Computer Speech and Language 77 (2023) 101427

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Cross-lingual multi-speaker speech synthesis with limited bilingual training data

Zexin Cai [a], Yaogen Yang [b], Ming Li [a,b,*]

[a] Department of Electrical and Computer Engineering, Duke University, Durham, NC, United States
[b] Data Science Research Center, Duke Kunshan University, Kunshan, China

ARTICLE INFO

Keywords:
Text-to-speech
Multi-speaker speech synthesis
Multilingual speech synthesis
Cross-lingual speech synthesis
Voice conversion
Code-switching

ABSTRACT

Modeling voices for multiple speakers and multiple languages with one speech synthesis system has been a challenge for a long time, especially in low-resource cases. This paper presents two approaches to achieve cross-lingual multi-speaker text-to-speech (TTS) and code-switching synthesis under two training scenarios: (1) cross-lingual synthesis with sufficient data, (2) cross-lingual synthesis with limited data per speaker. Accordingly, a novel TTS synthesis model and a non-autoregressive multi-speaker voice conversion model are proposed. The TTS model designed for sufficient-data cases has a Tacotron-based structure that uses shared phonemic representations associated with numeric language ID codes. As for the data-limited scenario, we adopt a framework cascading several speech modules to achieve our goal. In particular, we proposed a non-autoregressive many-to-many voice conversion module to address multi-speaker synthesis for data-insufficient cases. Experimental results on speaker similarity show that our proposed voice conversion module can maintain the voice characteristics well in data-limited cases. Both approaches use limited bilingual data and demonstrate impressive performance in cross-lingual synthesis, which can deliver fluent foreign speech and even code-switching speech for monolingual speakers.

1. Introduction

In the past few years, the end-to-end text-to-speech (TTS), which consists of an encoder–decoder-based text-to-spectrogram network and a neural vocoder, has allowed machines to synthesize high-fidelity speech that is as natural as human speech (Tan et al., 2021a; Shen et al., 2018). This type of TTS framework outperforms traditional frameworks like concatenative speech synthesis (Hunt and Black, 1996) and statistical parametric speech synthesis (SPSS) (Zen et al., 2009). It quickly became the state-of-the-art framework for speech synthesis and is widely applied in various TTS applications (e.g., audiobook readers, virtual assistants, navigation systems) in our daily lives.

Nonetheless, this kind of model, like vanilla Tacotron2 (Shen et al., 2018) and Fastspeech (Ren et al., 2019, 2021), keeps a certain level of limitations in controllability regarding latent speech attributes when it is proposed. It renders the speech attributes and learns to model those attributes implicitly during training. In this case, the model is not robust enough to synthesize speech with specific target characteristics, such as emotion, timbre, and prosody. Then researchers propose novel extensions on the end-to-end framework to improve the model's robustness in controlling speech attributes. For example, Wang et al. model the latent speech attributes by incorporating the unsupervised global style tokens (GSTs) to the Tacotron2 (Wang et al., 2018). This allows the model to control speaking speed and clone speaking styles via GSTs. As for speaker identity, some research works extend the Tacotron2

[*] Corresponding author at: Data Science Research Center, Duke Kunshan University, Kunshan, China.
E-mail addresses: zexin.cai@duke.edu (Z. Cai), ming.li369@duke.edu (M. Li).

https://doi.org/10.1016/j.csl.2022.101427
Received 27 September 2021; Received in revised form 19 June 2022; Accepted 2 July 2022
Available online 14 July 2022

for multi-speaker TTS with conditioned features extracted from a speaker verification system to achieve voice cloning (Jia et al., 2018; Cai et al., 2020).

On the other hand, language is an important attribute for multilingual speech synthesis. As bilinguals and polyglots are commonly seen in today's world, the speech communication scenario becomes more complicated (Titus et al., 2020; European Commission, 2012). It is essential for speech analysis tools, including speech recognition and speech synthesis, to adapt to this change to maintain their current performance (Rallabandi and Black, 2017). The challenge is that languages generally have different grapheme sets and pronunciations. This challenge motivates researchers to investigate shared representations between languages for speech signal analysis (Association, 1999; Gales et al., 2015; Li et al., 2019). Even with appropriate representations for multiple languages, however, the model architecture needs to be upgraded to achieve the multilingual processing of most speech analysis systems (Li et al., 2019). There are existing studies of multilingual synthesis and cross-lingual synthesis for text-to-speech based on classical statistical parametric speech synthesis (SPSS) (Li and Zen, 2016; Sitaram et al., 2016; Ming et al., 2017). Nevertheless, the synthesis performance is restricted by the relatively complex pipeline and the vocoder in terms of SPSS approaches (Tan et al., 2021a). As the end-to-end TTS models can generate speech with higher fidelity than classical methods, extensions on the end-to-end TTS frameworks are also explored for multilingual modeling (Lee et al., 2018; Zhang et al., 2019; Zhou et al., 2020; Chen et al., 2019). As a special case in multilingual synthesis, the cross-lingual synthesis, where we can generate speech with foreign text for monolingual speakers, is more challenging, especially in low-resource cases. Regarding that case, Zhang et al. achieve high-quality cross-lingual synthesis among three languages in a sufficient-data manner (Zhang et al., 2019). Liu et al. investigate cross-lingual synthesis with limited data for each speaker, while the synthesized speech has moderate quality due to the data sparsity issue (Liu and Mak, 2020).

