Cross-lingual Voice Conversion with Disentangled Universal Linguistic Representations

Zhenchuan Yang[1], *Weibin Zhang*[2], *Yufei Liu*[1], *Xiaofen Xing*[1]*

[1]South China University of Technology, China
[2] VoiceAI Technologies Co. Ltd., China
xfxing@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 methods, such 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 pure GAN based model 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 models, 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 by bilingual speakers. 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 (PPGs)([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 ([19, 20, 22]) cannot convert the input speech into an unseen language. (2) Usually the amount of data used to train the mix-lingual acoustic model is not balanced. Thus the PPGs 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 and can be shrunk to a much smaller size than that of PPGs without hurting the conversion performance. This can reduce the model complexity.

This paper presents an approach for cross-lingual voice conversion by using non-parallel training data. Related works are summarized in Section 2. The proposed approach is elaborated in Section 3, followed by experimental setups and results in Section 4. Finally we draw a conclusion in Section 5.

*Corresponding author.

2. Related work

Recent works on cross-lingual voice conversion usually use PPGs as the linguistic representations. Sun used monolingual PPGs to reconstruct the target speech [22] . However, the system proposed in [22] is mainly designed for unidirectional 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 recognizers 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 increases the dimension of the linguistic vectors. Latter, mix-language PPGs, which serve as the baseline in our experiments, are proposed in [20].

3. Methodology

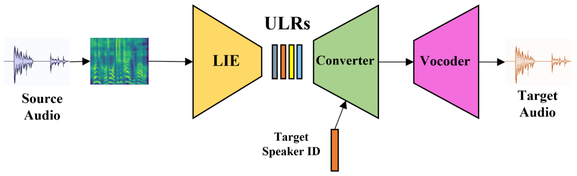

Figure 1: *The voice conversion system consists of a linguistic information extractor (LIE), a conversion model to generate the target Mel-spectrograms and a vocoder to reconstruct speech from Mel-spectrograms.*

3.1. System architecture overview

As shown in Figure 1, the VC system consists of a linguistic information extractor (LIE), a converter and a vocoder. The LIE 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 the target Mel-spectrograms. Finally, the speech wave is reconstructed by the vocoder.

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

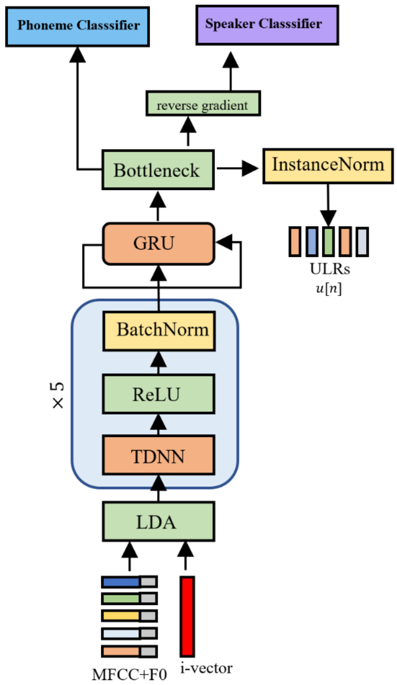

Figure 2: *The proposed linguistic information extractor(LIE). It aims to predict phoneme labels and remove speaker information. ULRs are extracted from a linear bottleneck layer after GRU.*

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 bottleneck features extracted from a mix-language phone recognizer. However, compared with PPGs that contains no speaker information, bottleneck 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 (DAT [24]). Traditionally, the loss function used to train the phone recognizer is the cross entropy between the recognition outputs $\hat{\mathbf{l}}_n$ and the one-hot phone labels $\mathbf{l}_n$, i.e.

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

where N is the total number of speech frames. An auxiliary classifier is added to predict the correct speaker. The cross entropy loss 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 neural network 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 that we propose to eliminate the speaker information is through instance normalization(IN), which has been used in [25]. Suppose the output of the bottleneck 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. W is the dimention of $\mathbf{x}_t$ (and also $\mathbf{u}_t$), and T is the number of frames.

