Odyssey 2018 The Speaker and Language Recognition Workshop
26-29 June 2018, Les Sables d'Olonne, France

Average Modeling Approach to Voice Conversion with Non-Parallel Data

Xiaohai Tian[1,2], *Junchao Wang*[3], *Haihua Xu*[4], *Eng Siong Chng*[1,2,4] *and Haizhou Li*[5]

[1]School of Computer Engineering, Nanyang Technological University (NTU),Singapore
[2]Joint NTU-UBC Research Center of Excellence in Active Living for the Elderly, NTU, Singapore
[3]College of Information Science and Engineering, Xinjiang University, China
[4]Temasek Laboratories, NTU, Singapore
[5]Department of Electrical and Computer Engineering, National University of Singapore

Abstract

Voice conversion techniques typically require source-target parallel speech data for model training. Such parallel data may not be available always in practice. This paper presents a non-parallel data approach, that we call average modeling approach. The proposed approach makes use of a multi-speaker average model that maps speaker-independent linguistic features to speaker dependent acoustic features. In particular, we present two practical implementations, 1) to adapt the average model towards target speaker with a small amount of target data, 2) to present speaker identity as an additional input to the average model to generate target speech. As the linguistic feature and the acoustic feature can be extracted from the same utterance, the proposed approach doesn't require parallel data in either average model training or adaptation. We report the experiments on the voice conversion challenge 2018 (VCC2018) database that validate the effectiveness of the proposed method.

Index Terms: Voice conversion, Non-parallel data, Average modeling approach (AMA)

1. Introduction

Voice conversion (VC) aims to modify the one's voice (source) to sound like that of another (target). As the spectral formant contains the representative information of a speaker, the voice conversion studies have focused on spectral features transformation. In [1, 2], Gaussian mixture model (GMM) was employed to build a statistical parametric model to learn a linear mapping from source spectral features to target. Some other statistical parametric methods, such as neural network [3, 4, 5, 6] and kernel partial least squares regression [7] learn a nonlinear feature mapping. To preserve the spectral details, frequency warping approaches were studied, such as weighted frequency warping [8], dynamic frequency warping [9], correlation-based frequency warping [10] and sparse representation based frequency warping [11].

Most of the voice conversion techniques rely on parallel data during training. As such parallel data are not always available, there have been attempts to find source-target paired frames from non-parallel corpus. In [12, 13], the INCA algorithm was proposed to align the non-parallel source and target data iteratively. However, the conversion performance was affected by the inaccurate alignment on non-parallel data [12]. An alternative is to learn a general model from parallel data of other speaker pairs, then to adapt the general model towards a new speaker pair. In [14], the adaptation was achieved by maximum a posterior (MAP) method, while in [15], the eigenvoice technique was employed for voice conversion. In [16], an i-vector and average model adaptation approach was proposed. These methods demonstrated the ability to perform the adaptation with non-parallel data. Unfortunately, they continue to rely on parallel training data from multiple speaker pairs when training the average model. Recently, some parallel data free approaches were developed using adaptive restricted boltzmann machine [17] variational autoencoding [18] and energy-based speaker clustering model [19].

Inspired by [20], we introduce a novel average modeling approach (AMA) to voice conversion without the need of parallel data. Instead of training a full conversion model for each target speaker [20], we adapt a general model towards the target, therefore reduce the required amount of target speech. The proposed AMA is a departure from previous adaptation-based approaches where the average models are trained on parallel data of multiple speaker pairs. It learns the mapping between linguistic features to acoustic features of the same speaker. We consider linguistic features are speaker independent that describe the phonetic content of the utterance, while acoustic features are speaker dependent. Therefore, we are able to learn a linguistic-acoustic mapping from the same utterance. In other words, we do not need any parallel data in either average model training or adaptation.

We consider AMA having three advantages,

- Without the need of parallel training data, we can easily make use of publicly available speech corpora for model training;
- Using model adaptation technique instead of full fledged training, we reduce the required amount of target speech;
- As the source data is not required during the training and adaptation, AMA can be easily applicable for many to one conversion.

