

Latent linguistic embedding for cross-lingual text-to-speech and voice conversion

Hieu-Thi Luong^{1,2}, Junichi Yamagishi^{1,2}

¹National Institute of Informatics, Tokyo, Japan

²SOKENDAI (The Graduate University for Advanced Studies), Kanagawa, Japan

{luonghieuthi, jyamagis}@nii.ac.jp

Abstract

As the recently proposed voice cloning system, NAUTILUS, is capable of cloning unseen voices using untranscribed speech, we investigate the feasibility of using it to develop a unified cross-lingual TTS/VC system. Cross-lingual speech generation is the scenario in which speech utterances are generated with the voices of target speakers in a language not spoken by them originally. This type of system is not simply cloning the voice of the target speaker, but essentially creating a new voice that can be considered better than the original under a specific framing. By using a well-trained English latent linguistic embedding to create a cross-lingual TTS and VC system for several German, Finnish, and Mandarin speakers included in the Voice Conversion Challenge 2020, we show that our method not only creates cross-lingual VC with high speaker similarity but also can be seamlessly used for cross-lingual TTS without having to perform any extra steps. However, the subjective evaluations of perceived naturalness seemed to vary between target speakers, which is one aspect for future improvement.

Index Terms: cross-lingual, text-to-speech, voice conversion, speaker adaptation, unsupervised adaptation

1. Introduction

Text-to-speech (TTS) and Voice Conversion (VC) are technologies that generate speech with a target voice from a text or a source utterance input. Recent studies have shown that we can clone new voices for these speech generation systems using a small amount of speech data either with or without transcription [1] while maintaining a high quality and speaker similarity to the target speakers [2, 3]. Voice cloning can be regarded as a classical scenario of speech generation in which the objective is to generate utterances that are as similar to the natural speech of the target speaker as possible. This framing is straightforward and easy to evaluate, but it ignores the potential of the speech synthesis system which is the ability to create something new and/or better than the original under certain terms and conditions. Dysarthric speech reconstruction [4, 5] and accent-reduction voice conversion [6, 7] are two examples of scenarios in which we want to create a speech generation model with voices that are better than the originals while maintaining a certain level of speaker individuality. Cross-lingual speech generation [8, 9], the scenario in which speech is generated in a language not spoken by the target speaker, is another one and it is also the main topic of this paper.

In the case of cross-lingual TTS, we could adapt an HMM-based TTS model to a target speaker whose data is not in the target language by establishing a phoneme mapping between the two languages [8, 10, 11], training a bilingual state-sharing HMM model [12], or factorizing the language and speaker

components as transformation functions [13]. Similar principles have been applied to neural TTS systems. For example, several works have proposed training a multi-speaker multi-lingual neural TTS model capable of cross-lingual speech generation by utilizing factorizing speaker and language components [14, 15, 16]. Unsupervised speaker adaptation methods, such as speaker-adaptive TTS model conditioning on neural speaker embedding [17], have also shown promising results for the cross-lingual scenario. Cross-lingual TTS is also the foundation for more interesting applications such as code-mixing speech synthesis [18, 19].

In the case of cross-lingual VC, the systems are generally based on non-parallel VC systems, which are developed using the non-parallel utterances of source and target speakers. Specifically, Phonetic PosteriorGram (PPG) based models are often used for the cross-lingual scenario [9]. Even though a monolingual PPG trained on the target language is good enough for cross-lingual VC, it was reported that a bilingual PPG [20, 19] or mixed-lingual PPG [21] can significantly improve the performance. Non-parallel VC systems based on a Variational Autoencoder (VAE) [22] or Generative Adversarial Network (GAN) [23, 24] are also applicable to the cross-lingual scenario. Cross-lingual VC was also the main topic of the Voice Conversion Challenge 2020 (VCC2020) which is the basis of the evaluations we present in this paper.

