Many-to-many Cross-lingual Voice Conversion with a Jointly Trained Speaker Embedding Network

Yi Zhou, Xiaohai Tian, Rohan Kumar Das and Haizhou Li
National University of Singapore, Singapore
E-mail: yi.zhou@u.nus.edu, {eletia, rohankd, haizhou.li}@nus.edu.sg

***Abstract*—Among various voice conversion (VC) techniques, average modeling approach has achieved good performance as it benefits from training data of multiple speakers, therefore, reducing the reliance on training data from the target speaker. Many existing average modeling approaches rely on the use of i-vector to represent the speaker identity for model adaptation. As such i-vector is extracted in a separate process, it is not optimized to achieve the best voice conversion quality for the average model. To address this problem, we propose a low dimensional trainable speaker embedding network that augments the primary VC network for joint training. We validate the effectiveness of the proposed idea by performing a many-to-many cross-lingual VC, which is one of the most challenging tasks in VC. We compare the i-vector scheme with the speaker embedding network in the experiments. It is found that the proposed system effectively improves the speech quality and speaker similarity.**

I. Introduction

Voice conversion (VC) aims to modify the speech of one speaker (source) to make it sound like another speaker (target). Based on the availability of parallel training data, VC can be broadly classified into parallel and non-parallel systems [1]. In comparison to parallel VC, non-parallel VC is more practical as it does not require the source and target speakers to record the same set of speech utterances for system training [2]. As a special case in nonparallel system, cross-lingual VC is even more demanding since the speech utterances in different languages are not possible to be parallel, and the phonetic information of the involved languages can be very different [3].

Over the last few decades, a number of techniques have been proposed aiming to improve the converted speech in two aspects: speech quality and speaker similarity. Many conventional approaches like vector quantization (VQ) [4], [5], frame selection [6]–[8], Gaussian mixture modeling (GMM)-based methods [2], [3], [9], [10] and vocal tract length normalization (VTLN) [11] have shown their success in cross-lingual VC. Nevertheless, the converted voice quality is still far from the natural speech [2].

Neural network approaches [12]–[21] are effective as they are powerful to model and generalize the complex spectral mapping function from the input to output features. While, to achieve a good conversion result, they usually require a large number of training data from the target speaker, which is expensive and inconvenient in practice. Average modeling approach is then proposed to leverage the large database from many other publicly available speakers during training [22]–[24]. As an average model learns from multiple speakers, it generates a voice that represents the average voice of all speakers in the training database. Although an average model generates good quality speech signals in general [24], [25], it requires an adaptation step to obtain the converted samples in a target speaker's voice.

The commonly used speaker adaptation techniques can be summarized into three main categories. First, the conversion model can be adapted from the average model with few arbitrary sentences from the target speaker in a different language [26]. Such techniques usually suffer from distortion due to the differences in two language systems. Second, a speaker embedding vector like i-vector [24] could be appended to the input features to condition the speaker identity. Thus the model can be trained to learn the speaker-dependent mapping function. However, the i-vector extraction model is following the speaker verification formulation [27], [28], which is not jointly optimized for VC for optimum voice conversion performance. Last, a trainable speaker embedding network is introduced in [29] for multi-speaker text-to-speech (TTS). The low-dimensional speaker code is trained jointly with input features via backpropagation for model adaptation. Nevertheless, this method is only capable to model speakers seen in the training data.

In our recent work, an average modeling approach has been proposed for cross-lingual VC [30], where i-vector is utilized as the speaker identity representation to achieve many-to-many VC and it is extracted by a separate model designed for speaker verification purpose. Although it works reasonably for cross-lingual VC, the conversion performance is highly dependent on the quality of i-vector. Moreover, we may not be able to obtain the optimal results since i-vector is not optimized together with the VC system.

In this paper, we propose a jointly trained speaker embedding network to encode the speaker information into the primary VC network for system optimization. The speaker embedding is obtained by an auxiliary network from acoustic features, that is called speaker embedding network. The resulting speaker embedding is then repeatedly concatenated to the transformed input linguistic features at each frame. The joint optimization strategy is applied to the speaker embedding network and primary VC network to map the linguistic information conditioned on the speaker embedding to its corresponding acoustic information via backpropagation algorithm. In this way, the speaker embedding network is able to directly model the relevant features for the conversion task.