2. Phonetic PosteriorGrams based VC

In this section, we discuss the advantage and limitation of the Phonetic PosteriorGrams (PPG) based voice conversion [20] technique.

2.1. Methodology

The PPG based voice conversion [20] models the relationship between the PPG features, that is called the linguistic feature, to the acoustic feature. As the PPG feature is considered to be speaker independent, the source speaker information is not required in the training process. Figure. 1 illustrates the training and conversion process. The PPG can be derived by a general purpose automatic speech recognition system (DNN-HMM

10.21437/Odyssey.2018-32

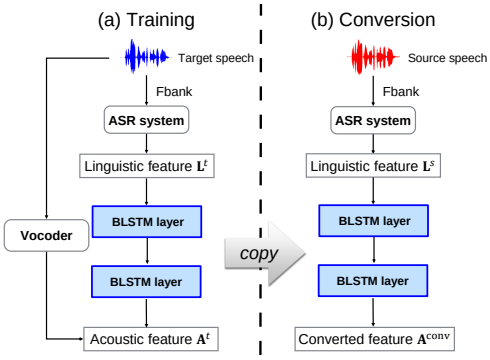

Figure 1: Block diagram of PPG based voice conversion.

ASR).

During offline training, as in Figure. 1(a), two types of features are extracted from target speaker t, the speaker independent linguistic feature, PPG $\mathbf{L}^t \in \mathbb{R}^{D1\times N}$ and the speaker dependent acoustic feature, Mel Cepstral Coefficients (MCC) $\mathbf{A}^t \in \mathbb{R}^{D2\times N}$. We have N to denote the number of frames, and D1 and D2 to denote the dimension of linguistic feature and acoustic feature respectively. As $\mathbf{L}^t$ and $\mathbf{A}^t$ are extracted from the same utterance, these two feature sequence are initially aligned. Then, the feature mapping $\mathcal{F}^t(\cdot)$ of target speaker is trained by the bidirectional long short-term memory (BLSTM) by the back-propagation through time (BPTT) algorithm, denoted as

$$\mathbf{A}^t = \mathcal{F}^t(\mathbf{L}^t) + \mathbf{e}, \tag{1}$$

At run-time, the trained BLSTM model is used for voice conversion as shown in Figure. 1 (b). Given a utterance from the source speaker, we first extract the PPG features $\mathbf{L}^s \in \mathbb{R}^{D2\times M}$ using the same ASR system. As the $\mathbf{L}^s$ is already normalized between speakers, the converted MCCs $\mathbf{A}^{conv}$, are generated by

$$\mathbf{A}^{\text{conv}} = \mathcal{F}^t(\mathbf{L}^s). \tag{2}$$

2.2. Limitation

While the PPG-based voice conversion approach doesn't require parallel data for source-target pair, it still requires a large amount of speech samples from target speaker for conversion model training, that is not practical in many real-world applications. Moreover, for each target speaker, the entire model training must be repeated, which is usually computationally expensive.

3. Average Modeling Approach (AMA) to Voice Conversion

Different from [20], we introduce a speaker independent (SI) average modeling technique as a solution to the non-parallel training data problem. Here, we proposed two average modeling approaches: 1) to adapt the average model with a small mount of target data (model-based average modeling approach), 2) to present speaker identity as the input to the average model to generate target speech (feature-based average modeling approach).

3.1. Model-based AMA

The Model-based AMA consists of three steps: (1) average model training, (2) adaptation and (3) conversion.

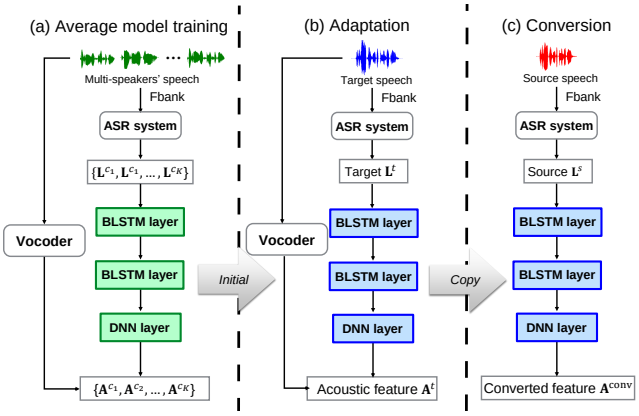

Figure 2: Block diagram of the model-based average modeling approach (AMA) to voice conversion .

