INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Cross-Lingual Voice Conversion with a Cycle Consistency Loss on Linguistic Representation

Yi Zhou[1], *Xiaohai Tian*[1], *Zhizheng Wu*[2] *and Haizhou Li*[1]

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]Facebook Inc, United States

yi.zhou@u.nus.edu, xiaohai171@gmail.com, wuzhizheng@gmail.com, haizhou.li@nus.edu.sg

Abstract

Cross-Lingual Voice Conversion (XVC) aims to modify a source speaker identity towards a target while preserving the source linguistic content. This paper introduces a cycle consistency loss on linguistic representation to ensure the speech content unchanged after conversion. The proposed XVC model consists of two loss functions during optimization: a spectral reconstruction loss and a linguistic cycle consistency loss. The cycle consistency loss seeks to maintain the source speech's linguistic content. Specifically, we utilize Phonetic PosteriorGram (PPG) to represent the linguistic content. XVC experiments were conducted between English and Mandarin. Both objective and subjective evaluations demonstrated that with the proposed cycle consistency loss, converted speech is more intelligible.

Index Terms: Cross-Lingual Voice Conversion (XVC), Cycle Consistency Loss, Phonetic PosteriorGram (PPG)

1. Introduction

Cross-Lingual Voice Conversion (XVC) is to convert the speaker identity of one speaker (source) to match that of another (target), where the source and target speak different languages. XVC can enable many applications, such as language education, foreign movie dubbing and so on [1]. The conversion seeks to achieve two objectives: one is to convert the speaker identity from the source to the target [2]; the other one is to maintain the linguistic content of the source speech. Most of the previous studies are focusing on achieving the former objective. While this paper aims to the optimize both objectives simultaneously.

Due to the language difference in XVC, it is difficult to obtain parallel data, where the source and target data have the same speech content [3]. Hence, non-parallel techniques become a mainstream solution for XVC. Early works explored various alignment techniques to find speech segment pairs between the source and target speech across different languages [4–8]. With the advance of deep learning, neural XVC networks have shown their strengths in generating high quality converted speech. According to the submitted systems in Voice Conversion Challenge 2020 (VCC2020) [9], XVC approaches can be grouped into three main categories including generative adversarial networks (GAN) [10, 11], sequence-to-sequence (seq2seq) mapping methods [12–14] and the encoder-decoder frameworks [15–17]. The encoder-decoder structure is reported as one of the most popular solutions. In such frameworks, speech signals are first encoded into speaker-independent linguistic representations, and then they reconstruct acoustic features or time-domain speech signals from the linguistic representations [15–17]. Among the encoder-decoder structures, Phonetic PosteriorGram (PPG) [18, 19] has been adopted by the majority VCC2020 systems as the linguistic representation, that could be either the posterior probability or bottleneck features from a speech recognition neural network [9]. In PPG-based XVC frameworks, the conversion model learns a feature mapping between linguistic features and the corresponding acoustic features [20]. Besides, there are also successful attempts combining seq2seq method with the GAN model to achieve rhythm-flexible voice conversion by modifying the PPG length [19].

Despite their success, most existing XVC approaches mainly focus on spectral feature optimization, but do not explicitly force the linguistic representation of the converted speech to match that of the source speech [9]. Although spectral features contain linguistic information, they may not be sufficient to preserve the speech content. In the informal listening test, we found the converted speech by XVC sometimes sounds not native [9, 21, 22]. The reason might be that spectral reconstruction loss is optimized to match the speaker identity of the target, but the loss also leads the conversion model to capture the articulation of the target speech from a different language. As a result, the conventional XVC cannot preserve the original source speech's native pronunciation or articulation.

This paper introduces a cycle consistency loss on the linguistic representation. The loss explicitly forces the converted speech to maintain the same linguistic content with the source speech. The proposed XVC framework utilizes a modularized neural network [22, 23] to reconstruct acoustic features from PPG features in two languages. Specifically, the network is optimized with two losses: an acoustic feature reconstruction loss to ensure that the converted spectral features are close to that of the target, and a cyclic loss on PPG features to force the converted speech to carry the same linguistic content as the natural input speech. The extra loss on linguistic representation is expected to help the network generate highly intelligible speech.

2. PPG-based XVC with a Modularized Neural Network

A Modularized Neural Network (MNN) was proposed in [22]. It maps input PPGs to output acoustic features across English and Mandarin with multi-task learning [24]. A MNN consists of a language-independent module and a language-specific output module. The language-independent module is shared by two languages, while the language-specific output module models the acoustic features in each language individually via two branches: one for English and the other one for Mandarin [22].

