Cross-lingual Voice Conversion with Disentangled Universal Linguistic Representations

Zhenchuan Yang, Weibin Zhang, Yufei Liu, Xiaofen Xing[*]

South China University of Technology

eezhenchuan@mail.scut.edu.cn

Abstract

Intra-lingual voice conversion has achieved great progress recently in terms of naturalness and similarity. However, in cross-lingual voice conversion, there is still an urgent need to improve the quality of the converted speech, especially with nonparallel training data. Previous works usually use Phonetic Posteriorgrams (PPGs) as the linguistic representations. In the case of cross-lingual voice conversion, the linguistic information is therefore represented as PPGs. It is well-known that PPGs may suffer from word dropping and mispronunciation, especially when the input speech is noisy. In addition, systems using PPGs can only convert the input into a known target language that is seen during training. This paper proposes an any-to-many voice conversion system based on disentangled universal linguistic representations (ULRs), which are extracted from a mix-lingual phoneme recognition system. Two methods are proposed to remove speaker information from ULRs. Experimental results show that the proposed method can effectively improve the converted speech objectively and subjectively. The system can also convert speech utterances naturally even if the language is not seen during training.

Index Terms: voice conversion, cross-lingual, disentangled universal linguistic representation

1. Introduction

Speech signal carries abundant information, among which the linguistic information and speaker characteristics are the most important ones[1].Voice conversion (VC) aims to change the speaker characteristics of speech signals while keeping the linguistic information unchanged[2]. Voice conversion is useful in various tasks such as speech enhancement[3][4], movie dubbing[5] or language learning[6], etc.

There are many works on intra-lingual voice conversion. Some traditional methodssuch as VTLN[7] and GMM[8], are conducted on parallel data i.e. the same content spoken by different speakers). Recently, deep neural networks have enabled systems built on non-parallel training data to generate converted speech with good quality. Some researchers propose to use generative adversarial networks (GANs) such as CycleGAN [9] [10] and StarGAN [11]. But the training of GAN is known to be much more sophisticated and unstable. In addition, the converted voice from a GAN model is not guaranteed to be of good quality [12]. Last but not least, methods based on GANs are usually end-to-end model, making it difficult to leverage other data resources, e.g. data for speech recognition. Therefore, the methods based on disentangling linguistic and speaker representations from the input speech are also interested to many researchers. During conversion, the linguistic content in the speech is preserved while the source speaker representation is replaced with that of the target speaker [12, 13, 14]. Among these approaches, phonetic posteriorgrams (PPGs [13, 15]) and its variants [16] are widely used.

Compared with intra-lingual VC, there are much less researches on cross-lingual voice conversion. Previous work [17] relies on paired data recorded from bilingual speakers speak both the source and target languages. Collecting large amount of training data from bilingual speakers is difficult. The corpus size is thus usually very small[18].

Cross-lingual VC based on phonetic posteriorgrams (PPG)([13], [19], [20]) can effectively leverage large amount of data from other tasks. PPGs are extracted from a speaker-independent automatic speech recognition (ASR) model. The ASR model itself is trained with a large variety of data, such as clean, noisy, far-field and near-field data, and with various accents. There is a large amount of ASR training data publicly available on the internet (e.g. the Librispeech [21]). On the other hand, acquiring high-quality VC training data is harder.

When using PPGs in the cross-lingual scenarios, an phone recognizer is trained with the combination of several different languages, such as English and Mandarin[20]. A hybrid dictionary is used to distinguish different words. The acoustic neural networks have a mixed output layer to predict different language modeling units. The disadvantages of using PPGs as the linguistic representations include: (1) Current works based on PPGs ([22], [19], [20]) cannot convert the input speech into a unseen language. (2) Usually the amount of data used to train the mix-lingual acoustic model is not balance. Thus the PPG outputs may be more biased towards the language with more training data. (3) We found that systems built on PPGs will suffer from word dropping and mispronunciation, especially when the input speech is noisy.