Figure. 2(a) shows the average model training process. Given speech data of multiple speakers $\{c_1, ..., c_K\}$, we first extract PPG $\{\mathbf{L}^{c_1}, ..., \mathbf{L}^{c_K}\}$ and MCC features $\{\mathbf{A}^{c_1}, ..., \mathbf{A}^{c_K}\}$, respectively. The paired PPG and MCC features from all the speakers, $\{\{\mathbf{L}^{c_1}, \mathbf{A}^{c_1}\}, ..., \{\mathbf{L}^{c_K}, \mathbf{A}^{c_K}\}\}$, are used to training the SI average model following the same training criterion as Section 2.1.

For a specific target speaker, we adjust the trained SI average model by a small amount of target data (e.g. less than 100 utterances), as shown in Figure. 2(b). The adaptation can be achieved by fine-tuning either a part of the model or the whole network. After adaptation, the feature mapping $\mathcal{F}^t(\cdot)$ of target speaker is obtained.

Similar to the process in Section 2.1, at run-time, we first extract the linguistic feature $\mathbf{L}^s$, from source speech. Then we convert $\mathbf{L}^s$ to acoustic features $\mathbf{A}^{\text{conv}}$ by Eq. (2), as shown in Figure. 2(c).

3.2. Feature-based AMA

As model-based AMA requires model adaptation towards the target speaker, an adaptation step is required for each target speaker. Alternatively, we propose a feature-based AMA that doesn't require adaptation of models. We present the speaker identity (SpeakerID) as part of the input features.

SpeakerID is a low-dimensional vector, which can be i-vector or one-hot vector, representing speaker identity. During the training of average model, we introduce the SpeakerID as a part of the input by augmenting it with the PPG features (see Figure. 3(a)). With the SpeakerID, the acoustic features from different speakers can be distinguished. At run-time conversion, the target speakerID augmented with PPG features are presented to the average model to generate the voice of target speaker, as shown in Figure. 3(b).

4. Experimental Setup

4.1. Database and feature extraction

Both the Wall Street Journal corpus (WSJ) [21] and the VCC2018 [22] corpus[1] were used in the system implementation.

The WSJ corpus consists of 37,318 utterances, including

[1] http://www.vc-challenge.org/

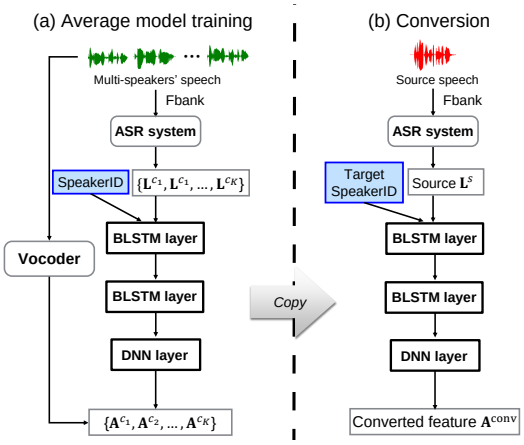

Figure 3: Block diagram of feature-based average modeling approach (AMA) to voice conversion .

18,596 speech samples from 140 female speaker and 18,722 speech samples from 142 male speakers. The VCC2018 corpus consists of 12 speakers, including 8 source speakers and 4 target speakers, with 81 utterances for each speaker.

The ASR system was trained on WSJ corpus; while the speaker independent (SI) average models were trained on the combination of the speech samples of 4 target speakers in VCC2018 corpus and the WSJ corpus. All the audio files were resampled at 16 kHz.

