporating phonetic information

A typical way to incorporate phonetic information in exemplar-based VC is to construct a phonetically structured dictionary [14, 16]. Given K phonemes, $\mathbf{A_s}$ and $\mathbf{A_t}$ are further divided into K sub-dictionaries. For each sub-dictionary, a number of speech frames are first selected according to their phoneme labels, and a clustering algorithm is used to find M cluster centroids, which then become the M exemplars for that phoneme. Formally, $\mathbf{A_s}$ and $\mathbf{A_t}$ can be expressed as follow,

$$\mathbf{A_s} = [\mathbf{P_s^1}, \mathbf{P_s^2}, \dots, \mathbf{P_s^K}] \tag{4}$$

$$\mathbf{A_t} = [\mathbf{P_t^1}, \mathbf{P_t^2}, \dots, \mathbf{P_t^K}] \tag{5}$$

where $\mathbf{P_s^i} \in \mathbb{R}^{D\times M}$ and $\mathbf{P_t^i} \in \mathbb{R}^{D\times M}$ denote the source and the target exemplar sub-dictionaries of the i-th phoneme, respectively, and $i \in \{1, 2, \dots, K\}$. In practice, phoneme labels for the speech frames are derived from force alignment or ASR.

Given the constructed structured dictionary, VC can be performed as described in [14-16]. In this paper, we use a similar approach at runtime as in [14]. Namely, for the l-th frame $\boldsymbol{x}_l$ of source utterance, we compute the sparse weight $\mathbf{w}^i$ with respect to each sub-dictionary,

$$\mathbf{w}^i = \underset{\mathbf{w}}{\operatorname{argmin}} \left\|\boldsymbol{x}_l - \mathbf{P_s^i}\mathbf{w}\right\|_2^2 + \lambda\|\mathbf{w}\|_1, \qquad s.t.\ \mathbf{w} \geq 0 \tag{6}$$

and then select a sub-dictionary with largest weight as,

$$i^* = \underset{i}{\operatorname{argmax}} \left\|\mathbf{w}^i\right\|_2^2 \tag{7}$$

Finally, the corresponding target frame $\mathbf{y}_l$ is generated as,

$$\mathbf{y}_l = \mathbf{P_s^{i^*}}\mathbf{w}^{i^*} \tag{8}$$

3.2. Problems with labeled data

As described in Section 1, building structured dictionaries using phoneme labels produced by FA or ASR can be expensive and degrade the VC performance if the labels are not correct. To avoid the need for phoneme labels, we proposed an iterative dictionary-learning algorithm inspired by "hard-decision Expectation Maximization" algorithms [18-22].

4. Proposed approach: hard decision dictionary learning

Let $\mathbf{X} \in \mathbb{R}^{D\times L}$ and $\mathbf{Y} \in \mathbb{R}^{D\times L}$ be the source and target utterances of a time-aligned parallel training corpus, where D is the number of spectral features and L is the number of speech frames. Following [26], we concatenate the time aligned source and target utterances as $\mathbf{Z} = [\mathbf{X}, \mathbf{Y}]^T$. Our goal is to learn a concatenated dictionary $\mathbf{A} = [\mathbf{A_s}, \mathbf{A_t}]^T$, where $\mathbf{A_s}$ and $\mathbf{A_t}$ consist of sub-dictionaries, as defined in eqs. (4-5). For notation simplicity, we define the concatenated sub-dictionary as $\mathbf{P}^i = [\mathbf{P_s^i}, \mathbf{P_t^i}]^T$. Consequently, we have $\mathbf{A} = [\mathbf{P}^1, \mathbf{P}^2, \dots, \mathbf{P}^K]$. We solve this dictionary-learning problem through an iterative algorithm. At each iteration, we perform two steps: cluster update and dictionary update. Details of each step are provided in following subsections. The overall algorithm is summarized in Table 1.

4.1. Cluster update

Let us denote the concatenated dictionary and sub-dictionary in the t-th iteration as $\mathbf{A}^{(t)}$ and $\mathbf{P}^{i,(t)}$. In the cluster update step, all the $\mathbf{P}^{i,(t)}$ are fixed. For each frame $\mathbf{z}_l$ in $\mathbf{Z}$, we assign $\mathbf{z}_l$ to the cluster whose sub-dictionary $\mathbf{P}^{i,(t)}$ represents $\mathbf{z}_l$ with the

lowest residual computed. Formally, we denote the residual of $\mathbf{z}_l$ respect to $\mathbf{P}^{i,(t)}$ as,