In this paper we propose an improved cross-lingual VC method based on disentangled universal linguistic representations (ULRs). A well-trained mix-languages acoustic model is used to extract ULRs from the input speech. We demonstrate that ULRs are even effective for unseen languages during training. When compared with the PPGs baseline on cross-lingual VC, our method can effectively improve the converted speech objectively and subjectively. The dimension of the ULRs features can also be easily adjusted. The dimension of ULRs is shrunk to a value much smaller than that of PPGs without hurting the conversion performance. This can reduce the model complexity.

This paper presents an approach of cross-lingual voice conversion on non-parallel data. The proposed approach will be elaborated in Section 2. Experimental setups and results are presented in Section 3. Finally we draw a conclusion in Section 4.

[*] Corresponding author. Email:xfxing@scut.edu.cn

2. Related work

Recent works on cross-lingual voice conversion usually use PPGs as the linguistic representations. [22] used monolingual PPGs to reconstruct the target speech. However, the system proposed in [22] is mainly designed for on-direction conversion from the source language to the target language. It is not suitable for conversions in both directions. To deal with this problem, Zhou *et. al.* proposes to use bilingual PPGs [19], which is obtained by combining the English PPGs and the Mandarin PPGs. Since the phone recognisers are trained separately, the bilingual PPGs may not provide a unified view on the input speech from two languages. Combining multiple PPGs from different languages also significantly increase the dimension of the linguistic vectors. Latter, mix-language PPGs, which serve as the baseline in our experiments, are proposed in [20].

3. Methodology

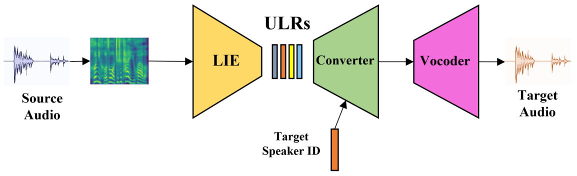

Figure 1: *The voice conversion system consists of a linguistic information extractor (LIE), a conversion model to generate target mel-spectrograms and a vocoder to reconstruct speech from mel-spectrograms.*

3.1. System architecture overview

As shown in Figure 1, the VC systems consists of a linguistic information extractor, a conversion model and a vocoder. The linguistic information extractor is used to extract the universal linguistic representations (ULRs). The converter takes the ULRs, together with the target speaker's ID as inputs, and converts them into target mel-spectrograms. Finally, the waveform will be generated from the converted mel-spectrograms by the vocoder.

All three parts are trained independently with different data sets. The linguistic information extractor and the converter will be elaborated in Subsecion 3.2 and Subsection 3.3 respectively. As for the vocoder, a standard speaker-independent multi-band WaveRNN is used. Details about WaveRNN can be found in [23].

3.2. Universal linguistic representations

We aim to find a linguistic representation that is language independent (i.e. universal) in order to do cross-lingual voice conversion between any language pair. To this end, we propose to use bottle-neck features extracted from a mix-language phone recognizer. However, compared with PPGs that contains no speaker information, bottle neck features extracted from a middle hidden layer may contain speaker information from the training data. This will affects the performance of the VC system. We propose two methods to eliminate the speaker information as much as possible. The first approach is based on domain adversarial training. Traditionally, the loss function used to train the phone recognizer is the cross entropy L_{pr} between the recognition outputs $\hat{\mathbf{l}}_n$ and the one-hot phone labels $\mathbf{l}_n$, i.e.