Figure 1(a) illustrates the PPG-based XVC framework with MNN. It utilizes both English and Mandarin speech data from multiple speakers during training. The English and Mandarin PPGs are first extracted from acoustic features by English and Mandarin phone recognizers, respectively. Two linear projection layers are employed to project the English and Mandarin PPGs into latent features, which are then concatenated to form

http://dx.doi.org/10.21437/Interspeech.2021-687

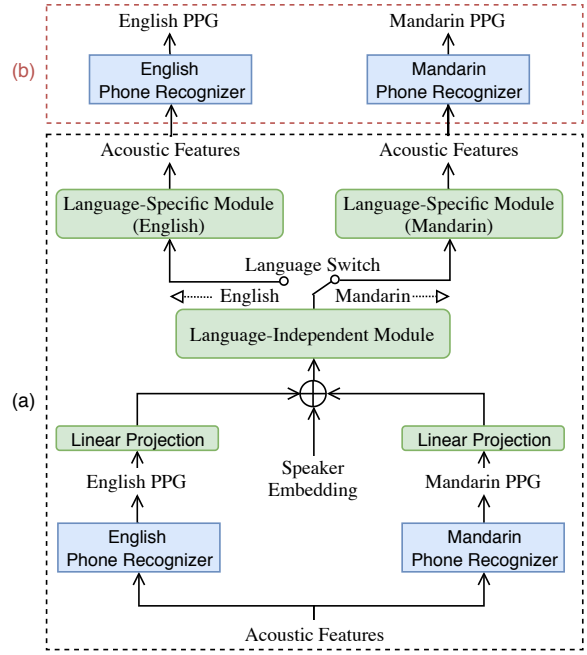

Figure 1: *(a) PPG-based XVC framework with the modularized neural network (MNN). (a) and (b) form the proposed XVC framework with linguistic cycle consistency loss.*

a bilingual linguistic representation [20]. Concatenating with a speaker embedding, these features are forwarded into the shared language-independent module consisting of two bidirectional long short-term memory (BLSTM) layers. Parameters in this shared module are updated with both languages. Selected by the language switch, the shared module output features will only pass through the branch in the language of the input speech. Each branch in the language-specific module has two dense layers, where parameters are only updated with its corresponding language. The English and Mandarin phone recognizers are set as non-trainable networks whose model parameters are frozen during training. The training objective is to optimize root mean square errors (RMSE) between reconstructed acoustic features Y_{Recon} and the reference ones Y_{Ref} :

$$\mathcal{L}_{\text{acoustic}} = \sqrt{\sum \frac{(Y_{\text{Recon}} - Y_{\text{Ref}})^2}{N}}, \tag{1}$$

where N is the number of total frames in the training data.

At run-time, taking the case of converting a Mandarin speech to an English speaker's voice as an example, acoustic features extracted from the source Mandarin speech are first encoded as the bilingual linguistic representation. Speaker embedding is obtained from English speech samples uttered by the target speaker, and concatenated with the bilingual linguistic representation. After passing through the shared BLSTM layers, the intermediate features are forwarded to the Mandarin branch. Last, a vocoder is used to generate speech waveform from the acoustic features [22].

3. XVC with a Linguistic Cycle Consistency Loss

Even though spectral reconstruction loss works well in the PPG-based XVC framework, we find the converted speech sometimes sounds not as native as the source natural speech [22]. It motivates us to consider introducing an extra loss on the linguistic representation, which is used to force the linguistic content of the converted speech to match that of the source speech. This will give better intelligibility, thereby further improve the converted speech quality. In this section, we first discuss the cycle consistency loss, which motivates our research. We then detail the proposed XVC with a cycle consistency loss on PPG.

3.1. Linguistic Cycle Consistency Loss

Cycle consistency loss assumes a reversible feature transformation process [25]. It has been applied to a number of speech-related applications including machine translation [26], speech recognition [25], speech synthesis [27] and voice conversion [17, 28, 29]. In the context of XVC, we expect that the model only changes the speaker identity while keeping the same underlying linguistic content. To facilitate the conversion model preserving linguistic content, the cycle consistency loss is designed to minimize the difference between linguistic representations obtained from input and converted acoustic features.

3.2. XVC with a Linguistic Cycle Consistency Loss

As shown in Figure 1(b), during training, we incorporate an extra component to extract PPGs in the XVC framework to explicitly model linguistic content for high intelligibility. It adds on phone recognizers to the language-specific output module in the MNN. In particular, the language of the phone recognizer and MNN language-specific module are in accordance with the input speech.