$$r_l^{i,(t)} = \left\| \mathbf{z}_l - \mathbf{P}^{i,(t)} \mathbf{w}_l^{i,(t)} \right\|_2^2 \quad (9)$$

where $\mathbf{w}_l^{i,(t)}$ are the coefficients of the sparse representation. We compute $\mathbf{w}_l^{i,(t)}$ as,

$$\mathbf{w}_l^{i,(t)} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\| \mathbf{z}_l - \mathbf{P}^{i,(t)} \mathbf{w} \right\|_2^2 + \lambda \|\mathbf{w}\|_1 \quad (10)$$

which we solve using the Least Angle Regression (LARS) [27] algorithm. Once the residuals are updated, we can assign $\mathbf{z}_l$ a latent cluster label $p_l^{(t)}$ as,

$$p_l^{(t)} = \underset{i}{\operatorname{argmin}}\, r_l^{i,(t)} \quad (11)$$

Then, we divide $\mathbf{Z}$ into K clusters based on their labels $p_l^{(t)}$ as,

$$\mathbf{Z}^{i,(t)} = \left\{ \mathbb{I}\left(p_l^{(t)} = i\right) \mathbf{z}_l \right\}, \qquad l = 1, 2, \ldots, L \quad (12)$$

where $\mathbf{Z}^{i,(t)}$ denotes all the speech frames in the i-th cluster, and $\mathbb{I}(\cdot)$ is the indicator function.

4.2. Dictionary update

In the dictionary update step, we fix the clusters and update the sub-dictionaries. For all the speech frames in the i-th cluster, we wish to find a sub-dictionary $\mathbf{P}^{i,(t+1)}$ that provides a sparse representation with minimum residual. In other words, we solve the following problem for each sub-dictionary $\mathbf{P}^{i,(t+1)}$:

$$\mathbf{P}^{i,(t+1)} = \underset{\mathbf{P}^i}{\operatorname{argmin}} \left\| \mathbf{Z}^{i,(t)} - \mathbf{P}^i \mathbf{w} \right\|_2^2 + \lambda \|\mathbf{w}\|_1 \quad (13)$$

which we solve using the online dictionary-learning algorithm proposed in [28].

5. Experiments

5.1. Corpus

We used four non-native English speakers from the L2-ARCTIC dataset[1]: ABA (Arabic male), HKK (Korean male), LXC (Chinese female), and SKA (Arabic female). Each speaker produced the full set of ARCTIC [29] prompts. The non-native English speakers in L2-ARCTIC are intermediate-level English learners, so their speech contains disfluencies and mispronunciations. As such, phoneme labels produced by force alignment can be inaccurate, which becomes a problem for conventional exemplar-based VC methods. We also used two native English speakers from ARCTIC: BDL (male), CLB (female). For each speaker, we selected three sets of utterances: 20 utterances for training, 10 utterances for validation, and 50 utterances for testing[2].

5.2. Implementation details

We used STRAIGHT [30] to extract a 1,025-dimensional spectral envelope, fundamental frequency (F0) and aperiodicity for each utterance. We compressed the STRAIGHT spectrum using 24 MFCCs (25 Mel-filterbanks, 25 coefficients, removing MFCC_0, which is energy), and used the MFCCs as the spectral feature. No dynamic contextual features were used. We set the number of clusters to 41, the number of phonemes in ARCTIC.

Table 1: HDDL algorithm

Inputs: concatenated training utterances $\mathbf{Z}$, the number of clusters K

Outputs: learned structured dictionary $\mathbf{A}^* = [\mathbf{P}^{1,*}, \mathbf{P}^{2,*}, \ldots, \mathbf{P}^{K,*}]$

Initialization: randomly assign a latent cluster label to each training frame and divide the training frames to K clusters according to the latent cluster labels, as in eq. (12). Then initialize the dictionary $\mathbf{A}^{(0)} = [\mathbf{P}^{1,(0)}, \mathbf{P}^{2,(0)}, \ldots, \mathbf{P}^{K,(0)}]$ by solving eq. (13).

Repeat until convergence:

- **Cluster update:** compute $\mathbf{w}_l^{i,(t)}$ by solving eq. (10), compute $r_l^{i,(t)}$ as in eq. (9), assign each training frame $\mathbf{z}_l$ a latent cluster label $p_l^{(t)}$ as in (11), and divide the training data into K clusters as in (12).
- **Dictionary update:** update each sub-dictionary $\mathbf{P}^{i,(t+1)}$ in $\mathbf{A}^{(t+1)}$ by solving eq. (13).