This paper aims to achieve cross-lingual multi-speaker TTS in English and Mandarin while using limited bilingual data. Since collecting huge data for multi-speaker cross-lingual synthesis is costly, it is essential to address the cross-lingual synthesis under low-resource data scenarios. Therefore, we also investigate solutions to such scenarios in our work. Two synthesis frameworks are proposed for two different scenarios: the data-sufficient and the low-resource scenario. In the data-sufficient scenario, we propose a Tacotron-based model conditioned on speaker embedding and numeric language ID codes. Similar pronunciations between languages are related through shared phoneme input. The proposed model can generate high-fidelity speech for all speakers in their native language. In addition, we investigate cross-lingual synthesis with the same model by involving a bilingual TTS dataset. Results show that linguistic knowledge can be transferred from a bilingual speaker to monolingual speakers, enabling us to generate fluent, high-fidelity, and intelligible speech in both Mandarin and English using monolingual speakers' voices. In the data-limited case, the training dataset contains hundreds of monolingual speakers, while the total recording of each speaker is less than half an hour. As the Tacotron-based TTS model is ineffective in this scenario, we adopt a series of speech modules to accomplish the cross-lingual synthesis. Specifically, we incorporate a linguistic feature extractor, a speaker representation extractor, a multi-speaker voice conversion system, and a neural vocoder to obtain the target speech. In particular, we propose a non-autoregressive network for the multi-speaker voice conversion module. The adversarial speaker classifier (Meng et al., 2019) and the speaker embedding consistency loss (Cai et al., 2020) are employed in the conversion network to improve the speaker similarity. We conduct objective evaluation and subjective evaluation on the synthesis performance. Results show that the VC system can generate fluent cross-lingual speech with satisfactory speaker similarity. Furthermore, both systems under two scenarios can tackle code-switching synthesis.[1] The contributions of our paper include:

- We investigate multi-speaker cross-lingual speech synthesis in two multilingual data setups using limited bilingual data.
- For the data-insufficient scenario, we adopt a synthesis framework that cascades a series of speech modules. Within the framework, we propose a parallel non-autoregressive model for voice conversion.

This paper is organized as follows. Section 2 introduces the related work regarding multilingual multi-speaker TTS and voice conversion. Section 3 presents our proposed model for data-sufficient scenario while Section 4 presents the speech modules we employ for the data-insufficient scenario. Experimental details and results are presented in Section 5. Finally, we hold a discussion in Section 6, and our paper is concluded in Section 7.

2. Related works

2.1. Multilingual and cross-lingual TTS

Developing a Multilingual Multi-speaker (MLMS) TTS model can relieve the efforts of training multiple TTS models used for several voices with different languages. While the voice can be controlled by a text-independent speaker embedding in a multi-speaker TTS system (Jia et al., 2018; Cooper et al., 2020), TTS regarding multiple languages is more complicated due to different grapheme representations across languages.

However, similar pronunciations between different languages can help reduce the gap of cross-lingual text-to-speech. Linguistic representation across languages has been investigated for years in MLMS TTS. Li et al. propose an MLMS TTS approach based on conventional statistical parametric speech synthesis (SPSS) (Li and Zen, 2016). They use the international pronunciation Alphabet (IPA) (Association, 1999) as the input representation and applied cluster adaptive language networks for generating language-dependent linguistic feature vectors, followed by speaker-dependent output layers for different voices. Ming et al. present a

[1] Audio samples are available online for listening: https://caizexin.github.io/mlms-syn-samples/index.html.

light-weight bilingual synthesis system that adopts concatenated vectors in the linguistic-feature level to manage two languages in one model (Ming et al., 2017).

More recently, Li et al. use a sequence of Unicode bytes to represent the text in multilingual speech recognition and multilingual speech synthesis (Li et al., 2019). This representation, called Bytes, allows speech recognition models and speech synthesis models to achieve multilingual processing. The use of Bytes and other representations in multilingual TTS are conducted and evaluated thoroughly later (Zhang et al., 2019), which shows that using phoneme units as the input representation is better than using Bytes. In addition, with large-scale training data (more than 500 h), the MLMS model can achieve cross-lingual synthesis with a high naturalness rate (Zhang et al., 2019). The shared phoneme input is one of the keys to cross-lingual synthesis. By using shared phoneme input neural network-based synthesis systems, similar pronunciations across languages result in close linguistic embedding vectors (Lee et al., 2018). We also propose a TTS framework using shared phonetic representations for cross-lingual multi-speaker speech synthesis, and it is archived at Cai et al. (2020b).[2] Compared to the MLMS system proposed by Zhang et al. (2019), we used the language embedding for a different purpose. The language embedding in this paper specifies the phoneme-level pronunciation for different languages instead of controlling the language-specific accent at the utterance level. In terms of the dataset setup, we reduce the data usage for achieving cross-lingual synthesis and code-switching synthesis by incorporating a limited bilingual dataset.

Likewise, there are more research works in this field. Liu et al. also use shared phoneme representation and extend the Tacotron2 by incorporating conditional embeddings for MLMS TTS (Liu and Mak, 2020), which has a similar structure as our proposed model. However, we have the language-dependent Tacotron encoder designed for allowing the TTS model to synthesize code-switching text. Zhou et al. present a novel method to merge context information between languages by adopting word embeddings from a pre-trained language model. Nevertheless, the cross-lingual synthesized speech has moderate quality, as shown in the figures from Zhou et al. (2020). Fu et al. present a code-switching speech synthesis system based on a language-dependent style token (Fu et al., 2020). It applies a dynamic soft windowing mechanism on the decoder module to implicitly improve the consistency in bilingual synthesis, which improves the performance concerning naturalness and intelligibility. The experiments conducted in Fu et al. (2020) mainly focus on the bilingual speaker, while we also look into the code-switching synthesis performance for the monolingual speakers in this paper.