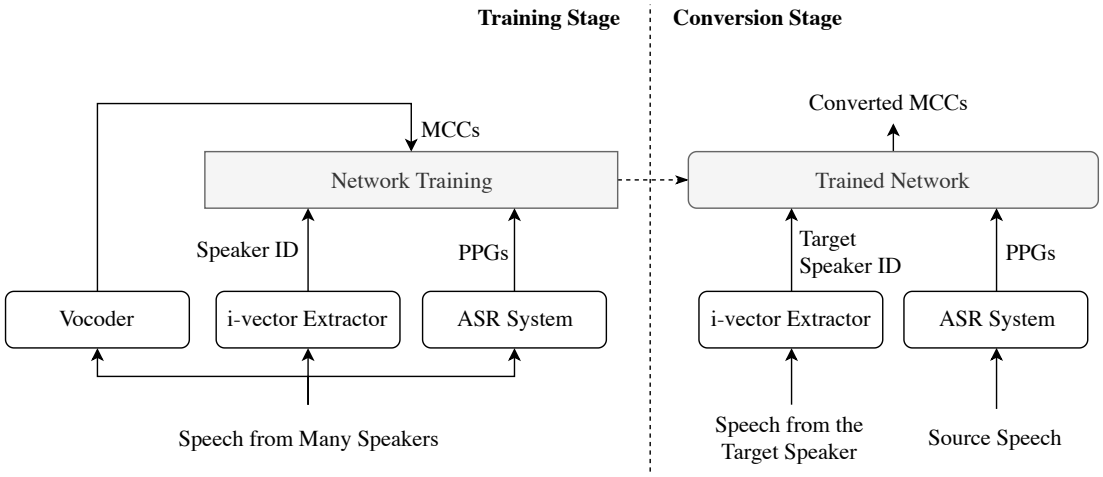

Fig. 1. The training stage and conversion stage of the cross-lingual voice conversion system with the average model conditioned on i-vector.

Furthermore, our proposed method does not rely on the target speaker's data during training, so it is considered flexible and convenient in practice.

II. Average Modeling VC System with i-vector

Fig. 1 shows the Phonetic PosteriorGram (PPG)-based average modeling VC framework. PPG is a time-versus-class vector representing the phonetic classes at frame level [24], [31]–[33], which is derived from an automatic speech recognition (ASR) system as the linguistic features to represent the input speech. The average model learns to map PPG linguistic features to Mel Cepstral Coefficients (MCCs). As it is trained on multiple speakers and different languages, it represents an average voice. The average model is conditioned on i-vector input to project the average voice to a target speaker identity.

A. i-vector Based Speaker Embedding

i-vector is a compact representation for an utterance representing the speaker characteristics [27]. It is derived by a factor analysis approach. In particular, GMM supervectors obtained from some feature representation are projected into a low dimensional space. It is done by creating a total variability space that covers all sorts of variability like speaker, channel and session information. This space is learned by expectation maximization (EM) algorithm using a large amount of background data. The low dimensional representations also carry channel/session information that are required to be compensated by using techniques like linear discriminant analysis (LDA). The final low dimensional representation captures the speaker's identity. This i-vector is used as a speaker embedding for the average model.

B. Training Stage

Speech data from many speakers are first passed into the i-vector extractor, ASR system and vocoder to extract the i-vectors, PPGs and MCCs, respectively. Then, input linguistic features can be formed by concatenating i-vectors with PPGs, and MCCs are used as the output acoustic features. The average model is trained to learn the transformation function from input PPGs with i-vectors to output MCCs by minimizing the mean square error between the original and predicted MCCs via backpropagation.

C. Conversion Stage

Firstly, PPGs are extracted from the source speech using the same ASR system; and the target i-vector is extracted from the target speaker's speech using the same i-vector extractor. Then, PPGs and i-vector are concatenated and fed into the trained network for acoustic feature (MCCs) generation. Finally, the converted MCCs will be used to reconstruct the target waveform as in [30].

D. Limitation

The conversion performance of the i-vector based speaker embedding may not be optimal for the reason that the i-vector extractor is not jointly trained and optimized for VC task.

III. Voice Conversion with a Jointly Trained Speaker Embedding Network

Inspired by the study of speaker auxiliary network in speaker extraction [34], we propose to employ a speaker embedding network for speaker adaptation in the average modeling VC system. Different from the i-vector framework discussed in Section II, our proposed method does not rely on the i-vector extractor to obtain the speaker embedding. Instead, we utilize the acoustic features from the same training speaker to learn a trainable speaker embedding. The primary VC network is based on an average model, that is conditioned on the speaker embedding from the speaker embedding network.