In this study, we extend the versatile voice cloning method of the NAUTILUS system [1] to the cross-lingual speech generation scenario. While we follow the VCC2020 theme, our primary focus is the development of a unified framework for both cross-lingual TTS and VC. The proposed system is expected to generate speech with target voices in a language that is not spoken by them by using either the TTS or VC input interface. Moreover, the performances when switching between the two modes should be relatively consistent. In Section 2 of this paper, we describe our procedure to create a unified cross-lingual TTS and VC system for a target speaker. Section 3 provides details about the experimental setup, particularly the data conditions and in Section 4 we discuss the results of the VCC2020 and those of our own listening test. We conclude in Section 5 with a brief summary and mention of future works.

2. Cross-lingual TTS and VC system with latent linguistic embedding

Our methodology for cross-lingual speaker adaptation is based on the unsupervised voice cloning strategy using Latent Linguistic Embedding (LLE) [1]. Previously, we showed that this type of cross-lingual VC system can deal with several sub-scenarios thanks to the imbalance of data demands in each step [25]. In this work, we focus on the cross-language speaker

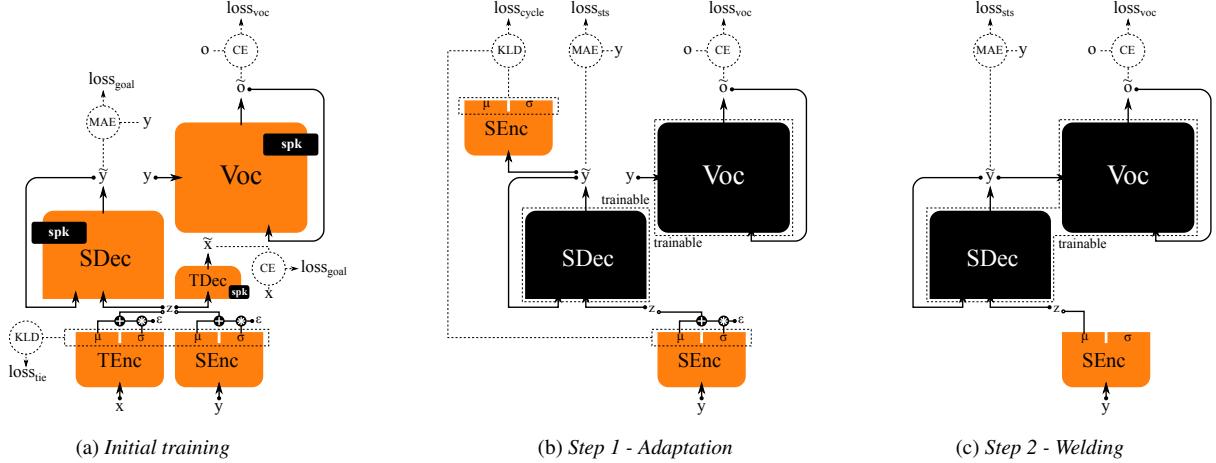


Figure 1: Steps of the intra-language/cross-language speaker adaptation procedure for the proposed unified TTS/VC system. The proposed system comprises of a text encoder ($T\text{Enc}$), a speech encoder ($S\text{Enc}$), a text decoder ($T\text{Dec}$), a speech decoder ($S\text{Dec}$), and a neural vocoder (Voc). \mathbf{x} is phoneme, \mathbf{y} is acoustic, \mathbf{o} is waveform, and \mathbf{z} is latent linguistic embedding. The $\text{loss}_{\text{goal}}$ term is a placeholder for either loss_{sts} , loss_{sts} , loss_{stt} , or loss_{stt} depending on the encoder/decoder combination.

adaptation scenario of TTS and VC to align with the theme of VCC2020. The steps to develop a unified cross-lingual TTS/VC are summarized in this section, while further details can be found in the original paper [1].