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 a time delay neural networks (TDNN) of five layers. A gated recurrent unit layer (GRU) is used after the TDNN layers. The GRU is followed by the bottleneck layer. Finally, two classification tasks are built on top of the bottleneck layer. Compared to PPGs, it is much easier to adjust the dimension of the bottleneck layer and thus the dimension of ULRs.

3.3. The conversion model

As shown in the Figure 3, the proposed conversion model is based on Tacotron [26]. The Prenet, CBHG, and the autoregressive RNN decoder are same as [26]. The attention structure between the encoder and the decoder is removed since the input and output features are naturally aligned in voice conversion. On the other hand, the conversion model includes a trainable lookup table to encode the target speakers' ID.

We hope the conversion model can not only synthesize the Mel-spectrograms, but also the rhythmic information $F0$ and voiced/unvoiced. Thus during training, the conversion model takes both the target speaker's embedding and the ULRs as input, and firstly transforms them into the latent representations. Then the latent representations are fed into the auto-regressive decoder to generate the Mel-spectrograms and the rhythmic features. At test time, the rhythmic outputs from the converter are discarded and only the Mel-spectrograms are fed into the vocoder.

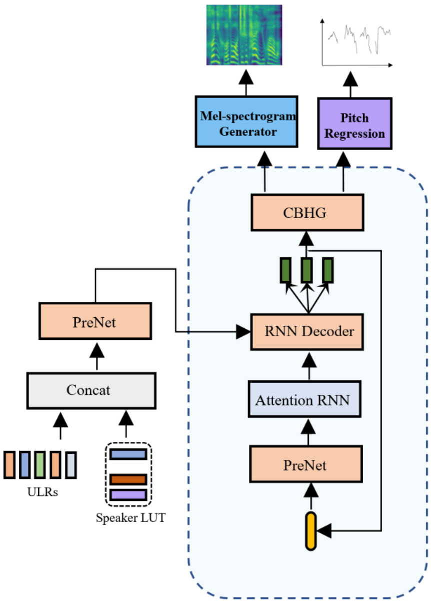

Figure 3: *The proposed conversion model. The prenet and auto-regressive decoder are the same as Tacotron [26].*

The conversion model is trained using a unsupervised way. 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 using the following loss function

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

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

4. Experiments

4.1. Dataset

The mix-language phone recognizer was trained with both Librispeech [21] and AiShell2 [27]. Librispeech contained 960 hours of reading English speech, while Aishell2 contained 1000 hours of reading Mandarin speech.

We conducted cross-lingual voice conversion experiments by using the dataset of task 2 in VCC2020 [28]. The source set contained 4 English speakers (SEF1, SEF2, SEM1, SEM2). The target set included 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 consisted of 25 utterances from 4 English speaker. For cross-lingual voice conversion, we performed $4 \times 6 = 24$ different source-target conversions . We mainly compared the proposed method with the traditional mix-language PPGs [20]. For fair comparison, only the linguistic representations were different.

4.2. Experiment 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 features, including $F0$, voiced and unvoiced. As a typical setup in Kaldi [29], we also extracted the 100-dimensional i-vector feature using a pretrained GMM model. As for the supervised features used for training the conversion model, we kept the 24kHz sampling rate of the target speech to ensure the quality. 80-dimensional Mel-spectrogram features with the same window length and window shift as above were extracted.

During training of the mix-phone recognizer, λ in Equation (3) was set to 0.5. The dimension of ULRs was 850 in our experiments. Both DAT and IN were used to remove the speaker information from the ULRs. 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. $gamma$ in Equation (7) was set to 0.1. The batch size was set to 32. The total number of training iterations was 150000. We used WaveRNN as the vocoder to synthesize the final speech. The vocoder was trained 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.

[*] Speech samples are available on https://pigzach.github.io/crosslingual-ulr-vc/.