On the other hand, low-resource synthesis is a common issue in TTS due to the difficulty of collecting data. In this case, there are studies investigating the MLMS synthesis for resource-poor languages recently. Staib et al. investigate phonological features that could adapt to untrained languages with zero-shot adaptation (Staib et al., 2020). Similarly, Korte et al. look into how different strategies work for resource-poor language synthesis with data from resource-rich languages (de Korte et al., 2020). In our paper, we pay attention to the low-resource scenario when each speaker contains limited data for training. Xue et al. also investigate cross-lingual synthesis systems in a data-limited case where there are only 500 utterances for each speaker in training (Xue et al., 2019). They build a female cross-lingual system with only monolingual data and show that it is still better to use mixed-lingual data. Our work, however, investigates the performance of systems trained with another dataset composition. Moreover, speaker consistency is an important criterion for code-switching synthesis under low-resource scenarios. Multiple voices may appear in the same sentence when synthesizing mixed-lingual text for one target speaker in limited-data scenarios (Xue et al., 2019; Xin et al., 2021). There are multiple ways to address this issue. Xin et al. proposed a method that disentangles language and speaker representations by minimizing mutual information (Xin et al., 2021), which greatly improves speaker consistency. However, the voice cloning of unseen speakers is not good enough due to the limited number of speakers in TTS training data. Thus we adopt another method from our previous works in this paper to improve the speaker consistency, where an external speaker encoder is incorporated to construct speaker embedding consistency loss during TTS training (Cai et al., 2020).

2.2. Voice Conversion

Voice Conversion (VC) is a speech technique that changes the voice characteristics of an audio signal to the desired voice while keeping the linguistic contents unchanged. Generally, the source speaker refers to the original voice of an utterance, and the target speaker is the expected voice the system converts to. According to whether the source speaker and the target speaker speak the same language, VC can be divided into intra-lingual VC and cross-lingual VC. For intra-lingual VC, variational auto-encoder (VAE)-based methods and generative adversarial network (GAN)-based approaches are widely used (Kameoka et al., 2019; Tobing et al., 2019; Kameoka et al., 2018; Lee et al., 2020).

The cross-lingual VC is nonparallel in nature. Since the source and target speakers speak different languages, the speech utterances are inherently different in content. To achieve cross-lingual VC, we are supposed to disentangle speaker characteristics and the content of the source-speech data in the source language and then replace the speaker characteristics with those from the target speaker regardless of what languages the target speaker speaks (Yi et al., 2020). Vector quantization (VQ)-based method is used for cross-lingual VC between Japanese and English (Abe et al., 1990). However, this approach is not robust enough in preserving the speakers' identity, where the feature space of the converted envelope is limited to a discrete set of envelopes. Ramani et al. propose a GMM-based cross-lingual VC to generate a polyglot speech corpus (Ramani et al., 2014). For the GMM-based approach, phonemes from the source language are accordingly replaced by acoustically similar phonemes from the target language under GMM-based VC. Later, Phonetic PosteriorGram (PPG) based methods (Zheng et al., 2016; Sun et al., 2016; Zhou et al., 2019; Zhao et al., 2021),

[2] Our preliminary methods and experimental results are shared in our archived paper https://arxiv.org/abs/2005.10441.

which take advantage of the linguistic information from a large amount of speech data, also achieve high performance in cross-lingual VC. The PPG obtained from a speaker-independent automatic speech recognition (ASR) system can be regarded as a bridge feature across boundaries between speakers and language (Sun et al., 2016). Those aforementioned methods focus on one-to-one cross-lingual VC and use the conventional vocoder WORLD (Morise et al., 2016) to reconstruct the waveform from the predicted spectrum, which leads to relatively lower naturalness and speaker similarity. In our paper, we aim to achieve many-to-many voice conversion such that the model can be used for multi-speaker synthesis. Similar to text-to-speech, speaker verification models have been incorporated in VC such that the VC system can generalize to unseen speakers' voices (Tan et al., 2021b). Different from Tan et al. (2021b), our voice conversion model is non-autoregressive, and we employ two modules, adversarial speaker classifier and embedding consistency loss, during training to further improve the speaker similarity performance.

3. Data-sufficient scenario

This section describes our proposed method for cross-lingual speech synthesis under the data-sufficient scenario. Generally, we have more than 8 h of data per speaker for training.

3.1. Input representation

Code-switching is defined as more than one language occurring in one sentence or between sentences. With the world's globalization, code-switching patterns in speech have become a common case in many countries and regions (Bernardo, 2005). The globalized language environment leads to more bilinguals and polyglots, which motivates researchers to develop speech processing systems that can handle multilingual challenges. Furthermore, code-switching corpora are collected and released for research related to speech communication in the recent decade (Lyu et al., 2010; Shen et al., 2011), followed by various approaches proposed to address complex speech analysis, including multilingual automatic speech recognition (ASR), language identification, and language diarization concerning the multilingual scenario (Ahmed and Tan, 2012; Vu et al., 2012; Lyu and Lyu, 2008; Lyu et al., 2013). Likewise, TTS systems need to be improved for synthesizing natural speech for code-switching sentences (Zhou et al., 2020).

One of the main challenges of code-switching TTS is that the grapheme or phoneme set between languages is different. However, some phonetic pronunciations between different languages are close. Thus exploring a multilingual TTS model with minimum data requirement, including textual and vocal data, is possible and essential. Previous approaches proposed for addressing multilingual issues in TTS indicate that shared input representation across languages is one of the keys to realizing cross-lingual synthesis (Li et al., 2019; Li and Zen, 2016; Lee et al., 2018). The shared representations include shared phoneme set, international pronunciation alphabet (IPA), and the Bytes coding (Li et al., 2019), where the phoneme representation can perform better (Zhang et al., 2019).