2.1. Initial training

This step focuses on training a robust and speaker-disentangled LLE in a target language as well as initializing all text/speech encoders/decoders using a large-scale multi-speaker corpus (Section III-A in [1]). We achieve this by jointly training the modules of the text-speech multimodal system using joint-goal and tied-layer objectives as follows:

$$\begin{aligned} \text{loss}_{\text{train}} &= \text{loss}_{\text{goals}} + \beta \text{loss}_{\text{tie}} \\ &= \text{loss}_{\text{sts}} + \alpha_{\text{sts}} \text{loss}_{\text{sts}} + \alpha_{\text{stt}} \text{loss}_{\text{stt}} \\ &\quad + \beta \text{loss}_{\text{tie}}. \end{aligned} \quad (1)$$

Given the VAE-like structure of the encoders, we use symmetrized Kullback-Leibler divergence between the encoder outputs as a tied-layer loss:

$$\begin{aligned} \text{loss}_{\text{tie}} &= \frac{1}{2} L_{\text{KLD}}(T\text{Enc}(\mathbf{x}), S\text{Enc}(\mathbf{y})) \\ &\quad + \frac{1}{2} L_{\text{KLD}}(S\text{Enc}(\mathbf{y}), T\text{Enc}(\mathbf{x})). \end{aligned} \quad (2)$$

A multi-speaker WaveNet vocoder is separately trained using the same corpus in this step. It is used as the initial model for speaker adaptation in later steps.

$$\text{loss}'_{\text{train}} = \text{loss}_{\text{voc}}, \quad (3)$$

Figure 1a shows the conceptual actions performed in the initial training step. The same initial model obtained in this step will be reused for all target speakers.

2.2. Step 1 - Adaptation

As cross-lingual adaptation can be considered a special case of unsupervised speaker adaptation in which the transcribed speech of target speakers is unobtainable, as it is in a foreign/unseen language, we can use the unsupervised voice

cloning strategy to tackle the cross-lingual scenario (Section III-B1 in [1]). Specifically, we adapt the speech decoder to the speech data of the target speaker as illustrated in Fig. 1b:

$$\text{loss}_{\text{adapt}} = \text{loss}_{\text{sts}} + \beta \text{loss}_{\text{cycle}}, \quad (4)$$

where $\text{loss}_{\text{cycle}}$ is the KL divergence between the LLE distributions extracted from natural and converted acoustic features. As LLE is a latent representation of sound, it is expected to be generalizable to other languages. The neural vocoder is also adapted to the target speaker, as

$$\text{loss}'_{\text{adapt}} = \text{loss}_{\text{voc}}. \quad (5)$$

The fact that the initial model was never trained on data in the languages other than English might negatively affect the performance of the cross-language speaker adaptation. However, we decided not to use additional data of the languages spoken by the target speakers in order to keep data demand low and simply test the generalization ability of the LLE.

2.3. Step 2 - Welding

Although tuning the speech decoder and neural vocoder separately is sufficient, we perform an additional step in which they are jointly tuned in order to increase their compatibility, as shown in Fig. 1c. The optimizing loss is formulated as follows:

$$\text{loss}_{\text{weld}} = \text{loss}_{\text{sts}} + \gamma \text{loss}_{\text{voc}}. \quad (6)$$

The loss_{sts} is included to maintain the acoustic space for the autoregressive speech decoder. Several training tactics are applied in this step to enable the model to learn fine-grained details while avoiding overfitting [1].

2.4. Step 3 - Inference

The adapted speech decoder and neural vocoder can be used along with either the text or speech encoder to form a complete TTS or VC system. For the standard intra-language speaker adaptation scenario, the NAUTILUS system has demonstrated a highly consistent performance between the TTS and VC inference [1]. In this work, we check if the same statement holds true in the cross-lingual scenario.