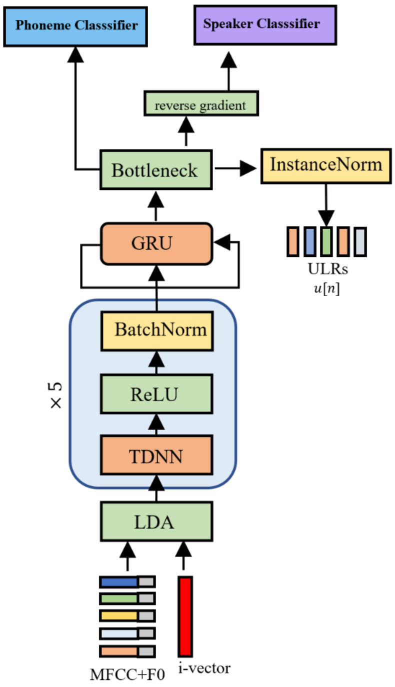

Figure 2: *Proposed linguistic information extractor(LIE). It aims to predict phoneme sequence and remove speaker information. It contains 5 TDNN layer and 1 GRU layer, and LDA is a linear transforms layer. ULRs are extracted from a linear bottleneck layer after GRU.*

$$L_{pr} = \frac{1}{N}\sum_{n=1}^{N} cross\_entropy(\mathbf{l}_n, \hat{\mathbf{l}}_n) \tag{1}$$

where N is the total number of speech frames. An auxiliary classifier is added to predict the correct speaker. The cross entropy loss L_{sr} between the predicted speaker probabilities $\hat{\mathbf{s}}_k$ and the target labels $\mathbf{s}_k$ is used to train the speaker classifier, i.e.,

$$L_{sr} = \frac{1}{M}\sum_{k=1}^{M} cross\_entropy(\mathbf{s}_k, \hat{\mathbf{s}}_k) \tag{2}$$

where M is the number of speech segments. The overall loss function used to train the mix-language recognizer is

$$L_{ulr} = L_{pr} - \lambda L_{sr} \tag{3}$$

where λ is used to tune the relative importance of the two terms.

The second approach is through instance normalization(IN), which has been used in [24] to remove speaker information. Suppose the output of the bottle neck layer is $\mathbf{x}_t$ at time t, then the proposed universal linguistic representations $\mathbf{u}_t$ can be calculated by normalizing $\mathbf{x}_t$, i.e.,

$$\sigma = \frac{1}{TW}\sum_{t=0}^{T}\sum_{i=0}^{W}\left(x_t^i - \mu\right)^2 \tag{4}$$

$$\mu = \frac{1}{TW}\sum_{t=0}^{T}\sum_{i=0}^{W} x_t^i \tag{5}$$

$$\mathbf{u}_t = \frac{\mathbf{x}_t - \mu}{\sqrt{\sigma + \epsilon}} \tag{6}$$

where x_t^i is the ith element of the input vector $\mathbf{x}_t$, u_t^i is the ith element of the output vector $\mathbf{u}_t$, ϵ is a small value to avoid numerical instability.

As for implementation details, Figure 2 shows the network architecture used in our experiments. The input of the phone recognizer includes MFCCs, $F0$ and i-vectors. LDA is used to reduce the dimension before feeding them into five layers of time delay neural networks (TDNN). A gated recurrent unit layer (GRU) is used after the TDNN layers. The GRU is followed by the bottle neck layer. Finally two classification tasks are built on top of the bottle neck layer. Compared to PPGs, it is much easier to adjust the dimension of the bottle neck layer and thus the dimension of ULRs.

3.3. The conversion model

The the conversion model takes the ULRs as input. It consists of a trainable lookup table that stores speaker's embeddings, a prenet that transforms both the ULRs and the speaker embedding into latent representations, and finally an auto-regressive decoder that generates the mel-spectrograms of the target speaker. The prenet and auto-regressive decoder are the same as that of Tacotron [25]. The overall architecture of the conversion model is shown in Figure 3.