4.1.1. ASR system training

To train the ASR system, 40-dimensional filterbank feature was extracted using a 25 ms hamming window with 5 ms shift. The input layer was composed of filterbank feature and its dynamics with a context window of 11 frames, 120 $\times$ 11 = 1,320 dimensions. The DNN model contained 5 hidden layers with 1,024 hidden units in each layer and the soft-max output layer had 42 units, representing the 42 phoneme. The frame accuracy of the ASR system is 76.03%. Kaldi toolkit [23] was used to train the DNN-HMM ASR system.

4.1.2. Speaker independent (SI) average model training

To ensure the quality of the average models, gender dependent models were trained in our systems. For female average model, we chose 17,828 utterances for training and another 930 non-overlapping utterances for validation. while, for male average model, we chose 17,894 utterances for training and 990 non-overlapping utterances for validation.

WORLD vocoder [24] was used to extract the 513-dimensional spectrum, 1-dimensional aperiodicity coefficients and F_0 with 5 ms frame step. Then 40-dimensional MCCs were calculated from the spectrum using Speech Signal Processing Toolkit (SPTK) [2]. The 42-dimensional phonetic posteriorgram (PPG) features were extracted by the ASR system trained in Section 4.1.1.

- **Model-based AMA:** The average model consisted of two BLSTM layers of 1,024 hidden units, one feed foreword layer of 1,024 hidden units and a linear output layer. The network input was PPG features (42-dim); While the dimension of output was 127, which was the vuv flag (1-dim) combining with MCC (40-dim), lf0 (1-dim) and bap (1-dim) with their dynamic and accelerate features.
- **Feature-based AMA:** The average model architecture and output was the same as that in model-based implementation. The network input was the combination of PPG features (42-dim) with speakerID. In our system, one-hot vector was used to represent the speakerID. Hence, the speakerID were 142 and 144 for female and male average model respectively.

For all the average model training, the minibatch size was set to 10, while the momentum and learning rate were set as 0.9 and 0.002 respectively.

4.2. Baselines and setup

The voice conversion experiments were conducted on the VCC2018 database. There are totally 8 source speakers and 4 target speaker.

To fairly compare with parallel data baseline, 4 source speakers with parallel data were selected consisting of two female speakers, *VCCSF1* and *VCCSF2*, and two male speakers, *VCCSM1* and *VCCSM2*. 4 target speakers were selected consisting of two female speakers, *VCCTF1* and *VCCTF2*, and two male speakers, *VCCTM1* and *VCCTM2*. The training and evaluation data set contains 81 utterances and 35 utterances of each speaker, respectively. For the feature extraction, we used the same configuration as in Section 4.1. The details of baseline and proposed methods were introduced as follows.

- **ML-GMM**: We implemented the joint-density Gaussian mixture model with maximum likelihood parameter conversion. The source and target MCC features were aligned using dynamic time warping (DTW) [25]. 81 parallel utterances from source-target speaker pair were selected for model training. Both static and its dynamic features were used in this implementation. The mixtures number of GMM is set to 64. The global variance enhancement [26] was performed to post filter the converted results for listening test.
- **PPG-LSTM**: The PPG based BLSTM system with non-parallel data [20]. The network contained two BLSTM layers with 256 hidden units of each layer. The network input and output were the same as fine-tune based adaptation described in Section 4.1.2. The minibatch size was set to 10, while the momentum and learning rate were set as 0.9 and 0.002 respectively. 70 and another 11 non-overlapping utterances from target speaker were selected for training and validation respectively.
- **M-AMA**: refers to the proposed model-based AMA. The average model was described in Section 4.1.2. The minibatch size set to 10, the momentum and learning rate of 0.9 and 0.002 respectively. We will examine the effect of adaptation with different settings in Section 5.1.1.
- **F-AMA**: refers to the proposed feature-based AMA. The average model architecture was described in Section 4.1.2.

During conversion, aperiodicity coefficients were copied from source data, while F_0 was converted by a global linear transformation in log-scale. The Maximum Likelihood Parameter Generation (MLPG) algorithm was employed to refine the spectral parameter trajectory [27], followed by spectral enhancement post-filtering in the cepstral domain. The Merlin [28] was used for both average model and conversion model training.