Return $\mathbf{A}^* = [\mathbf{P}^{1,*}, \mathbf{P}^{2,*}, \ldots, \mathbf{P}^{K,*}]$

We used 100 basis vectors for each sub-dictionary (i.e., 4,100 in total). Source and target utterances were time-aligned using dynamic time warping [31].

We used the SPAMS sparse coding toolbox [28, 32] to solve for eqs. (10) and (13). Additionally, we normalized the source F0 to match the target space using log-scale mean and variance normalization [5]. We estimated the converted spectral envelope from the converted MFCCs, and finally synthesized the converted speech with the converted spectral envelope, normalized F0 and source aperiodicity.

5.3. Experimental design

We conducted both objective and subjective experiments. In each experiment, we compared the VC performance of the proposed method (HDDL) with two baselines: a conventional GMM approach [4], and the voice conversion method proposed by Aihara et al.: Phoneme-Categorized Dictionaries (PCD) [14]. To evaluate the proposed structured dictionary, we used a similar approach as [14] at runtime except we did not use phoneme labels to build the dictionaries, so it is straightforward to compare our method with PCD. We used similar parameters for PCD (41 sub-dictionaries, 100 basis vectors per sub-dictionary) as our approach to guarantee a fair comparison. The GMM based method in [4] is one of the most widely used parametric VC models. 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]. We consider four VC directions: BDL to ABA, BDL to HKK, CLB to LXC, and CLB to SKA. All the reported results are averaged over these four VC directions.

6. Results

6.1. Objective evaluation

We evaluated the three systems objectively by computing the Mel Cepstral Distortion (MCD) [33] between VC syntheses and the time-aligned target utterances, which served as ground-truth. Results are shown in Figure 2 (a). HDDL outperforms PCD (2.82 vs. 3.05 dB), which shows that the learned structured dictionaries can improve VC performance. Though GMM has lower MCD than HDDL, GMMs suffer from over-smoothing issues [5], and previous studies indicate that GMMs do not achieve better VC performance even if they have lower MCD [15, 34].

[1] The L2-ARCTIC dataset is available at https://psi.engr.tamu.edu/l2-arctic-corpus/.

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

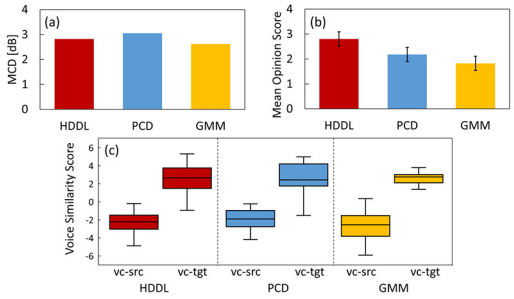

Figure 2: *(a) Average MCD for HDDL, PCD, and GMM. (b) Acoustic quality results with 95% confidence intervals. (c) Voice similarity results (vc-src: VSS between VC and source speaker. vc-tgt: VSS between VC and target speaker).*

6.2. Subjective evaluations

We conducted subjective listening tests on Amazon Mechanical Turk to measure acoustic quality and voice similarity:

- ***Acoustic quality (Mean Opinion Score)***. We recruited 20 participants to rate the 5-point MOS (1: bad, 5: excellent) of 60 utterances from three VC systems: 20 utterances per system, 5 utterances per VC direction.
- ***Voice Similarity Score (VSS)*** [35]. We recruited 20 participants to rate the VSS of 120 utterance pairs from three VC systems: 40 pairs per system, 10 pairs per VC direction. Half of them were VC-source pairs, and the other half were VC-target pairs. Following [1], we played utterances in reverse to reduce the influence of accents in the perception of voice identity. Participants were required to decide if a pair of utterances were produced by the same speaker and rate their confidence on a 7-point scale. VSS results were computed by collapsing the responses into a 14-point scale (-7: definitely different speakers, +7: definitely the same speaker).

MOS results are shown in Figure 2 (b). Participants consistently rated HDDL as having higher MOS than both PCD and GMM, and the results were statistically significant ($p \ll 0.001$ in both cases). VSS results are shown in Figure 2 (c). Listeners are "confident" that the speaker of the converted speech is different from the source speaker (VSS≈ -2.4) and it is the same as the target speaker (VSS$\approx +2.6$). A t-test reveals no statistically significant differences in VSS between HDDL and either PCD or GMM ($p \gg 0.05$ in both cases). The reasons of a low VSS could be two-fold. First, Munro and Derwing [36] have shown that playing utterances in reverse does not entirely eliminate the perception of accent. Second, it is harder for listeners to rate reversed utterances than the original utterances, so they are less confident about their choices.