As shown in Fig. 2, the VC framework is similar to the one discussed in Section II, while the proposed VC system does not require the i-vector extractor. Rather than appending i-vector as speaker ID to PPGs as input features, the speaker embedding is jointly trained by presenting MCCs to the network. The schematic diagram of the jointly trained network is shown

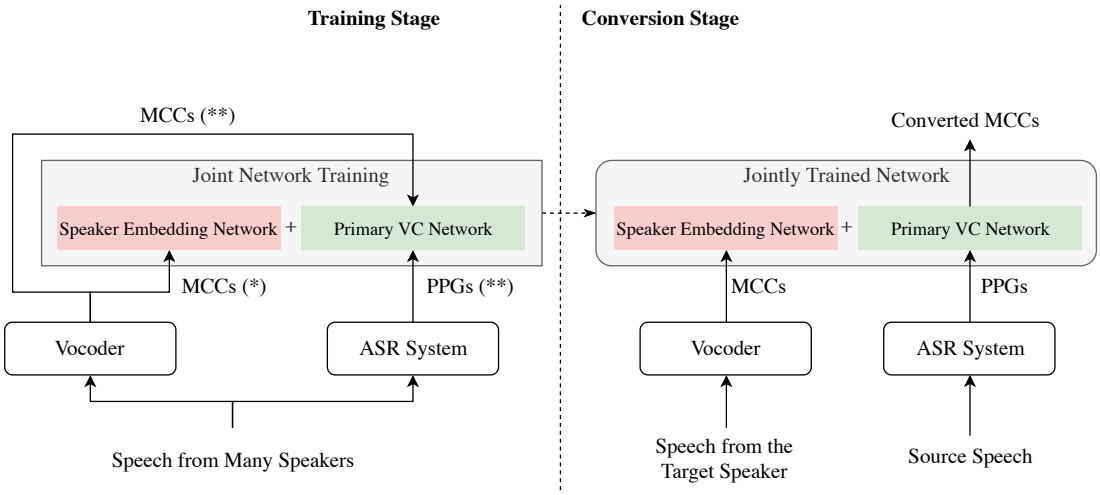

Fig. 2. The block diagram for training stage and conversion stage of the jointly trained many-to-many cross-lingual voice conversion framework using the proposed speaker embedding network. The Joint Network contains two networks: Speaker Embedding Network and Primary VC network. PPGs (**) and MCCs (**) indicate they are extracted from the same utterances. MCCs(*) and MCCs(**) can be obtained from either the same or different utterances, but they are required to be extracted from the same speaker.

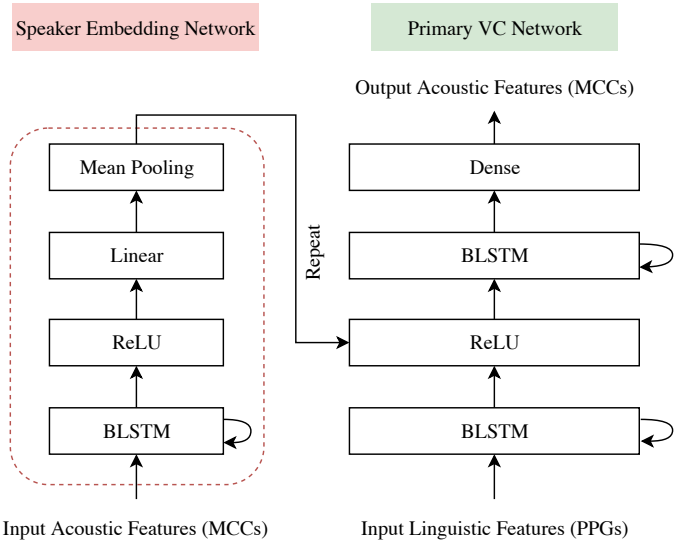

Fig. 3. The schematic diagram of the jointly trained Primary VC Network with the proposed Speaker Embedding Network.

in Fig. 3, and there are two blocks in the proposed network: primary VC network and speaker embedding network.