In this paper, we choose to use a shared phoneme set from the CMU dictionary (The Carnegie Mellon, 2015) to investigate bilingual multi-speaker TTS and cross-lingual synthesis between Mandarin and English. As for Mandarin, the pronunciation representation called pinyin can be converted to CMU phoneme by the pinyin-to-cmu mapping table (Chao Weng, 2012). Since Mandarin is a tone-language, digits 1 to 6 are used to denote different tones for Mandarin phonemes, and '0', '1', '2' are used to mark the lexical stress for English phonemes. In this case, part of the tones and stress share the same annotations, which may cause ambiguity. However, we also have numeric language ID codes as another input stream. Language ID codes are used to obtain language-dependent encoding features while preserving the shared information between languages, like close pronunciations. Specifically, '0', '1', '2' are used for language identification for every corresponding input phoneme, where '0' specifies that the corresponding phoneme or stress annotation is from English, '1' is for Mandarin, and '2' is for language-unrelated symbols like punctuation marks. Take the phrase 'speech 合成.' (speech synthesis.) as an example; two input sequences are obtained after the front-end text processing. One is the phoneme sequence 'S P IY 1 CH HH ER 2 CH AH 2 NG 2 .', and the other is the corresponding numeric language ID codes '0 0 0 0 0 1 1 1 1 1 1 1 1 2', which has the same length as the phoneme sequence. We break up phonemes with their corresponding tones, e.g., 'AH2' is converted to 'AH 2', to allow our proposed model to share close pronunciations between Mandarin and English.

3.2. Proposed model

Our proposed bilingual multi-speaker TTS model is illustrated in Fig. 1. The input text is converted into phoneme sequence and language token sequence, as introduced in Section 3.1. The phoneme sequence is converted to a phoneme embedding sequence by a learnable lookup table. Correspondingly, the numeric language ID codes are converted to a 64-dimensional language embedding sequence through another learnable embedding table. Two embedding sequences are concatenated together as the input of the Tacotron encoder, which accumulates the linguistic and context characteristics of the input vector sequence with several convolutional layers and a bi-directional long short-term memory (BLSTM) layer.

256-dimensional speaker embedding is concatenated with the encoder outputs for conditioning the network to synthesize expected voices. For the speaker embedding, we use the mean embedding derived from all embeddings extracted with a pre-trained ResNet-based speaker verification model (Cai et al., 2020a) by feeding all training utterances of each speaker. We believe it can lead to the same performance as using a trainable lookup table yet costs less training time. Mel-spectrogram is used as the predicted acoustic feature in our bilingual multi-speaker TTS model.

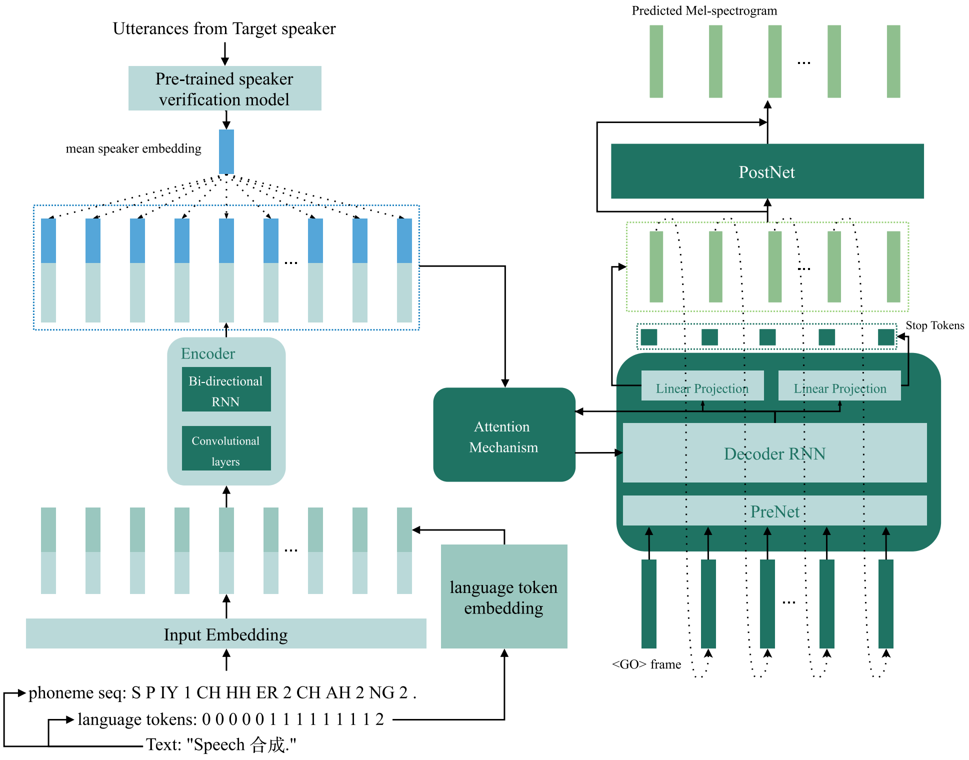


Fig. 1. Proposed multilingual multi-speaker TTS model.

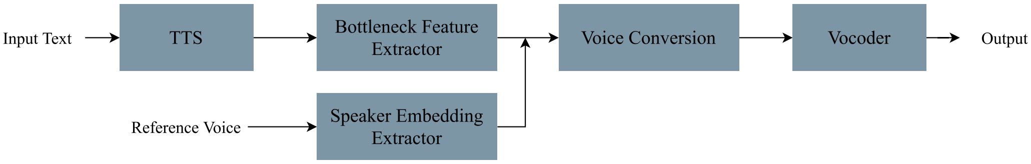


Fig. 2. Synthesis pipeline for data-insufficient scenario.