During training, the XVC network first extracts PPGs from input acoustic features, which will be transformed into latent bilingual linguistic representations for acoustic feature reconstruction. Reconstructed acoustic features are further passed into the phone recognizer to extract PPGs, which are called cyclic PPG. Whereas PPGs extracted from input acoustic features are referred to as reference PPG. The bottleneck features are utilized as PPGs, and details can be found from Section 4.1. The linguistic cycle consistency loss can be calculated as,

$$\mathcal{L}_{\text{PPG}} = \sqrt{\sum \frac{(PPG_{\text{Cyc}} - PPG_{\text{Ref}})^2}{N}}, \tag{2}$$

where PPG_{Cyc} is cyclic PPG and PPG_{Ref} stands for reference PPG. N represents the total frame number in the training data. The XVC network is trained to jointly minimize the spectral reconstruction loss and linguistic cycle consistency loss as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{acoustic}} + \lambda \mathcal{L}_{\text{PPG}} \tag{3}$$

where $\lambda (0 < \lambda < 1)$ is the coefficient on the linguistic cycle consistency loss.

At run-time, the extra component in Figure 1(b) for linguistic cycle consistency is not necessary. Therefore, the conversion process is exactly the same as the PPG-based XVC framework described in Section 2.

4. Experiments and Results

We conducted XVC experiments between English and Mandarin to compare performance between the PPG-based XVC frameworks without/with the linguistic cycle consistency loss.

4.1. Data and Feature Extraction

- **Phone Recognizer**: Two phone recognizers for English and Mandarin were pretrained for PPG extraction. The English phone recognizer was trained with the Librispeech database (460 hours) [30]. The Mandarin phone recognizer was trained with multiple Mandarin database (1238 hours) including aidataTang [31], AISHELL-1 [32], MagicShell [33], PrimeWords [34], ST-CMDS [35] and THCHS-30 [36]. Both English and Mandarin phone recognizers shared the same model architecture. It consisted of three 5×1 convolutional layers with 512 channels in each layer. Each convolutional layer was followed by batch normalization and ReLU activation. Then the output was fed into three BLSTM layers of 512 units, followed by a bottleneck dense layer with 256 units. The output softmax layers had 5,808 and 9,864 units for English and Mandarin, respectively. The input was 80-dimensional Mel-spectrogram, and the output was the senone class. The frame accuracy of the English phone recognizer was 73.84%, while that was 76.37% of the Mandarin phone recognizer. Both English and Mandarin PPG dimensions were 256.

- **Speaker embedding extractor**: The speaker embedding extractor was trained on the TIMIT database [37] from 630 speakers, each speaker contributed 10 utterances. The network had two BLSTM layers with 768 units in each layer, followed by a bottleneck dense layer with 256 units [38]. The input was 80-dimensional Mel-spectrogram, and the extracted speaker embedding dimension was 256. The equal error rate (EER) on 24 speakers in the TIMIT test data was 3.21%.

- **XVC database**: XVC experiments utilized a multi-lingual and multi-speaker database consisting of 100 speakers (50 English speakers and 50 Mandarin speakers), each speaker contributed 150 utterances. 50 English speakers were randomly selected from the VCTK database [39] while 50 Mandarin speakers were from the Data-Baker Mandarin Library[1]. During testing, 4 bilingual speakers (MF2, MF4, MM1 and MM2) from the EMIME database [40] were selected, with 20 utterances from each speaker.

All speech signals were resampled to 16 kHz. The Mel-spectrogram were extracted at 50 ms window size and 12.5 ms frameshift with the dimension of 80.

4.2. Experimental Systems

- **Baseline**: The PPG-based XVC framework discussed in Section 2 is the baseline system **without** linguistic cycle consistency loss. Both projection dense layers had 256 units. The BLSTM layers in the MNN language-independent module had 768 units. The language switch was a 0/1 flag. Two dense layers in the language-specific module had 256 units and 80 units, respectively. The network was trained with 0.001 learning rate. The batch size was set to 16.

- **CyclePPG-XVC**: The XVC framework introduced in Section 3.2 is our proposed XVC system **with** linguistic cycle consistency loss. It shared the same neural network architecture and hyperparameters with the Baseline system. The λ value was chosen to be 0.7 by trial and error.