In the primary VC network, PPGs and MCCs extracted from the same utterances are used as paired input and output features. In the speaker embedding network, we present the MCCs extracted from the training speaker's utterances as input features. In particular, the utterances used for acoustic feature (MCCs) extraction can be different from those used in the primary VC network. By doing so, the speaker embedding network can generally learn the speaker embedding from any given speech data from a target speaker. Thus it is effective to alleviate the mismatch problems caused by language difference between source and target speech during testing time in cross-lingual conversions. The speaker embedding vector is then repeated and concatenated to the transformed linguistic features in the hidden layer of the primary VC network to all frames.

During conversion, we extract PPGs as input features from the speech of a source speaker. While, we extract MCCs from a target speaker's speech to be used as speaker embedding network input features. Both PPGs and MCCs will be passed to their corresponding network, and the jointly trained model will generate the converted MCCs in the target speaker's voice.

IV. Experiments

A. Database and Feature Extraction

In our experiment, VC was performed between English and Mandarin speakers. All selected speech data is native and monolingual, and the details are shown in TABLE I. For training, 64 speakers were used including 32 female and 32 male speakers with 150 utterances from each speaker. The other 12 utterances from each speaker in the training data were used for validation. In total, we used 9,600 utterances to train the average models. For testing, 20 non-overlapped utterances from each of 8 target speakers were chosen.

For the i-vector based speaker embedding, the universal background model (UBM) contained 502 speakers including 251 male and 251 female speakers from Switchboard II corpus. In total, there were 1,872 utterances, and each utterance was about 5 minutes. For the proposed speaker embedding network, the network used the same training data as discussed above, i.e., 9,600 utterances from 64 speakers, and each utterance was few seconds.

The Kaldi toolkit [39] was used for ASR system training. WORLD vocoder [40] was used for MCC feature extraction. The network configurations and other feature details were all the same as in [30].

TABLE I
DETAILS OF DATA USED IN THE EXPERIMENTS.

Stage	Database	Data	Selected Speaker
Training	VCTK [35]	English 1,296 utterances 8 male, 8 female	294, 297, 299, 300, 301, 303, 305, 306 302, 311, 315, 316, 334, 345, 360, 363
	CMU ARCTIC [36]	English 972 utterances 3 male, 3 female	slt, clb, lnh bdl, jmk, rms
	VCC2016 [37]	English 810 utterances 5 male, 5 female	SF1, SF2, SF3, TF1, TF2 SM1, SM2, TM1, TM2, TM3
	Mandarin Library	Mandarin 2,592 utterances 16 male, 16 female	01F, 02F, 03F, 04F, 05F, 06F, 10F, 11F 17F, 18F, 19F, 20F, 22F, 23F, 25F, 26F 07F, 08F, 09F, 12F, 13F, 35F, 39F, 42F 46F, 47F, 48F, 52F, 56F, 58F, 59F, 61F
Testing	VCC2018 [38]	English 80 utterances 2 male, 2 female	TF1, TF2 TM1, TM2
	Mandarin Library	Mandarin 80 utterances 2 male, 2 female	14F, 15F 16M, 24M

We used mel frequency cepstral coefficient (MFCC) features to derive the i-vectors. We used gender-independent UBM of 1024 mixture components and total variability matrix with 400 speaker factors. The 400-dimensional i-vectors obtained from the framework were again reduced to 150 dimensions by applying LDA.

B. Experimental Setups

We compared two systems, namely i-vector scheme, and speaker embedding network, in cross-lingual voice conversion experiments.

- **iSE:** We implemented a cross-lingual VC system with the **i**-vector based **S**peaker **E**mbedding (iSE) as the baseline. As described in Section II, the model was trained with Merlin toolkit [41]. The input feature dimension was 491 including 341-dimensional bilingual PPG and 150-dimensional i-vector. Two BLSTM layers were used and each layer had 512 nodes. The minibatch size, momentum and learning rate were set to 20, 0.9 and 0.002, respectively. The output acoustic feature dimension was 127, which consisted of MCCs (40-dim), log fundamental frequency ($F0$) (1-dim), Aperiodicity (AP) (1-dim) and their dynamic features, and voiced/unvoiced flag (1-dim).
- **SEN:** We employed the proposed jointly trained **S**peaker **E**mbedding **N**etwork (SEN) as discussed in Section III. For the primary VC network, the input feature dimension was 341, which only contained the bilingual PPG. Two BLSTM layers were used and each layer had 512 hidden units. For the speaker embedding network, the input acoustic feature dimension was 127, which was same with that of output acoustic features. Using the same minibatch size of 20, another 20 utterances from from the same speaker were fed into the speaker embedding network. Two BLSTM layers were used and each layer had 256 nodes. The feed-forward hidden layer with ReLU activation function also had 256 nodes. Last, a linear layer with 30 nodes was used with a mean pooling over all frames to produce a 30-dimensional speaker embedding vector. Other parameters and features were the same as our baseline iSE system.