[2] https://sourceforge.net/projects/sp-tk/

5. Evaluations

We compare the proposed approaches with the state-of-the-art through both objective and subjective evaluations. In addition, the results submitted to VCC2018 were also reported.

5.1. Objective evaluation

The Mel-Cepstral Distortion (MCD) [29] was employed as the objective measure. For j^{th} frame, the MCD was calculated as:

$$\text{MCD[dB]} = 10/\ln 10 \sqrt{2 \sum_{i=1}^{I} (mc_{i,j} - mc_{i,j}^{conv})^2}, \quad (3)$$

where $mc_{i,j}$ and $mc_{i,j}^{conv}$ are the i^{th} dimension of target and converted MCCs for frame j, respectively. I is the dimension of MCC feature.

In the experiments, we only report the averaged MCDs of all the conversion pairs. The lower MCD indicates the smaller distortion.

5.1.1. *Effect of adaptation setups on model-based AMA*

Table 1: *Mel-Cepstral Distortions (MCDs) as a function of different adaptation settings for model-based AMA.*

Model-based AMA adaptation			MCD (dB) with different number of adaptaion utterances		
Output layer	DNN layer	BLSTM layers	20 utts	50 utts	70 utts
Y			6.72	6.57	6.55
Y	Y		6.59	6.46	6.42
Y	Y	Y	6.38	6.28	**6.28**

Firstly, we examine the effect of different adaptation setups in model-based AMA. Specifically, we performed the adaptation on different hidden layers and varies the number of adaptation utterances.

The results are presented in Table 1. Across the three adaptation setting, the adaptation over the whole network consistently outperforms another two approaches with lower Mel-Cepstral Distortions (MCDs).

We then compare the performance of these adaptation settings over different amounts of adaptation data, e.g. 20, 50 and 70 utterances for training with 3, 8 and 11 non-overlapping utterances for validation respectively. As shown in Table 1, MCD decreases as the adaptation utterances increases. While the improvement between 50 and 70 adaptation utterances is marginal. The setting of using 70 utterances to adapt the whole network achieves the lowest distortion,with MCD of 6.28 dB. Hence, this adaptation setting is used in the rest of the experiments.

5.1.2. *Comparative studies across competitive methods*

Then, we further compare the performance of AMA approaches with the baseline methods, ML-GMM-GV and PPG-LSTM. The MCD results are demonstrated in Figure 4. We observe that the M-AMA performs similar to PPG-LSTM approach. The averaged MCDs over all the testing pairs of these two approaches are 6.28 dB and 6.27 dB respectively. While the ML-GMM and F-AMA achieves the best and worst results according to in terms of Mel-Cepstral Distortion.

According to the objective evaluation, the proposed method does not outperform the baselines. However, the objective results do not always correlated with the subjective evaluation. Such observations were reported in previous works [8, 11, 30].

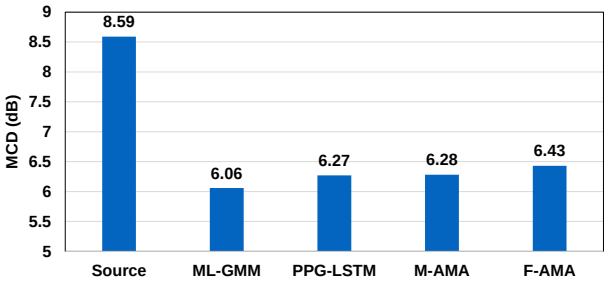

Figure 4: Comparison of Mel-Cepstral Distortions (MCDs) of different conversion methods. 'Source' indicates the MCD evaluated on source utterance without conversion.

In next section, we will examine the subjective evaluation to assess the effectiveness of the proposed AMA.

5.2. Subjective evaluation

AB preference tests and XAB test was conducted to assess the speech quality and speaker similarity respectively. 20 sample pairs were randomly selected from the $16 \times 35 = 560$ paired samples. 10 subjects participated in each tests.

In AB preference tests, each paired samples A and B were randomly select from the proposed method and one of the baseline methods, respectively. Each listener was asked to choose the sample with better quality.