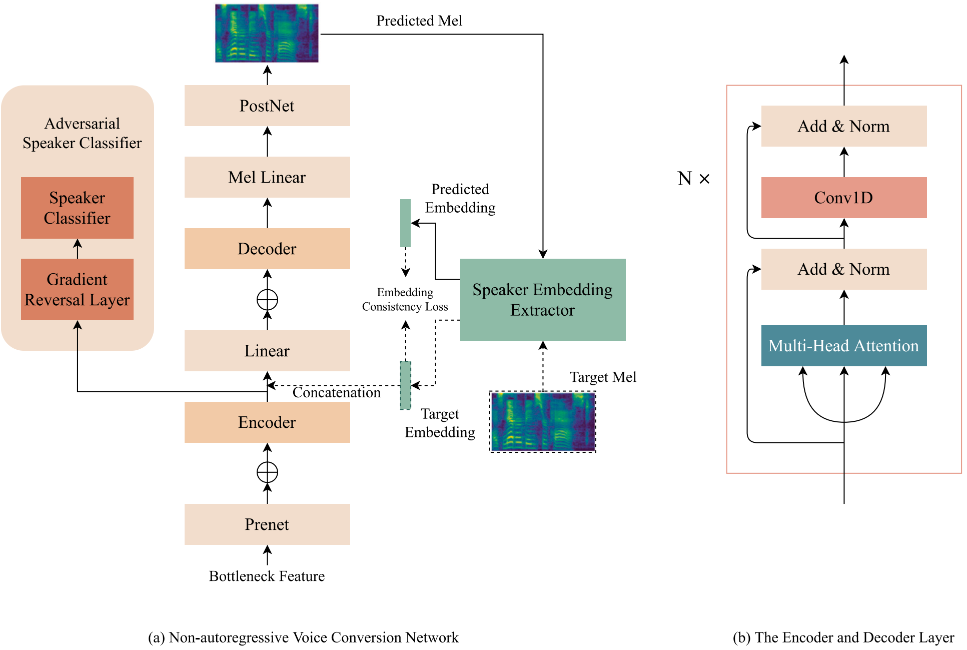
3.3. Vocoder

The vocoder participates in TTS systems to transform the acoustic features back to audio signals in the time domain. Both Griffin–Lim (Griffin and Lim, 1984) and neural vocoders (van den Oord et al., 2016; Kalchbrenner et al., 2018; Kumar et al., 2019) can be applied in our framework to reconstruct the waveform. This work uses MelGAN (Kumar et al., 2019) as our vocoder for the proposed methods in both scenarios since MelGan is much faster in waveform generation while maintaining high quality.

4. Data-insufficient scenario

One of the low-resource cases in cross-lingual multi-speaker synthesis is the utterance-limited scenario where we have limited data per speaker for training. The duration of the audio data per speaker is less than 30 min. However, we still have hundreds of voices to model. It is difficult for the end-to-end TTS framework to model such varieties regarding the speaker space and language characteristics with such limited data.

We have investigated the cross-lingual TTS performance in such cases using the proposed framework in Section 3. It turns out that the model performs well in multilingual multi-speaker synthesis. However, the performance on cross-lingual synthesis is poor. It is hard to generate accurate speech with foreign text for monolingual speakers. For example, the system cannot synthesize English speech with Mandarin speakers' voices. Therefore, we adopt a synthesis pipeline that consists of several speech modules for cross-lingual synthesis. As shown in Fig. 2, it contains a bilingual TTS system, a bottleneck feature extractor, a speaker embedding extractor, a voice conversion (VC) system that converts the voice to the expected voice, and a vocoder for transforming the predicted acoustic features to audio signals.



(a) Non-autoregressive Voice Conversion Network

(b) The Encoder and Decoder Layer

Fig. 3. The architecture of the multi-speaker voice conversion system. (a) The non-autoregressive voice conversion network. (b) The structure of encoder and decoder layer.

4.1. Bottleneck feature extractor

The intermediate linguistic feature vector used in the VC system is important for synthesis performance. Here we adopt the speaker-independent bottleneck feature extracted from a bilingual speech recognition model trained with Kaldi (Povey et al., 2011). Typically, speech recognition is trained on audio–text pairs. The recognition process can break down to acoustic feature extraction, phonetic unit prediction, and decoding via maximum likelihood estimation on context models like language models. The key module we borrow from the speech recognition system is the acoustic model that predicts phonetic probabilities from acoustic features. Here the acoustic model contains a bottleneck layer that we use as the linguistic feature vector. Therefore it is adopted as the bottleneck feature extractor.

In our work, the acoustic model is constructed by time-delayed neural networks (TDNN), where the linear layer before the output layer is designed to be a low-dimensional layer, also known as the bottleneck layer (Grézl et al., 2007). Since the acoustic model is trained to maximize the probability on the true phonetic label for each acoustic frame, the output from the bottleneck layer in a well-trained model should contain precise linguistic information. Thus we can adopt the output of the bottleneck layer as the linguistic feature vector for voice conversion. The acoustic model is trained with multilingual data to extract language-independent features for multilingual scenarios.

4.2. Speaker embedding extractor

The speaker embedding extractor comes from models designed for speaker verification tasks. Speaker verification is the task of identifying persons from their voices. Recently, deep learning has revolutionized the speaker verification field. The X-vector-based system (Snyder et al., 2018) and its variant frameworks (Desplanques et al., 2020; Cai et al., 2018) have become the most popular architectures in speaker verification. Normally, a deep speaker verification model contains a front-end pattern extractor, an encoder layer, and a back-end classifier. The fully connected layer is used as the speaker embedding, which is a discriminative fixed-length vector to represent a speaker's identity. Here we employ the speaker embedding in our cross-lingual voice conversion system to render the target speaker's voice characteristics. In addition, the dataset used in the data-insufficient scenario contains a rather large number of speakers, allowing us to train a network that is robust enough for unseen speaker synthesis. Therefore, unlike the data-sufficient scenario, which uses mean speaker embedding as the representation, the voice conversion system uses utterance-level speaker embedding as shown in Fig. 3.