We used the WaveRNN neural vocoder [41] to generate speech waveform from Mel-spectrogram for both experimental systems. The details of WaveRNN model can be found in [23].

[1] http://www.data-baker.com/hc_pm_en.html

Table 1: *A summary of objective evaluation metrics including PPG RMSE, MSD, RMSE and WER. 'Source' indicates the source speech. en→cn is the case when we convert a source English utterance to a Mandarin target speaker, and vice-versa.*

		Source	Baseline	CyclePPG-XVC
PPG RMSE	en→cn	N.A	13.89	9.64
	cn→en		13.76	9.19
	average		13.83	9.42
MSD (dB)	en→cn	15.43	12.39	11.52
	cn→en	16.18	12.96	11.46
	average	15.81	12.68	11.49
RMSE (dB)	en→cn	17.48	14.02	12.99
	cn→en	19.96	13.20	13.17
	average	18.72	13.61	13.08
WER (%)	en→cn	8.21	35.50	16.67
	cn→en	3.75	26.63	10.33
	average	5.98	31.07	13.50

4.3. Objective Evaluations

We first report PPG RMSE as a measurement of the linguistic content preservation. Then we compute Mel-spectrogram distortion (MSD) [42] to measure the spectral distance between the reconstructed and reference Mel-spectrogram. Also, we calculate speech RMSE to evaluate the signal distortion. Last, we use the speech-to-text APIs to measure a voice's intelligibility in terms of word error rate (WER). In all objective metrics, a lower value indicates a better result.

- **PPG RMSE**: It is calculated between cyclic PPG and reference PPG. From Table 1, it can be found that the average PPG RMSE value drops from 13.83 dB to 9.42 dB for Baseline system and CyclePPG-XVC system, respectively. It suggests that the linguistic cycle consistency loss constrains the system to generate acoustic features with linguistic content that is closer to the source speech.

- **MSD**: With Y_{Recon} and Y_{Ref} representing the reconstructed and reference Mel-spectrogram, MSD is defined as:

$$MSD[dB] = 10/\ln 10 \sqrt{2 \sum ((Y_{\text{Recon}} - Y_{\text{Ref}})^2}, \quad (4)$$

Average MSD results are presented in Table 1. It is observed that average MSD value decreases from 12.68 dB (Baseline) to 11.49 dB (CyclePPG-XVC). Compared to the Baseline, the proposed CyclePPG-XVC effectively reduces the spectral distortion by adding the linguistic cycle consistency loss.

- **RMSE**: The RMSE between converted and natural reference speech can be computed as:

$$RMSE[dB] = \sqrt{\frac{1}{M} \sum (20 \log_{10}(\frac{|F_{\text{Gen}}|}{|F_{\text{Ref}}|}))}, \quad (5)$$

where M is the number of frequency bins, $|F_{\text{Gen}}|$ and $|F_{\text{Ref}}|$ are the corresponding magnitude values of the generated and reference speech, respectively. As presented in Table 1, CyclePPG-XVC achieves a lower RMSE value (13.08 dB) than the Baseline (13.61 dB). It confirms that with linguistic cycle consistency loss, the proposed approach is able to generate speech signals with a smaller distortion.

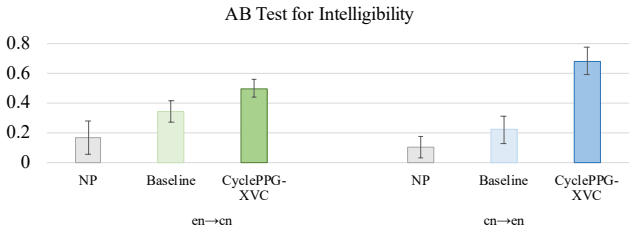

Figure 2: *AB test result on intelligibility with* 95% *confidence intervals. A higher value accounts for a more intelligible speech. 'NP' means no preference.*

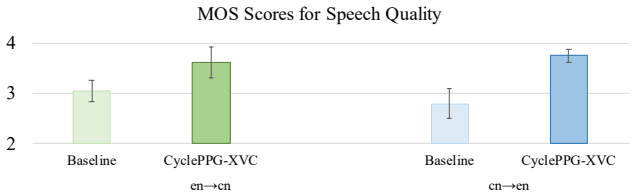

Figure 3: *MOS test result on speech quality with* 95% *confidence intervals.*

- **WER**: First, we obtain the transcripts of the English and Mandarin utterances using the Google Cloud Speech-to-Text[2] and iFLYTEK Open Platform[3] APIs, respectively. Then we further calculate the Levenshtein distance for WER. In Table 1, we find that the Baseline achieves an average WER of 31.07% while the proposed CyclePPG-XVC obtains 13.50%, which shows that using linguistic cycle consistency loss greatly enhances the synthesized voice intelligibility.