In XAB preference tests, X indicated the reference target sample, A and B were the converted samples randomly selected from the comparison methods. We note that X, A and B have the same language content. The listeners were asked to listen to the samples, then decided A and B which is closer to the reference sample or no preference.

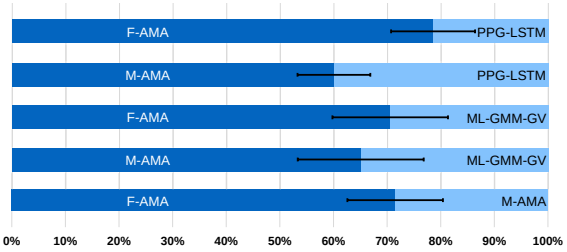

Figure 5: Preference t-test result of speech quality with 95% confidence intervals for proposed AMA and baseline methods.

The subjective results of quality preference tests are presented in Figure 5. The results suggest that the proposed approaches, F-AMA and M-AMA, significantly outperform the baseline systems, ML-GMM-GV and PPG-LSTM. While the F-AMA achieves significantly better performance than M-AMA in terms of quality.

The subjective results of speaker identity are presented in Figure 6. It is observed that the M-AMA obtains the similar score as PPG-LSTM and outperforms ML-GMM-GV. While, the F-AMA consistently outperform the baseline methods and M-AMA.

The subjective results confirm the effectiveness of the proposed average modeling approach (AMA) in terms of both quality and similarity.

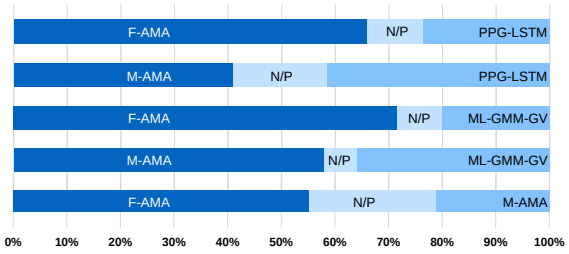

Figure 6: Preference score of speaker identity for proposed AMA and baseline methods. 'N/P' indicates no preference.

5.3. Evaluation results in VCC2018

There are 23 teams participated in VCC2018 [22], including 18 normal submission teams (represented as N03 to N20) and 5 delay submission teams (represented as D01 to D05). We submitted the results of F-AMA system (N04) to the HUB task (parallel data). The standard Mean Opinion Score (MOS) test and similarity test were adopted to evaluation the system performance.

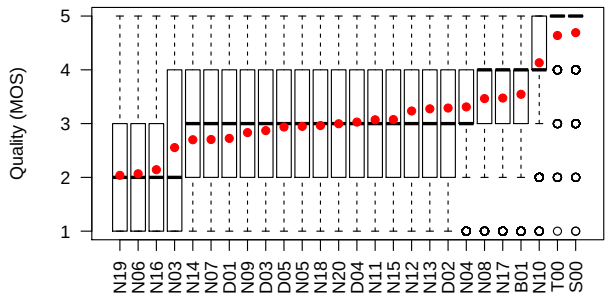

Figure 7: Boxplot for the HUB task according to the MOS score. System N04 denotes the proposed F-AMA.

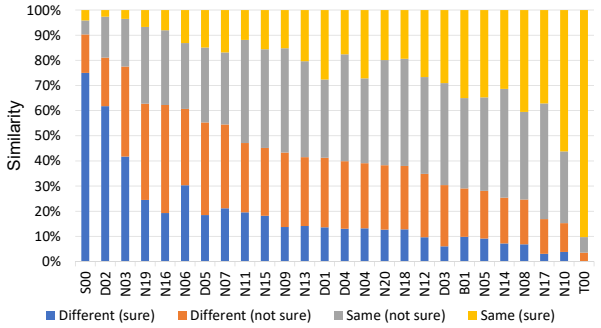

Figure 8: The results for the HUB task according to the Similarity score. System N04 denotes the proposed F-AMA.