4.3. Cross-lingual voice conversion

We propose a non-autoregressive model for voice conversion. As shown in Fig. 3, the framework is a variant of the synthesis network FastSpeech (Ren et al., 2019). We remove the length regulator module since the input sequence and the output sequence share the same length in the VC task. The network comprises a speaker encoder, an encoder–decoder structure with a multi-head attention mechanism, and an adversarial speaker classifier.

FastSpeech is first proposed for converting text-embedding sequence to acoustic features, while VC converts linguistic feature vectors to acoustic features. We use Mel-spectrogram as the acoustic feature. For the encoder–decoder structure, we replace the character-embedding layer with a PreNet that contains two fully connected layers, each with 256 hidden units. We add triangular position encoding (Vaswani et al., 2017) to the input sequences of the encoder and decoder to provide the location information. The encoder contains a stack of $N = 4$ identical blocks. Each block has two multi-head self-attention modules, followed by two 1D convolutional layers. Residual connections and layer normalization are applied in each convolutional layer. To perform multi-speaker VC, we condition the decoder with speaker embeddings. The speaker embedding is concatenated with the encoder output to provide speaker information. The decoder has the same feed-forward network structure as the encoder, which significantly speeds up the training and inference process compared to autoregressive models. Finally, a Post-Net module consisting of 5-layer convolution is added to obtain the residual coefficients from the predicted acoustic features to improve the overall reconstruction quality.

The bottleneck feature vector, which is extracted from source speech, has been proven to contain voice characteristics from the source speaker in many-to-many VC systems (Ding et al., 2020). To further eliminate speaker information and prevent the converted voice from resembling source speaker's voice, we employ an adversarial speaker classifier in our proposed framework (Meng et al., 2019; Ding et al., 2020). The adversarial module contains a gradient reversal layer and a speaker classifier. Both are linear layers, while the latter is used to produce probabilities for speakers from the training set. The gradient reversal layer scales the gradient flowing to the encoder reversely by an adjustable factor λ during backward propagation. The adversarial speaker classifier is optimized to reduce the cross-entropy loss of speaker classification during training.

We also use the embedding consistency loss (Cai et al., 2020) in our framework, which is proposed to improve the speaker similarity between the synthesized speech and its reference voice. Simply concatenating the speaker embedding may not transfer enough speaker information learned by the verification system, especially for cross-lingual VC. Therefore, we incorporate the speaker verification model in our VC training to reinforce the voice cloning ability. We use the embedding consistency loss between the ground truth speaker embedding and the one extracted from the predicted Mel-spectrogram as one component of the loss functions for optimizing the VC network. Hyperparameter α is used to control the weight of the embedding loss. During the training stage, the parameters of the speaker encoder network are frozen.

5. Experiments and results

5.1. Data-sufficient scenario

For the data-sufficient scenario, our experiments are conducted with the framework illustrated in Section 3. Three TTS datasets are used to investigate the cross-lingual synthesis performance, including the publicly available LJ Speech (LJS) dataset (Ito, 2017) and two Chinese datasets, Female DB-1 and Female DB-4, from Data Baker[3] (**LJS**, **DB-1** and **DB-4** are notated for both speaker identity and dataset in this section). DB-1 is an open-source dataset,[4] while DB-4 is a commercial one. LJS contains approximately 24 h of English audio-transcript pairs recorded by a female English native speaker. The DB-1 has approximately 12 h of Mandarin speech synthesis data recorded by a Chinese female speaker. The DB-4 is a bilingual dataset containing 12 h of Chinese audio-transcript pairs, 6 h of English pairs, and 6 h of code-switching data from a Chinese female speaker.

The frequencies of all phonemes in the three datasets are shown in Appendix A. Most phonemes between two languages share the same representation in our experiments. This indicates that the pronunciation of intersecting shared phonemes may be less challenging to learn by a cross-lingual TTS system than phonemes that only exist in one language. Moreover, cross-lingual synthesis can be achieved when the model catches the pronunciation similarity of these phonemes between English and Mandarin.

5.1.1. Training setup

We trained two bilingual multi-speaker TTS systems with different datasets. The first system, notated as **BLMS**, is the bilingual multi-speaker TTS model trained with DB-1 and LJS. The other system, notated as **CLMS**, is the cross-lingual system trained with all three datasets, including the bi-lingual dataset DB-4. Although the latter system also can be used for bilingual multi-speaker synthesis, we focus on its capability of cross-lingual synthesis here. All training audio samples are downsampled to 16 kHz. The hyperparameter settings for acoustic feature extraction and network components are shown in Appendix B.

5.1.2. Subjective evaluations

The subjective evaluation is done by speech synthesis MOS-scale rating, a categorical score from 1 to 5, with 0.5 increments, where score 5 is the best. We ask 17 native Mandarin speakers (all evaluators speak fluent English) to rate the synthesized speech. We have three types of synthesized text for evaluating the TTS synthesis performance: Mandarin sentences, English sentences, and code-switching sentences that contain both Mandarin and English content in each sentence. Each type of text has 15 sentences for synthesis. Therefore, there are 90 synthesized utterances from BLMS and 135 utterances from CLMS. Each evaluator rates every chosen sentence regarding naturalness, similarity, and intelligibility. The naturalness is related to the quality of synthesized audio samples regardless of the content. The speaker similarity score measures how close the synthesized voice is to the expected speaker, while the intelligibility score evaluates the clarity level of the speech content.

Table 1
The mean opinion scores (MOS) with 95% confidence interval (CI) for all proposed systems under the data-sufficient scenario. BLMS is the bilingual multi-speaker TTS model trained with DB-1 and LJS, while CLMS is the cross-lingual TTS model trained with DB-1, DB-4, and LJS. For synthesis type, CN denotes Mandarin sentences, EN denotes English sentences, and CS denotes code-switching sentences that contain both Mandarin and English.