4.3. Objective evaluation

To evaluate the proposed method objectively, we mainly use Mel-cepstrum distortion (MCD) [1] to measure the spectral distance between two audio segments. MCD is defined as

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

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 is, the smaller the distortion, meaning that the two audio segments are more similar to each other.

We also adopt an automatic speech recognition (ASR) evaluation method. An ASR model is trained on Librispeech to evaluate the word error rate on the converted speech sentence recognition results. The conversion result of the utterance is measured by word error rate(WER) [30].

Table 1: *Objective comparison between the baseline PPG-VC system and the proposed ULR-VC system. The result is obtained from the official test set of 4 target speakers, each with 25 sentences.*

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

Table 1 compares the baseline PPG-VC system and the proposed ULR-VC system based on MCD and WER for cross-lingual voice conversion. Since the source set contains only English speaker, all the converted audios are in English. Therefore we can use a single English model to recognize the converted speech. We can see from Table 1 that the proposed ULR-VC system significantly outperforms the baseline PPG-VC system. We find that using mix-language PPGs leads to word dropping and mispronunciation, resulting in a bad word error rate. On the other hand, the proposed ULRs help making the converted speech sounds much more clear.

Table 2: *Results 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 data are not seen during the training of the LIE.*

	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 conduct a detailed analysis on the outputs generated by the proposed ULR-VC system and the results are shown in Table 2. For comparison, we also include the results of intra-lingual conversion (i.e. English → English) in Table 2. As can be seen, intra-lingual conversions are easier than cross-lingual conversions. In addition, converting men's voice to women's voice is the most difficult task, especially in cross-lingual conversion. Finally, even if German and Finnish data are not seen during the training of the LIE, the proposed system can accomplish the conversions with good-quality outputs.

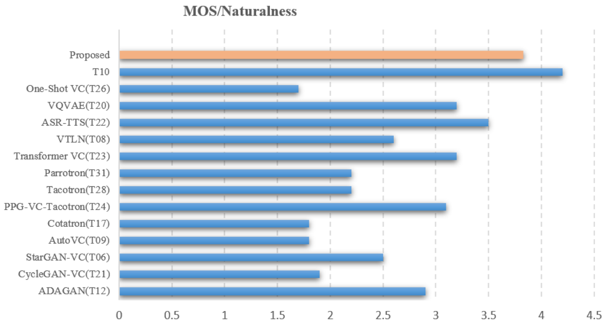

Figure 4: *MOS result compared with other systems in VCC2020 competition*

4.4. Subjective evaluation

The mean opinion score (MOS) is a common multi-level scoring mechanism used to evaluate the naturalness and the similarity of the converted speech. MOS scores are usually divided into 5 levels from excellent to worse. 20 listeners participated in this evaluation with randomly selected utterances from the experimental outputs. The results are shown in Table 3. The proposed ULRs significantly outperforms 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 4. The proposed system is comparable with the best system T10 [28]. However, the training of our model is much simpler.

5. Conclusion

In this paper, we propose an universal linguistic representations (ULRs) for cross-lingual voice conversion with non-parallel data. ULRs are extracted from a bottleneck layer of a linguistic information extractor. To remove speaker information from the bottleneck features, we propose to use domain adversarial training and instance normalization. Compared with the 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. Acknowledge

The work is supported by Key-Area Research and Development Program of Guangdong 2019B010154003, the National Natural Science Foundation of China U1801262, and the Science and technology Project of Guangzhou 202103010002.

7. 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 *Proc. ICASSP*, 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 *Proc. ISSPIT(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 *EUROSPEECH*, 2003.

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

[10] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *Proc. ICASSP*, 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 *SLT*, 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 *ICML*. 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 *ICME*, 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 *Proc. ICASSP*, 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," in *Proc. INTERSPEECH*, 2009, pp. 1635–1638.

[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 *Proc. ICASSP*, 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 *Proc. ASRU*, 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 *Proc. 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 *Proc. 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 *ICML*, 2018, pp. 2410–2419.

[24] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

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