4.4. Subjective Evaluations

We further conduct three subjective listening tests: AB test, Mean Opinion Score (MOS) test, and XAB test. 20 bilingual English and Mandarin listeners have participated in all listening tests. They are all university students and staffs aged between 20 and 40, proficient in both English and Mandarin. Each listener was presented with 12 randomly selected speech samples from 160 converted utterances.

- **AB Test for Intelligibility**: In AB test, we focus on intelligibility. A and B are two randomly selected converted speech samples from different experimental systems. Listeners were asked to listen and compare A and B, then choose the one that was more intelligible. In particular, they were requested to assess the extent of whether the converted speech are understandable at both word and sentence level including the pronunciation and articulation. The result is presented in Figure 2. It is obvious that CyclePPG-XVC significantly outperforms the Baseline in both en→cn and cn→en conversions. With linguistic cycle consistency loss, CyclePPG-XVC has demonstrated a performance improvement on speech intelligibility. Hence, we may declare the effectiveness of the proposed mechanism in preserving the source linguistic content.

- **MOS Speech Quality Test**: In MOS test, listeners needed to listen to each converted sample and rate the overall quality including intelligibility and naturalness at a 5-point scale. A higher MOS score indicates a better speech quality. Results are presented in Figure 3, which show that CyclePPG-XVC outperforms Baseline significantly in both en→cn and cn→en conversions. It is an evidence that introducing linguistic cycle consistency loss favors a XVC system to generate higher quality speech samples.

[2]https://cloud.google.com/speech-to-text
[3]https://www.iflyrec.com/

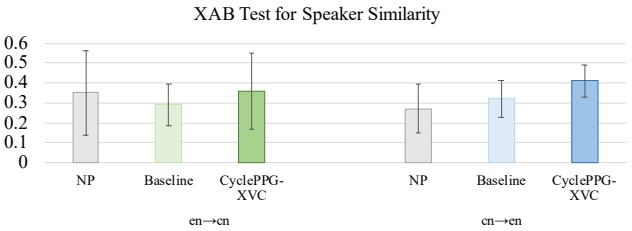

Figure 4: *XAB test result on speaker similarity with* 95% *confidence intervals. 'NP' is short for no preference.*

- **XAB Speaker Similarity Test**: In XAB test, X is given as reference, which is the natural speech from the target speaker. A and B are two converted samples from different experimental systems. The listeners were told to listen to both samples and to select the one that sounded closer to the reference speech in terms of speaker identity. As shown in Figure 4, the proposed CyclePPG-XVC outperforms the Baseline, although the improvements are not significant. It is reasonable since the proposed cyclePPG-XVC incorporates an extra cycle consistency loss to preserve the linguistic content whereas no dedicated effort has been made to improve the speaker identity characterization during conversion.

Overall, it is noted that the proposed CyclePPG-XVC obtains better results consistently than the Baseline in all objective and subjective evaluations for both en→cn and cn→en conversions. It confirms that CyclePPG-XVC achieves an enhanced intelligibility of the converted speech while keeping a similar result with the Baseline in terms of the speaker similarity[4].

5. Conclusion

We have proposed a cycle consistency loss on linguistic representation for XVC. The network is optimized with a spectral reconstruction loss and a linguistic cycle consistency loss jointly. The linguistic cycle consistency loss aims to ensure the converted speech maintain the same linguistic content as the input speech. Experiment results demonstrate that the proposed method is able to enhance the converted speech quality and intelligibility significantly.

6. Acknowledgements

This work was supported by the Science and Engineering Research Council, Agency of Science, Technology and Research, Singapore, through the National Robotics Program under Grant No. 192 25 00054 and Programmatic Grant No. A18A2b0046 from Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain). Project Title: Human Robot Collaborative AI for AME. Yi Zhou is also supported by the NUS Research Scholarship.

[4]The converted samples are available at this website: https://vcsamples.github.io/interspeech2021/.

7. References

[1] 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.

[2] M. Mashimo, T. Toda, H. Kawanami, K. Shikano, and N. Campbell, "Cross-language voice conversion evaluation using bilingual databases," *IPSJ Journal*, vol. 43, no. 7, pp. 2177–2185, 2002.

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

[4] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *IEEE ICASSP*, vol. 1, 1996, pp. 373–376.