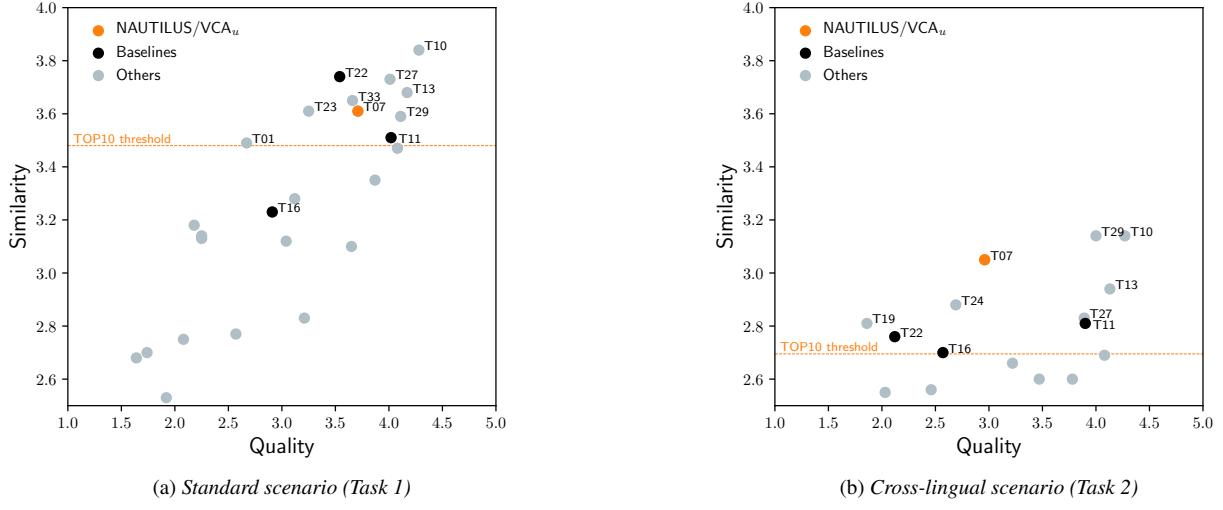


Figure 2: Subjective evaluations of quality and speaker similarity at VCC2020. Listeners are native English speakers.

3. Experiments

3.1. Model configuration and LLE training

The same NAUTILUS system used to evaluate scenario B in the original paper [1] was reused for the cross-language speaker adaptation experiments. Each module of the neural model consisted of several one-dimensional convolution layers to capture temporal context. Phonemes were used as the text representation and an 80-dimensional mel-spectrograms were used as the acoustic representation. The current system is not yet an end-to-end (E2E) system, as it requires the explicit phoneme duration information when generating speech from text input. This setup allows us to have the same duration condition between the TTS and VC system. The initial model used for scenario B (and reused in this paper) was first trained on 24 kHz English transcribed speech of LibriTTS [26], which has diverse linguistic content, and then on VCTK corpus [27], which recorded in a more controlled environment. This initial model trained on English data was adapted to target speakers whose data is in languages other than English. Additional data of languages spoken by the target speakers would likely improve the performance [16, 20], but we leave such investigation for future works.

3.2. Target speakers for evaluations

We tested the performance of the proposed cross-lingual speech generation system on target speakers from VCC2020, which consists of four native English speakers (Task 1) and six speakers who spoke languages other than English (Task 2). While the focus of the work is the cross-lingual scenario, we perform both tasks to establish a reliable baseline with English target speakers. In summary, the target speakers included four English, two Finnish, two German, and two Mandarin speakers with an equal number of male and female speakers in each language. Each speaker had 70 utterances (around five minutes) of untranscribed speech available for adaptation. We tested our cross-lingual TTS system on the same target speakers and compared its performance with our cross-lingual VC system to evaluate the consistency between the two. Even though the challenge provided speech data of source speakers, we decided not to use it for either training or adaptation.

4. Evaluations

4.1. Voice Conversion Challenge 2020

Figure 2a and 2b shows the subjective quality and speaker similarity results, judged by native English speakers, of the systems submitted to VCC2020. We recreated the figures based on results provided by the organizers to focus on the information relevant to our works¹. For the standard intra-language task, our system, T07, achieved moderate performance and ranked among the top ten systems with the highest speaker similarity (Fig. 2a). Our system had a higher speaker similarity but lower quality than T11, which was the best system (N10) of VCC2018 [3]. This is similar to the results presented in the original paper of the NAUTILUS system [1]. Compared with other baselines, our system had a higher quality but lower speaker similarity than T22 [29], while T16 [30] did not make it into the top ten ranking in the intra-language scenario. In summary, the subjective results of Task 1 confirm the competitive performance of our system when used as a VC system.