TABLE II
MCD RESULTS FOR INTRALINGUAL VOICE CONVERSION. ISE DENOTES THE BASELINE VC SYSTEM WITH I-VECTOR, AND SEN INDICATES THE PROPOSED JOINTLY TRAINED SPEAKER EMBEDDING NETWORK. M AND F DENOTE THE FEMALE SPEAKER AND MALE SPEAKER RESPECTIVELY. THE ARROW SHOWS THE CONVERSION DIRECTION.

Language	Gender	iSE	SEN
English	M $\rightarrow$ M	5.88	**5.76**
	F $\rightarrow$ F	6.53	**6.29**
	F $\rightarrow$ M	6.54	**6.48**
	M $\rightarrow$ F	6.87	**6.69**
Mandarin	M $\rightarrow$ M	5.91	**5.77**
	F $\rightarrow$ F	6.81	**6.68**
	F $\rightarrow$ M	6.73	**6.60**
	M $\rightarrow$ F	7.47	**7.21**

During conversion, we directly copied APs from source speech, while converted $F0$ by a global linear transformation in log-scale [32], [42], [43]. The MCCs were obtained by maximum likelihood parameter generation algorithm [44]. A post-filtering technique was also employed [45].

C. Evaluations

Both objective and subjective evaluations were conducted on the baseline and proposed systems. We covered all intra-gender and inter-gender conversions among the test speakers in two languages, and the average results will be reported in each language. As we only have monolingual speech data from all chosen speakers, the objective evaluation results will be discussed only for intralingual VC. However, subjective evaluations will focus on cross-lingual VC. The converted speech samples are available from the demo link[1].

1) Objective Evaluation: Mel-cepstral distortion (MCD) was used to measure the spectral distance between the ground truth and converted speech, which is defined as follows between two MCC frames,

$$MCD[dB] = \frac{10}{log10}\sqrt{2\sum_{d=1}^{D}(c_d - c_d^{converted})^2} \quad (1)$$

[1]https://vcsamples.github.io/APSIPA2019/

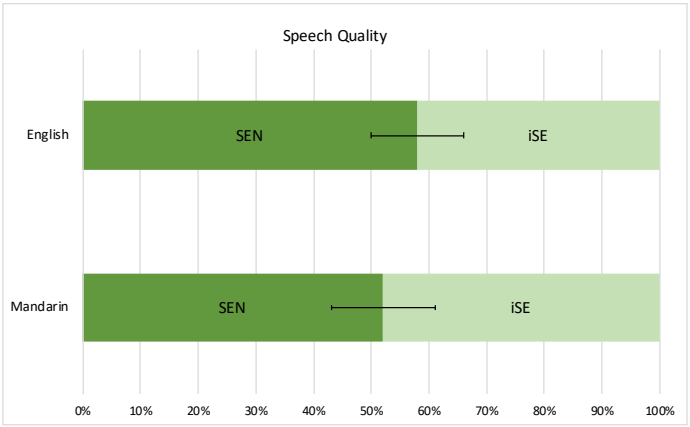

Fig. 4. AB preference test results for speech quality with 95% confidence intervals. iSE denotes the baseline VC system with i-vector, and SEN indicates the proposed jointly trained speaker embedding network.

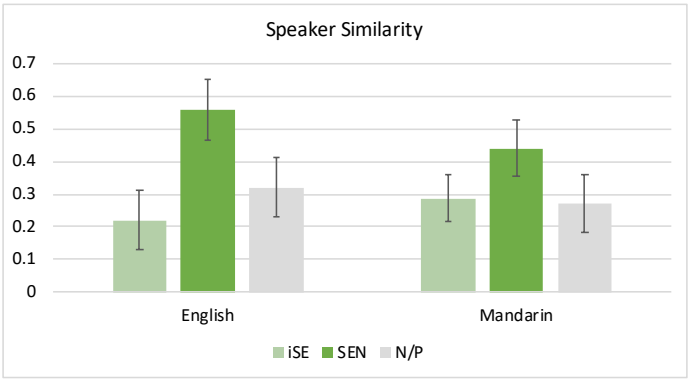

Fig. 5. XAB preference test results for speaker similarity with 95% confidence intervals. iSE and SEN denotes the baseline VC system with i-vector and the proposed jointly trained speaker embedding network, respectively. N/P means no preference.

where c_d and $c_d^{converted}$ are d-th dimension of the original and converted MCCs, and D indicates the MCC dimension. The lower value accounts for a smaller distortion.