[5] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," in *IEEE ASRU*, 2003, pp. 676–681.

[6] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg, "Text-independent cross-language voice conversion," in *INTERSPEECH*, 2009.

[7] 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.

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

[9] Z. Yi, 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—," in *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 80–98.

[10] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion," in *IEEE ICASSP*, 2019, pp. 6820–6824.

[11] M. Patel, M. Purohit, M. Parmar, N. J. Shah, and H. A. Patil, "AdaGAN: Adaptive GAN for many-to-many non-parallel voice conversion," 2019.

[12] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks." in *INTERSPEECH*, vol. 2017, 2017, pp. 1283–1287.

[13] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 540–552, 2019.

[14] W.-C. Huang, T. Hayashi, S. Watanabe, and T. Toda, "The sequence-to-sequence baseline for the voice conversion challenge 2020: Cascading ASR and TTS," *arXiv:2010.02434*, 2020.

[15] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *IEEE APSIPA*, 2016, pp. 1–6.

[16] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *IEEE ICASSP*, 2018, pp. 5274–5278.

[17] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AUTOVC: Zero-shot voice style transfer with only autoencoder loss," *arXiv:1905.05879*, 2019.

[18] 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*, 2016, pp. 1–6.

[19] C.-c. Yeh, P.-c. Hsu, J.-c. Chou, H.-y. Lee, and L.-s. Lee, "Rhythm-flexible voice conversion without parallel data using cycle-gan over phoneme posteriorgram sequences," in *IEEE SLT*. IEEE, 2018, pp. 274–281.

[20] 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*, 2019, pp. 6790–6794.

[21] Y. Zhou, X. Tian, R. K. Das, and H. Li, "Many-to-many cross-lingual voice conversion with a jointly trained speaker embedding network," in *IEEE APSIPA*, 2019.

[22] 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 *IEEE ASRU*, 2019, pp. 160–167.

[23] Y. Zhou, X. Tian, and H. Li, "Multi-Task WaveRNN With an integrated architecture for cross-lingual voice conversion," *IEEE Signal Processing Letters*, vol. 27, pp. 1310–1314, 2020.

[24] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[25] T. Hori, R. Astudillo, T. Hayashi, Y. Zhang, S. Watanabe, and J. Le Roux, "Cycle-consistency training for end-to-end speech recognition," in *IEEE ICASSP*, 2019, pp. 6271–6275.

[26] Y. Xia, D. He, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma, "Dual learning for machine translation," *arXiv:1611.00179*, 2016.

[27] R. Liu, B. Sisman, G. l. Gao, and H. Li, "Expressive tts training with frame and style reconstruction loss," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2021.

[28] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv:1711.11293*, 2017.

[29] H. Du, X. Tian, L. Xie, and H. Li, "Optimizing voice conversion network with cycle consistency loss of speaker identity," *arXiv:2011.08548*, 2020.

[30] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *INTERSPEECH*, 2019, pp. 1526–1530.

[31] "aidataTang 200zh, a free Chinese Mandarin speech corpus." Beijing DataTang Technology Co., Ltd.

[32] B. Hui, D. Jiayu, N. Xingyu, W. Bengu, and Z. Hao, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Oriental COCOSDA*, 2017, pp. 1–5.

[33] "MAGICDATA Mandarin Chinese read speech corpus." Magic Data Technology Co., Ltd, 2019. [Online]. Available: http://www.imagicdatatech.com/index.php/home/dataopensource/data_info/id/101

[34] "PrimeWords Chinese corpus set 1." Primewords Information Technology Co., Ltd, 2018. [Online]. Available: https://www.primewords.cn

[35] "ST-CMDS-20170001 1 free ST Chinese Mandarin corpus." Surfingtech.

[36] D. Wang and X. Zhang, "THCHS-30: A free chinese speech corpus," *arXiv:1512.01882*, 2015.

[37] "The DARPA TIMIT acoustic-phonetic continuous speech corpus." Linguistic Data Consortium and others, 1990, pp. 1–1.

[38] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *IEEE ICASSP*, 2018, pp. 4879–4883.

[39] 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.

[40] M. Wester and H. Liang, "The EMIME Mandarin bilingual database," The University of Edinburgh, Tech. Rep. EDI-INF-RR-1396, 2011.

[41] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *ICML*, 2018, pp. 2415–2424.

[42] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "WAVE-TACOTRON: Spectrogram-free end-to-end text-to-speech synthesis," *arXiv:2011.03568*, 2020.