For the cross-lingual task, the setup for the listening test was trickier due to the language mismatch between the generated and natural utterances. Based on information provided by the organizers, for the speaker similarity evaluation, listeners were presented with either a native language utterance or the English utterance of the target, the data of which was not provided to the challenge participants. The results shown in Fig. 2b are the averages over both cases. In general, the cross-lingual scenario seems to have had a lower speaker similarity than the standard scenario, which is not surprising. Our system (T07) was ranked 3rd in the similarity metric and higher than all three baselines. However, it received relative low quality scores compared to other systems. Interestingly, if we only consider the speaker similarity results between converted utterances and the native language utterance of the target speakers, our system ranked 1st in the case of Finnish speakers, 4th in the case of German speakers, and 4th in the case of Mandarin speakers. This result demonstrates the feasibility of using the NAUTILUS system for cross-language speaker adaptation, as it achieved very high speaker similarity without having to pre-train on data of languages spoken by the target speakers.

¹See the paper written by the organizers for detailed results [28]

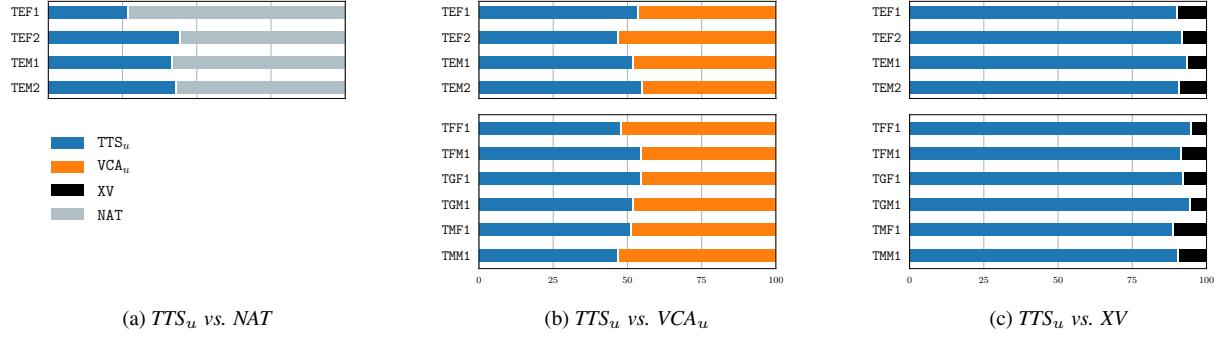


Figure 3: Subjective quality evaluations. Each speaker/system pair was judged 300 times by native Japanese speakers.

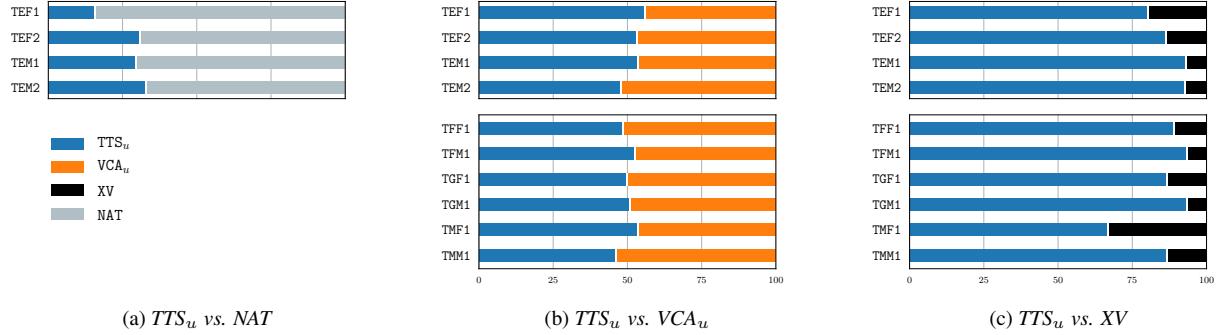


Figure 4: Subjective similarity evaluations. Each speaker/system pair was judged 300 times by native Japanese speakers.