MOS ± 95%CI		BLMS		CLMS		
		DB-1	LJS	DB-1	LJS	DB-4
Naturalness	CN	3.97 ± 0.1	2.73 ± 0.12	4.01 ± 0.1	3.02 ± 0.11	3.99 ± 0.1
	EN	2.86 ± 0.13	3.86 ± 0.09	3.86 ± 0.08	3.96 ± 0.08	4.04 ± 0.08
	CS	3.4 ± 0.11	3.05 ± 0.11	3.81 ± 0.1	3.24 ± 0.1	3.95 ± 0.1
	ALL	3.41 ± 0.07	3.21 ± 0.07	3.89 ± 0.06	3.41 ± 0.06	3.99 ± 0.05
Intelligibility	CN	4.58 ± 0.06	2.38 ± 0.13	4.64 ± 0.06	3.54 ± 0.13	4.65 ± 0.06
	EN	1.83 ± 0.12	4.17 ± 0.1	4.17 ± 0.09	4.37 ± 0.08	4.37 ± 0.08
	CS	3.4 ± 0.1	2.92 ± 0.13	4.29 ± 0.09	3.68 ± 0.12	4.41 ± 0.07
	ALL	3.27 ± 0.1	3.16 ± 0.09	4.37 ± 0.05	3.86 ± 0.07	4.47 ± 0.04
Similarity	CN	4.25 ± 0.08	3.16 ± 0.1	4.21 ± 0.07	3.18 ± 0.1	4.16 ± 0.08
	EN	3.37 ± 0.11	3.64 ± 0.09	3.91 ± 0.07	3.84 ± 0.09	4.08 ± 0.08
	CS	4.14 ± 0.07	3.26 ± 0.1	4.13 ± 0.08	3.31 ± 0.1	4.11 ± 0.08
	ALL	3.92 ± 0.06	3.35 ± 0.06	4.09 ± 0.04	3.44 ± 0.06	4.12 ± 0.04

Table 2
p-values obtained by the Mann–Whitney U test comparing MOS results between BLMS and CLMS. The test is performed under the null hypothesis that the distribution underlying MOS samples from system BLMS is the same as the distribution underlying samples from the CLMS.

p-value	DB-1			LJS		
	CN	EN	CS	CN	EN	CS
Naturalness	0.169	$<10^{-5}$	$<10^{-5}$	$<10^{-3}$	0.063	$<10^{-2}$
Intelligibility	0.121	$<10^{-5}$	$<10^{-5}$	$<10^{-5}$	$<10^{-2}$	$<10^{-5}$
Similarity	0.215	$<10^{-5}$	0.489	0.436	$<10^{-2}$	0.369

The mean opinion scores (MOS) on naturalness, intelligibility, and similarity are shown in Table 1. Accordingly, as MOS ratings are ordinal values rather than numeric (Rosenberg and Ramabhadran, 2017), we perform the Mann–Whitney U test to compare the results between system BLMS and CLMS. The statistical significance levels are shown by two-sided *p*-values in Table 2. As shown in Table 1, the quality of synthesized audio samples varies among different systems and speakers for the subjective evaluation of naturalness. Generally, the quality reaches around 4 when synthesizing audio samples in the target speaker's native language, while the performance degrades when generating cross-lingual speech for monolingual speakers. For example, for system BLMS, DB-1 obtains a MOS of 3.97 when synthesizing Mandarin sentences, but the score degrades to 2.86 for English sentences. The naturalness performances on the cross-lingual and code-switching synthesis are improved when we include the bilingual dataset for training. DB-1 yields a MOS of 3.86 in synthesizing English text and 3.81 in synthesizing code-switching text by CLMS, while DB-1 achieves a MOS of 2.86 in synthesizing English text and 3.4 in synthesizing code-switching text by BLMS, respectively. Similar results can be observed from the speaker LJS between the two systems. Those results are all statistically significant at the $p < 10^{-2}$ level. In addition, as shown by the similarity scores on Table 1, the speech synthesized by our proposed model can well preserve the speaker identity according to the speaker embedding. Most speaker similarity MOS are around 4, while scores lower than 4 can be observed in cross-lingual cases.

Most essentially, the code-switching performance can be observed from Table 1. Although BLMS can achieve bilingual multi-speaker synthesis, the cross-lingual synthesis performance is poor, which matches the result from Zhang et al. (2019). The cross-lingual synthesized speech is barely intelligible as the cross-lingual intelligibility MOS is pretty low. It achieves a score of 1.83 for DB-1 when synthesizing English sentences and a score of 2.38 for LJS when synthesizing Mandarin sentences. However, the CLMS system is able to generate cross-lingual speech, even in code-switching cases, with intelligible pronunciations for monolingual speakers. The cross-lingual synthesis performance on intelligibility is significantly improved in this case, where the CLMS system achieves a score of 4.17 for DB-1 in English sentences synthesis and a score of 3.54 for LJS in Mandarin sentence synthesis. Raters said that synthesized speech is exactly like a foreign speaker speaking another language with an accent from their native language. The result indicates that using a bilingual dataset with our proposed model can significantly improve cross-lingual speech synthesis for monolingual speakers.

5.2. Data-insufficient utterance-limited scenario

5.2.1. The bottleneck extractor

The English dataset Librispeech (Panayotov et al., 2015) and the Mandarin dataset AISHELL-2 (Du et al., 2018) are used to train our bilingual bottleneck extractor. The recipe to train Librispeech in Kaldi is used for our model training. The acoustic model, known

[3] https://www.data-baker.com/en.

[4] https://www.data-baker.com/open_source.html.

Table 3
The ASR performance of the bottleneck extractor.