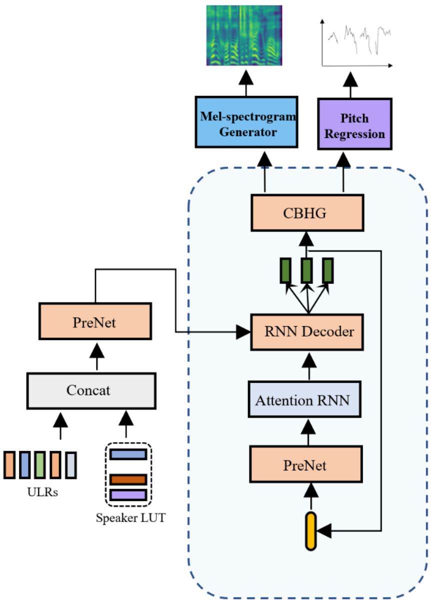

Figure 3: *Proposed conversion model. The prenet and auto-regressive decoder are same as Tacotron.*

We hope the conversion model can not only synthesize the mel-spectrograms, but also rhythmic information $F0$ and voiced/unvoiced. Thus during training, given a training example $\mathbf{e}$ (i.e. a speech segment) from a target speaker, mel-spectrograms $\mathbf{m}$ and rhythmic information $\mathbf{r}$ are extracted from $\mathbf{e}$. Then the model parameters $\Theta \triangleq \{\Theta_{pre}, \Theta_{lt}, \Theta_{decoder}\}$, where Θ_{pre} represents the parameters of the prenet, Θ_{lt} represents the parameters of the lookup table and $\Theta_{decoder}$ represents the parameters of the decoder, can be trained uing the following loss function

$$L = \sum \left(|\mathbf{m} - \hat{\mathbf{m}}| + \lambda \|\mathbf{r} - \hat{\mathbf{r}}\|^2\right) \tag{7}$$

where λ is a parameter to tune the importance of th two terms, $\hat{\mathbf{m}}$ and $\hat{\mathbf{r}}$ are the converter's predition outputs of the mel-spectrograms and rhythmic information respectively.

Our conversion model is a part of model based on Tacotron[25], in which the attention structure between encoder and decoder is removed, as shown in the Figure 3. It includes a lookup table speaker encoder, Prenet and CBHG, and an autoregressive RNN decoder same as [25]. It converts the Mel-spectrogram from source speaker to target speaker. The ULRs linguistic feature $u[n]$ and source speaker embedding e_s are concatenated into the converter, which restore those feature to $m_{s\to s}[n]$ to map the source audio features $m_s[n]$.

At test time, the rhythmic outputs from the converter is discarded. In addition, given the target speaker ID, we can retrieve the target speaker's embedding from the lookup table and concatenate it with the ULRs.

4. Experiments

4.1. Dataset

The mix-language phone recognizer is trained with both Librispeech[21] and AiShell2[26]. Librispeech contains 960 hours of reading English speech, while aishell2 contains 1000 hours of reading mandarin speech.

We conducted cross-lingual voice conversion experiments by using the dataset of task 2 in VCC2020[27]. The source set contains 4 English speakers (SEF1,SEF2,SEM1,SEM2). The target set includes 4 English speakers (TEF1,TEF2,TEM1,TEM2), 2 Finnish speakers(TFF1,TFM1), 2 German speakers (TGF1,TGM1) and 2 Mandarin speakers (TMF1,TMM1). The test set for evaluation consists of 25 utterances from 4 English speaker. For cross-lingual voice conversion, we performed $4 \times 6 = 24$ different source-target conversions using different models. We mainly compare the proposed method with the traditional mix-languager PPGs [20]. For fair comparison, only the linguistic representations are different.

4.2. experimental setup

In order to match the input of the mix-phone recognizer, the input source audios were down-sampled to 16kHz. 40-dimensional MFCC features were then extracted using a window length of 25ms and a window shift of 5ms. We also extracted 3-dimensional rhythmic feature, including $F0$, voice and unvoice. Also the 100-dimensional i-vector feature is extracted using a pretrained GMM model. For the target speech, we kept the sampling rate of 24KHz to ensure the quality, and adopted 80-dimensional Mel-spectrogram features with the same window length and window shift as above.