The MCD results are presented in TABLE II. We observe that the proposed SEN always outperforms the baseline iSE with lower MCDs for all conversion experiments in two languages, which indicates that our proposed SEN is more effective than iSE in intralingual VC. Although our focus is cross-lingual VC, the intralingual conversion results are also meaningful to evaluate the system performance [46].

2) Subjective Evaluations: AB preference test was conducted to assess speech quality, and XAB preference test was also conducted to evaluate speaker similarity. 12 listeners were invited to participate in all the tests. 20 samples were randomly selected from 160 converted samples from each system. In AB preference tests, the listeners were asked to compare the quality and naturalness of the converted speech samples from different systems, and select the better one. Fig. 4 shows the speech quality test results, which suggests that our proposed approach outperforms the baseline system, and the quality improvement is more remarkable in English.

In XAB preference tests, X was the reference target speaker's speech, A and B were the randomly selected converted samples from different systems. The listeners were asked to chose the sample that was closer to the reference speaker's voice. The speaker similarity test results are presented in Fig. 5. It is observed that our proposed SEN outperforms the i-vector system in both English and Mandarin, and the difference is statistically significant in English.

Both objective and subjective results demonstrate the proposed jointly trained speaker embedding network consistently outperform the baseline VC system using i-vectors, which confirms the effectiveness of our proposed approach in terms of quality and similarity.

V. Conclusions

In this paper, we proposed a jointly trained speaker embedding network by integrating a speaker embedding network to the primary voice conversion network and optimizing it jointly with the rest of the model. A many-to-many cross-lingual voice conversion framework is implemented to validate the effectiveness of our proposed technique. Experimental results show that the proposed network can effectively improve the conversion performance in terms of both speech quality and speaker individuality compared to the average modeling voice conversion system using i-vector.

Acknowledgment

This research is supported by Ministry of Education, Singapore AcRF Tier 1 NUS Start-up GrantFY2016, Non-parametric approach to voice morphing. Yi Zhou is also funded by NUS research scholarship. The Mandarin library of average model database is provided by Data-baker[2].

References

[1] E. Godoy, O. Rosec, and T. 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.

[2] B. Ramani, M. A. Jeeva, P. Vijayalakshmi, and T. Nagarajan, "A multi-level GMM-based cross-lingual voice conversion using language-specific mixture weights for polyglot synthesis," *Circuits, Systems, and Signal Processing*, vol. 35, no. 4, pp. 1283–1311, 2016.

[3] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2010.

[4] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[5] M. Abe, K. Shikano, and H. Kuwabara, "Cross-language voice conversion," in *INTERSPEECH*, pp. 345–348, 1990.

[6] D. Erro and A. Moreno, "Frame alignment method for cross-lingual voice conversion," in *INTERSPEECH*, pp. 1969–1972, 2007.

[7] Y. Qian, J. Xu, and F. K. Soong, "A frame mapping based HMM approach to cross-lingual voice transformation," in *IEEE ICASSP*, pp. 5120–5123, 2011.