Figure 7 and Figure 8 show the MOS and Similarity score plots of all the participating teams for the HUB task[3]. S00 and T00 indicate the source and target. B01 is the baseline system from the SPRocket Voice Conversion Software[4]. The proposed F-AMA is denoted as the the system N04. The results show that our system commands the 4^{th} place in terms of MOS with the score of 3.30 and the 10^{th} place in terms of similarity with the score of 61%.

[3]Figure 7 and Figure 8 are extracted from the official results release

[4]https://github.com/k2kobayashi/sprocket

6. Conclusions

This paper presents an average modeling approach to voice conversion which does not require parallel data in both training and adaptation. We proposed two AMA implementations, namely model-based AMA and feature-based AMA. The AMA method benefits from the use of publicly available speech corpora for average modeling training. Experiments results show that the AMA outperforms to the baseline methods in terms of quality. The results of VCC2018 further confirm the effectiveness of proposed method in both quality and similarity.

7. Acknowledgment

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative, and NUS Start-up Grant, Non-parametric approach to voice morphing.

8. References

[1] Yannis Stylianou, Olivier Cappé, and Eric Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

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

[3] Srinivas Desai, E Veera Raghavendra, B Yegnanarayana, Alan W Black, and Kishore Prahallad, "Voice conversion using artificial neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2009, pp. 3893–3896, IEEE.

[4] Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, and Li-Rong Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE Transactions on Speech and Audio Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[5] Feng-Long Xie, Yao Qian, Yuchen Fan, Frank K Soong, and Haifeng Li, "Sequence error (se) minimization training of neural network for voice conversion," in *INTERSPEECH*, 2014.

[6] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[7] Elina Helander, Hanna Silén, Tuomas Virtanen, and Moncef Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 806–817, 2012.

[8] Daniel Erro, Asunción Moreno, and Antonio Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[9] Elizabeth Godoy, Olivier Rosec, and Thierry Chonavel, "Voice conversion using dynamic frequency warping with

amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[10] Xiaohai Tian, Zhizheng Wu, Siu Wa Lee, and Eng Siong Chng, "Correlation-based frequency warping for voice conversion," in *9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2014, pp. 211–215.

[11] Xiaohai Tian, Zhizheng Wu, Siu Wa Lee, Nguyen Quy Hy, Eng Siong Chng, and Minghui Dong, "Sparse representation for frequency warping based voice conversion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[12] Daniel Erro, Asunción Moreno, and Antonio Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 944–953, 2010.

[13] Hanna Silén, Jani Nurminen, Elina Helander, and Moncef Gabbouj, "Voice conversion for non-parallel datasets using dynamic kernel partial least squares regression," *Convergence*, p. 2, 2013.

[14] Chung-Han Lee and Chung-Hsien Wu, "MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training.," in *Interspeech*, 2006.

[15] Tomoki Toda, Yamato Ohtani, and Kiyohiro Shikano, "Eigenvoice conversion based on gaussian mixture model," in *Interspeech*, 2006.

[16] Jie Wu, Zhizheng Wu, and Lei Xie, "On the use of I-vectors and average voice model for voice conversion without parallel data," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016.

[17] Toru Nakashika, Tetsuya Takiguchi, and Yasuhiro Minami, "Non-parallel training in voice conversion using an adaptive restricted boltzmann machine," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2032–2045, 2016.

[18] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *Interspeech*, 2017.

[19] Toru Nakashika, "Cab: An energy-based speaker clustering model for rapid adaptation in non-parallel voice conversion," *Interspeech*, pp. 3369–3373, 2017.

[20] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016.

[21] Douglas B Paul and Janet M Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992.

[22] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *Submitted to Odyssey*, 2018.

[23] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[24] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, pp. 1877–1884, 2016.

[25] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[26] Tomoki Toda, Takashi Muramatsu, and Hideki Banno, "Implementation of computationally efficient real-time voice conversion.," in *INTERSPEECH*, 2012.

[27] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2000.

[28] Zhizheng Wu, Oliver Watts, and Simon King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW, Sunnyvale, USA*, 2016.

[29] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[30] Xiaohai Tian, Siu Wa Lee, Zhizheng Wu, Eng Siong Chng, and Haizhou Li, "An exemplar-based approach to frequency warping for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1863–1876, 2017.