7. Discussion

Our experiments show that the proposed method can improve the VC acoustic quality over PCD on non-native speech (where phoneme labels may not be reliable) without sacrificing the target speaker's identity. In addition, the proposed method does not require phoneme labels in training nor cause extra computations at runtime. Compared with GMMs, methods based on sparse representations (HDDL and PCD) show an obvious improvement on acoustic quality in this low-resource setting, similar to results in previous studies [10, 11].

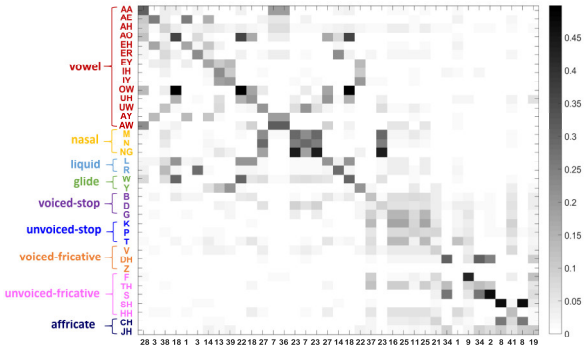

Figure 3: *Confusion matrix between forced-aligned phoneme labels and the matched clusters. Y-axis values are phonemes (sorted by manner of articulation), and X-axis values are the cluster IDs.*

In our final analysis, we explore the relationship between the ground-truth phoneme labels[1] and the learned clusters. In "hard-decision" algorithms, clusters commonly represent latent variables; phonemes can be thought of as latent variables in HDDL. For each speech frame, we computed its residual (eq. (9)) and assigned the frame to the cluster which minimized the residual. Then, using the corresponding forced-aligned phoneme labels for each frame, we computed the cluster distribution of each phoneme (i.e. which clusters were assigned to speech frames with the same phoneme label). For each phoneme, we matched it with the cluster that most frequently represented that phoneme. The confusion matrix of ground-truth phonemes vs. the matched clusters is shown in Figure 3. The dark diagonal elements indicate that each cluster is preferentially associated with a single phoneme label. Confusions occur but are usually restricted to be within the same manner of articulation. For example, the sub-dictionaries with the latent phoneme labels of "7" and "23" are both good at representing nasals. The sub-dictionaries "11", "16", "21", "25", "37" are all used for stops. Both "22" and "39" can represent /EY/, /IH/, /IY/ well, which are all front vowels. These results indicate that the proposed algorithm can learn latent structures of speech (e.g. phonemes) without using any supervision signal.

8. Conclusion

In this paper, we proposed an iterative algorithm to learn phonetically structured dictionaries. In contrast with previous studies, we did not use phoneme labels, thus avoiding the degradation of VC performance caused by inaccurate phoneme labels derived from FA or ASR. We conducted both objective and subjective experiments to evaluate the proposed algorithm, and compared it against two baselines: a conventional GMM approach and a state-of-the-art exemplar-based VC algorithm. Results showed that the proposed method improves the acoustic quality and retains the target speaker's identity.

9. Acknowledgements

This work was supported by NSF awards 1619212 and 1623750. We would like to thank Guanlong Zhao and Jie Zhang for providing the dataset and editing the paper.

[1] Labels for /AX/, /OY/, and /ZH/ are missing as they do not occur in our training data.

10. References

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

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

[3] 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.

[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] 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.

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

[8] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*, 2016, pp. 1-6.

[9] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 18, pp. 954-964, 2010.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] S. B. Cohen and N. A. Smith, "Viterbi training for PCFGs: Hardness results and competitiveness of uniform initialization," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 1502-1511.

[19] R. Samdani, M.-W. Chang, and D. Roth, "Unified expectation maximization," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012, pp. 688-698.

[20] D. J. MacKay, *Information theory, inference and learning algorithms*: Cambridge university press, 2003.

[21] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, pp. 281-297.

[22] Y.-C. Chen, V. M. Patel, J. K. Pillai, R. Chellappa, and P. J. Phillips, "Dictionary learning from ambiguously labeled data," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 353-360.

[23] 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.

[24] 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.

[25] 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.

[26] S.-W. Fu, P.-C. Li, Y.-H. Lai, C.-C. Yang, L.-C. Hsieh, and Y. Tsao, "Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery," *IEEE Transactions on Biomedical Engineering,* vol. 64, pp. 2584-2594, 2017.

[27] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics,* vol. 32, pp. 407-499, 2004.

[28] 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.

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

[30] 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.

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

[32] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 689-696.

[33] 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.

[34] 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.

[35] 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.

[36] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language learning,* vol. 45, pp. 73-97, 1995.