4.2. A unified cross-lingual system of TTS and VC

The strength of the NAUTILUS system is its ability to switch between TTS and VC modes. We evaluate its performance as a cross-lingual TTS system in this section. Specifically, we compared our cross-lingual TTS system (TTS_u) with the cross-lingual VC system (VCA_u), a simple TTS baseline based on x-vector (XV) [17], and the natural utterances of the target speakers (NAT)². The adapted models used in the previous section were reused here for TTS_u and VCA_u , while XV³ was the same system used in the original paper [1]. We conducted our own listening test by asking 170 participants, who did one to ten tests, to pick which of the two samples had better quality or similarity. Each test consisted of 24 quality and 24 similarity questions covering all speaker and system pairs. For the TTS_u /NAT pair, we only evaluated the English target speakers, as comparing an English and non-English utterances would be uninformative. The reference utterances used in the similarity question were in the native languages of the target speakers.

The quality and speaker similarity results are presented in Figures 3 and 4, respectively. As expected TTS_u was not comparable with the natural utterances, but neither was it totally dominated by NAT, which is an encouraging result. Interestingly, the listeners seemed to pick the natural samples more when presented with a reference utterance as shown in the similarity results. Between TTS_u and VCA_u , there were no clear winners across all speakers which suggests the highly consistent performance between the two modes of the proposed system even in the cross-lingual scenario. Last but not least, compared with the cross-lingual TTS baseline, XV, our system dominated

the results across all speakers, which validates our cross-lingual methodology on the TTS side. If we look at individual speakers, our system seems to have had noticeably worse results in the case of TEF1 and TMF1 in the similarity metric. This suggests that the performance of the proposed method may not be consistent between speakers.

5. Conclusions

The cross-lingual experiments presented in this paper have once again demonstrated the versatility of the proposed NAUTILUS system. Even though the initial LLE is only trained on English speech data, the proposed system can generalize to other languages spoken by the target speakers and generate utterances with high speaker similarity. This significantly reduces the data demands on the cross-lingual system and allows it to adapt to target speakers who only speak low-resource languages. However, as cross-lingual systems are expected to benefit from multilingual data [17, 20], our future work will focus on taking advantage of either transcribed or untranscribed speech from large-scale multilingual corpora. Moreover, although the VCC2020 has taken the first step in shifting the focus of the research community to developing speech generation system that are better and different from the original voice, concrete guidelines on how to evaluate such systems have not yet been established. Additional research on this new and exciting topic is required.

6. Acknowledgements

This work was partially supported by a JST CREST Grant (JP-MJCR18A6, VoicePersonae project), Japan, and MEXT KAKENHI Grants (16H06302, 17H04687, 18H04120, 18H04112, 18KT0051), Japan.

²Samples are available at <https://nii-yamagishilab.github.io/sample-cross-lingual-tts-vc/>

³XV system consists of the pretrained *libritts.tacotron2.v1* and *libritts.wavenet.mol.v1* models based on ESPNet toolkit [31]