During training of the mix-phone recognizer, λ in Equation 3 was set to 0.5. The learning rate decayed from 0.0015 to 0.00015. For the training of the conversion model, we used the adam optimizer with 0.001 learning rate. The batch size was set to 32. The total number of training iterations is 150000. We used WaveRNN as the vocoder to synthesize final speech. The vocoder was trained in a multi-speaker way to generate speech with a sampling rate of 24kHz. We also used the mel-spectrograms synthesized by the converter model to fine-tune the vocoder.

4.3. Objective evaluation

For objective evaluation method, we mainly uses mel-cepstrum distortion (MCD)[1] to measure the spectral distance between two audio segments, defined as

$$MCD[dB] = 10/ln10\sqrt{2\sum_{d=1}^{D}\left(\hat{Y}_d - Y_d\right)^2} \qquad (8)$$

where D is the dimension of the mel-cepstrum feature, and $\hat{Y}_d$ and Y_d are the dth elements of the vectors. The lower the MCD, the smaller the distortion, meaning that the two audio segments are more similar.

We also adopt an automatic speech recognition (ASR) evaluation method. An ASR model is trained on Librispeech to evaluate sentence recognition results. The conversion result of the utterance is measured by word error rate(WER)[28], which is defined as

$$WER = \frac{S + D + I}{N} \qquad (9)$$

where S is the number of words with replacement errors, D is the number of words with deletion errors, I is the number of words with insertion errors, and N is the total number of words in the reference transcript.

Table 1: *Objective comparison between the baseline PPG-VC and the proposed ULR-VC.*

System	**WER**/(%)	**MCD**/(dB)
PPG-VC	48.31	5.85
ULR-VC	**13.51**	**5.43**
ground truth	8.50	0.00

Table 2 compares the baseline PPG-VC system and the proposed ULR-VC system on mel-cepstrum distortion and word error rate for cross-lingual voice conversion. We can see that the proposed ULR-VC system significantly outperforms the baseline PPG-VC system. We found that using mix-language PPGs leads to word dropping and mispronunciation, leading to bad word error rate. On the other hand, the proposed linguistic representation can significantly improve intelligibility and thus a lower WER.

Table 2: *Result of WER/MCD on different conversion pairs and genders of ULR-VC. Utterances from English speakers are converted to target speakers speaking in the language listed in the table. "F" represents female and "M" represents male. Noted that German and Finnish are not seen during the training of the linguistic information extractor.*

	English	**German**	**Finnish**	**Mandarin**
F→F	11.68/5.32	11.45/5.17	13.32/5.62	11.92/5.46
F→M	11.33/5.05	13.32/5.61	11.45/6.90	10.05/5.67
M→M	14.72/4.06	16.59/5.54	15.42/6.95	11.68/5.53
M→F	16.47/5.31	17.01/5.20	16.82/5.70	14.96/5.38

To see how genders and languages affect the conversion performance, we conducted a detailed analysis on the output generated by the proposed ULR-VC system and the results are shown in Table 3. For comparison, we also include the results of intra-lingual (i.e. English → English) in Table 3. As can be seen, intra-lingual conversions are better than cross-lingual conversions. In addition, converting men's voice to women's voice is the most difficult task, especially in cross-lingual conversion.