[2]http://www.data-baker.com

[8] H. Wang, F. Soong, and H. Meng, "A spectral space warping approach to cross-lingual voice transformation in HMM-based TTS," in *IEEE ICASSP*, pp. 4874–4878, 2015.
[9] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.
[10] M. Charlier, Y. Ohtani, T. Toda, A. Moinet, and T. Dutoit, "Cross-language voice conversion based on eigenvoices," in *INTERSPEECH*, pp. 1635–1638, 2009.
[11] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," in *IEEE ASRU*, pp. 676–681, 2003.
[12] S. Desai, B. Yegnanarayana, and K. Prahallad, "A framework for cross-lingual voice conversion using artificial neural networks," in *7th ICON*, 2009.
[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 Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.
[14] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *IEEE ICME*, pp. 1–6, 2016.
[15] F.-L. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN approach to cross-lingual TTS," in *IEEE ICASSP*, pp. 5515–5519, 2016.
[16] Y. Ning, Z. Wu, R. Li, J. Jia, M. Xu, H. Meng, and L. Cai, "Learning cross-lingual knowledge with multilingual blstm for emphasis detection with limited training data," in *IEEE ICASSP*, pp. 5615–5619, 2017.
[17] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *INTERSPEECH*, pp. 3364–3368, 2017.
[18] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen, "Can we steal your vocal identity from the internet?: Initial investigation of cloning obama's voice using gan, wavenet and low-quality found data," *arXiv:1803.00860*, 2018.
[19] B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura, "Adaptive wavenet vocoder for residual compensation in gan-based voice conversion," in *IEEE SLT*, pp. 282–289, 2018.
[20] B. Sisman, M. Zhang, and H. Li, "Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1085–1097, 2019.
[21] M. Zhang, X. Wang, F. Fang, H. Li, and J. Yamagishi, "Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet," *arXiv: 1903.12389*, 2019.
[22] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A training method of average voice model for HMM-based speech synthesis," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 86, no. 8, pp. 1956–1963, 2003.
[23] J. Wu, Z. Wu, and L. Xie, "On the use of i-vectors and average voice model for voice conversion without parallel data," in *IEEE APSIPA ASC*, 2016.
[24] X. Tian, J. Wang, H. Xu, E.-S. Chng, and H. Li, "Average modeling approach to voice conversion with non-parallel data," in *Proceedings of Odyssey The Speaker and Language Recognition Workshop*, pp. 227–232, 2018.
[25] M. Zhang, B. Sisman, S. S. Rallabandi, H. Li, and L. Zhao, "Error reduction network for dblstm-based voice conversion," in *IEEE APSIPA ASC*, pp. 823–828, 2018.
[26] T. Hashimoto, D. Saito, and N. Minematsu, "Many-to-many and completely parallel-data-free voice conversion based on eigenspace dnn," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 332–341, 2018.
[27] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 788–798, May 2011.
[28] R. K. Das, Abhiram B., S. R. M. Prasanna, and A. G. Ramakrishnan, "Combining source and system information for limited data speaker verification," in *INTERSPEECH*, pp. 1836–1840, 2014.
[29] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in neural information processing systems*, pp. 2962–2970, 2017.
[30] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *IEEE ICASSP*, pp. 6790–6794, 2019.
[31] L. Sun, H. Wang, S. Kang, K. Li, and H. M. Meng, "Personalized, cross-lingual TTS using phonetic posteriorgrams," in *INTERSPEECH*, pp. 322–326, 2016.
[32] B. Sisman, H. Li, and K. C. Tan, "Sparse representation of phonetic features for voice conversion with and without parallel data," in *IEEE ASRU*, pp. 677–684, 2017.
[33] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder," in *INTERSPEECH*, pp. 1978–1982, 2018.
[34] C. Xu, W. Rao, E. S. Chng, and H. Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," *arXiv:1903.09952*, 2019.
[35] C. Veaux, J. Yamagishi, K. MacDonald, *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.
[36] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *SSW*, 2004.
[37] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," in *INTERSPEECH*, pp. 1632–1636, 2016.
[38] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel method," *arXiv:1804.04262*, 2018.
[39] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, "The kaldi speech recognition toolkit," in *IEEE ASRU*, no. EPFL-CONF-192584, 2011.
[40] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.
[41] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *SSW*, 2016.
[42] B. Sisman, H. Li, and K. C. Tan, "Transformation of prosody in voice conversion," in *IEEE APSIPA ASC*, pp. 1537–1546, 2017.
[43] B. Sisman, G. Lee, and H. Li, "Phonetically aware exemplar-based prosody transformation.," in *Odyssey*, pp. 267–274, 2018.
[44] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *IEEE ICASSP*, vol. 3, pp. 1315–1318, 2000.
[45] T. M. T. K. Takayoshi Yoshimura, Keiichi Tokuda and T. Kitamura, "Incorporating a mixed excitation model and post filter into HMM-based text-to-speech synthesis," *Systems and Computers in Japan*, vol. 36, no. 12, pp. 43–50, 2005.
[46] A. F. Machado and M. Queiroz, "A flexible and modular crosslingual voice conversion system.," in *ICMC*, 2014.