7. References

- [1] H.-T. Luong and J. Yamagishi, “NAUTILUS: a versatile voice cloning system,” *arXiv preprint arXiv:2005.11004*, 2020.
- [2] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, “Neural voice cloning with a few samples,” in *Proc. NIPS*, 2018, pp. 10 040–10 050.
- [3] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, “Wavenet vocoder with limited training data for voice conversion,” in *Proc. INTERSPEECH*, 2018, pp. 1983–1987.
- [4] A. B. Kain, J.-P. Hosom, X. Niu, J. P. Van Santen, M. Fried-Oken, and J. Staehely, “Improving the intelligibility of dysarthric speech,” *Speech communication*, vol. 49, no. 9, pp. 743–759, 2007.
- [5] D. Wang, J. Yu, X. Wu, S. Liu, L. Sun, X. Liu, and H. Meng, “End-to-end voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction,” in *Proc. ICASSP*, 2020, pp. 7744–7748.
- [6] S. Aryal and R. Gutierrez-Osuna, “Can voice conversion be used to reduce non-native accents?” in *Proc. ICASSP*, 2014, pp. 7879–7883.
- [7] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, “Accent conversion using phonetic posteriograms,” in *Proc. ICASSP*, 2018, pp. 5314–5318.
- [8] Y.-J. Wu, S. King, and K. Tokuda, “Cross-lingual speaker adaptation for hmm-based speech synthesis,” in *Proc. ISCSLP*, 2008, pp. 9–12.
- [9] 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.
- [10] H. Liang, J. Dines, and L. Saheer, “A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for hmm-based speech synthesis,” in *Proc. ICASSP*. IEEE, 2010, pp. 4598–4601.
- [11] K. Oura, K. Tokuda, J. Yamagishi, S. King, and M. Wester, “Unsupervised cross-lingual speaker adaptation for hmm-based speech synthesis,” in *Proc. ICASSP*, 2010, pp. 4594–4597.
- [12] Y. Qian, H. Liang, and F. K. Soong, “A cross-language state sharing and mapping approach to bilingual (mandarin–english) tts,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1231–1239, 2009.
- [13] H. Zen, N. Braunschweiler, S. Buchholz, M. J. Gales, K. Knill, S. Krstulovic, and J. Latorre, “Statistical parametric speech synthesis based on speaker and language factorization,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 6, pp. 1713–1724, 2012.
- [14] B. Li and H. Zen, “Multi-language multi-speaker acoustic modeling for lstm-rnn based statistical parametric speech synthesis,” *Proc. INTERSPEECH*, pp. 2468–2472, 2016.
- [15] Y. Fan, Y. Qian, F. K. Soong, and L. He, “Speaker and language factorization in dnn-based tts synthesis,” in *Proc. ICASSP*, 2016, pp. 5540–5544.
- [16] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, “Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning,” *Proc. INTERSPEECH*, pp. 2080–2084, 2019.
- [17] M. Chen, M. Chen, S. Liang, J. Ma, L. Chen, S. Wang, and J. Xiao, “Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding.” in *Proc. INTERSPEECH*, 2019, pp. 2105–2109.
- [18] S. K. Rallabandi and A. W. Black, “Variational attention using articulatory priors for generating code mixed speech using monolingual corpora.” in *Proc. INTERSPEECH*, 2019, pp. 3735–3739.
- [19] Y. Cao, S. Liu, X. Wu, S. Kang, P. Liu, Z. Wu, X. Liu, D. Su, D. Yu, and H. Meng, “Code-switched speech synthesis using bilingual phonetic posteriorgram with only monolingual corpora,” in *Proc. ICASSP*, 2020, pp. 7619–7623.
- [20] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, “Cross-lingual voice conversion with bilingual phonetic posteriogram and average modeling,” in *Proc. ICASSP*, 2019, pp. 6790–6794.
- [21] Y. Zhou, X. Tian, E. Yilmaz, 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.
- [22] S. H. Mohammadi and T. Kim, “Investigation of using disentangled and interpretable representations for one-shot cross-lingual voice conversion,” *Proc. INTERSPEECH*, pp. 2833–2837, 2018.
- [23] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, “Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks,” in *Proc. SLT*, 2018, pp. 266–273.
- [24] B. Sisman, M. Zhang, M. Dong, and H. Li, “On the study of generative adversarial networks for cross-lingual voice conversion,” in *Proc. ASRU*, 2019, pp. 144–151.
- [25] H.-T. Luong and J. Yamagishi, “Bootstrapping non-parallel voice conversion from speaker-adaptive text-to-speech,” in *Proc. ASRU*, 2019, pp. 200–207.
- [26] 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 *Proc. INTERSPEECH*, 2019, pp. 1526–1530.
- [27] C. Veaux, J. Yamagishi, and K. MacDonald, “CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit,” 2017, <http://dx.doi.org/10.7488/ds/1994>.
- [28] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinunen, 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] 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 preprint arXiv:2010.02434*, 2020.
- [30] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, “Non-parallel voice conversion with cyclic variational autoencoder,” *Proc. INTERSPEECH*, pp. 674–678, 2019.
- [31] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.-E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, “Espnet: End-to-end speech processing toolkit,” *Proc. INTERSPEECH*, pp. 2207–2211, 2018.