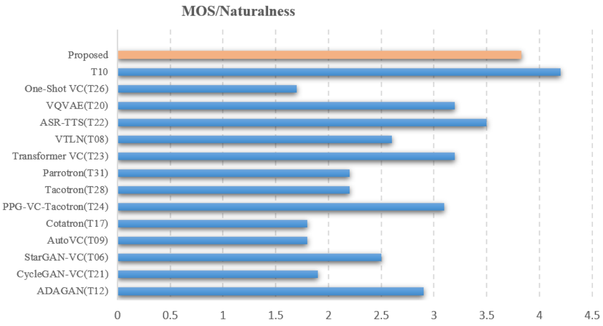

Figure 4: *MOS result compared with other systems in VCC2020 competition*

4.4. Subjective evaluation

The mean opinion score is a common multi-level scoring system used to evaluate the naturalness and similarity of the converted speech. It is usually divided into 5 levels from excellent to worse. 20 listener participated in this evaluation with randomly selected utterances from the experimental outputs. The results are shown in table 4. The proposed ULRs significantly outperform the PPGs in terms of speech naturalness and speaker similarity.

Table 3: *Mean opinion score (MOS) evaluation of different VC systems.*

	PPG-VC	**ULR-VC**
Naturalness	3.25±0.37	**3.83±0.27**
Similarity	3.16±0.32	**3.77±0.29**

We also compare the proposed method with some systems submitted for VCC2020 competition in Figure 3. The proposed system is comparable with the best system T10. However, the training of our model is much simpler.

5. Conclusion

In this paper, we propose an universal linguistic representation (ULR) for cross-lingual voice conversion with non-parallel data. ULRs are extracted from a bottle neck layer of a linguistic information extractor. To remove speaker information the bottle neck features, we propose to use domain adversarial training and instance normalization. Compared with commonly used PPGs, the proposed ULRs are much more compact and robust, rendering better voice conversion quality in both speech naturalness and speaker similarity.

6. References

[1] 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, no. 8, pp. 2222–2235, 2007.

[2] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[3] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, 2012.

[4] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Esophageal speech enhancement based on statistical voice conversion with gaussian mixture models," *IEICE TRANSACTIONS on Information and Systems*, vol. 93, no. 9, pp. 2472–2482, 2010.

[5] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4869–4873.

[6] 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.

[7] D. Sundermann and H. Ney, "Vtln-based voice conversion," in *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No.03EX795)*, 2003, pp. 556–559.

[8] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, "Voice conversion with smoothed gmm and map adaptation," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[9] A. Almahairi, S. Rajeshwar, A. Sordoni, P. Bachman, and A. Courville, "Augmented cyclegan: Learning many-to-many mappings from unpaired data," in *International Conference on Machine Learning*. PMLR, 2018, pp. 195–204.

[10] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6820–6824.

[11] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[12] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[13] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.

[14] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[15] S. Liu, L. Sun, X. Wu, X. Liu, and H. Meng, "The hccl-cuhk system for the voice conversion challenge 2018." in *Odyssey*, 2018, pp. 248–254.

[16] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *arXiv preprint arXiv:2009.02725*, 2020.

[17] M. Abe, K. Shikano, and H. Kuwabara, "Cross-language voice conversion," in *International Conference on Acoustics, Speech, and Signal Processing*, 1990, pp. 345–348 vol.1.

[18] M. Charlier, Y. Ohtani, T. Toda, A. Moinet, and T. Dutoit, "Cross-language voice conversion based on eigenvoices," 2009.

[19] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6790–6794.

[20] Y. Zhou, X. Tian, E. Yılmaz, R. K. Das, and H. Li, "A modularized neural network with language-specific output layers for cross-lingual voice conversion," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 160–167.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[22] L. Sun, H. Wang, S. Kang, K. Li, and H. M. Meng, "Personalized, cross-lingual tts using phonetic posteriorgrams." in *INTERSPEECH*, 2016, pp. 322–326.

[23] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2410–2419.

[24] J.-c. Chou, C.-c. Yeh, and H.-y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," *arXiv preprint arXiv:1904.05742*, 2019.

[25] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: A fully end-to-end text-to-speech synthesis model," *arXiv preprint arXiv:1703.10135*, vol. 164, 2017.

[26] J. Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[27] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," *arXiv preprint arXiv:2008.12527*, 2020.

[28] B. H. Juang and L. R. Rabiner, "Hidden markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, 1991.