Test set	WER/CER
Librispeech Dev-clean	3.5%
Librispeech Test-clean	3.9%
AIShell-2 Dev	4.29%
AIShell-2 Test	4.59%

as the chain model in Kaldi, has 17 TDNN layers, followed by the 256-dimensional bottleneck layer. The frames' sub-sampling factor is set to 1 so that the frame length of output bottleneck features matches the length of the input acoustic features. The phoneme set we used for building the recognition dictionary includes 39 English phonemes and 52 Mandarin phonemes. We use 50 ms window-length and 12.5 ms frame-shift for MFCC feature extraction, which is the same setting as in TTS. In Table B.9, the hop length with 200 samples for input audio samples in 16 kHz is the same as 12.5 ms frame-shift. We demonstrate the performance of our bottleneck extractor by applying the model in speech recognition. Here the performance of English speech recognition is reported by word error rate (WER), while those of Mandarin are reported by character error rate (CER). As shown in Table 3, our model achieves low recognition error rates on test sets from both languages. It achieves a WER of 3.5% on the Librispeech development set and a CER of 4.29% on the AISHELL-2 development set. Hence the quality of this acoustic model is acceptable for linguistic feature vector extraction.

5.2.2. Speaker embedding extractor

Our experiments deploy two different speaker embedding extractors trained by different datasets. One is the ECAPA-TDNN (Desplanques et al., 2020) model incorporated as the speaker verification model during voice conversion training, as shown in Section 4.3. It is trained with the AISHELL-2 and the VCTK dataset (Veaux et al., 2016). The training set contains 1901 speakers with more than 900,000 utterances from the AISHELL-2 database and 100 speakers from the VCTK dataset, while utterances from another 100 speakers in AISHELL-2 and 9 from VCTK are excluded as the test set to evaluate the verification performance. The other embedding extractor is the ResNet-based speaker verification model (Cai et al., 2020a) trained on the VoxCeleb2 (Chung et al., 2018) dataset. It is the same model mentioned in Section 3.2. The ResNet-based model here is only used for evaluation. Since the ECAPA-TDNN model is employed in the voice conversion system to improve the speaker similarity, we surely achieve a better verification performance when evaluating the converted speech with the same model. Hence we need the ResNet-based model with a different structure for a fair assessment. Normally, the speaker verification performance is measured by the equal error rate (EER) and the minimum detection cost function (mDCF). The ECAPA-TDNN model achieves an EER with 2.26% on the AISHELL-VCTK test set, and the mDCF on the test set is about 0.41. The result shows that most of the utterance pairs constructed from the test set are correctly verified. Therefore, the verification system we trained is able to extract discriminative representations as the conditioning feature for the multi-speaker voice conversion system.

5.2.3. Multilingual multi-speaker voice conversion

Three publicly available datasets are used in our experiments, including the LJ Speech (LJS) dataset (Ito, 2017) introduced in Section 5.1, the VCTK English dataset (Veaux et al., 2016), and the AISHELL-3 Mandarin dataset (Shi et al., 2021). The VCTK English corpus contains 109 speakers with various accents. 100 speakers are randomly chosen for training, while the rest speakers are used during the test phase. For AISHELL-3, we select 174 speakers for training. Each speaker from the two datasets contains approximately 400 utterances, about 20 min long in total, for training. All audio samples are downsampled to 16 kHz. Settings for extracting Mel-spectrogram are the same as the one used in the utterance-limited scenario (Table B.9). Hyperparameters λ, α are set to 1.0 and 5.0, respectively.

5.2.4. Subjective evaluation

We finetuned the cross-lingual multi-speaker (CLMS) system from the data-sufficient scenario with datasets DB-4, AISHELL-3, and VCTK, then used it as the baseline system for comparison. As for the synthesis pipeline using VC, we first use the CLMS model from the data-sufficient scenario to generate speech with DB-4's voice. Then we use the voice conversion system to convert the synthesized speech to our target voice. Two voices from VCTK and two from AISHELL-3 are randomly chosen from the training set for subjective evaluation. 20 native Mandarin speakers (all evaluators speak fluent English) are asked to rate the synthesized speech on MOS-scale. We randomly choose 5 utterances per target voice and synthesized sentence type for rating. However, the CLMS system trained with limited data fail to generate English (EN) speech with voices from AISHELL-3, nor do Mandarin sentence (CN) with voices from VCTK. There is no synthesized utterance for evaluation in those cases. Besides, the synthesized speech from the CLMS model has an early-stopping issue. We ask raters to evaluate the performance, especially the intelligibility, only regarding the unfinished sentences from the monolingual results. Hence each rater evaluates 35 utterances from the CLMS system and 60 utterances from the voice conversion system.

Results are shown in Table 4, and the corresponding statistical significance levels are shown in Table 5. Regarding naturalness and speaker similarity, the voice conversion performance is not as good as the CLMS system. Compared to the CLMS system, the naturalness MOS on synthesizing English sentences with voices from VCTK degrades from 3.73 to 3.24 with a statistically significant level at $p < 10^{-5}$. Similarly, the similarity MOS from the same case degrades from 3.93 to 3.63 with a statistically

Table 4
The mean opinion scores (MOS) with 95% confidence interval (CI) for system CLMS and the voice conversion pipeline under the data-insufficient scenario. CLMS is the cross-lingual TTS model trained with DB-4, VCTK, and AISHELL-3. VCTK denotes voices chosen from the VCTK dataset, and AISHELL-3 denotes voices chosen from the AISHELL3 dataset. For synthesis type, CN denotes Mandarin sentences, EN denotes English sentences, and CS denotes code-switching sentences that contain both Mandarin and English.

MOS ± 95%CI		CLMS (data-insufficient)			Voice conversion	
		DB-4	VCTK	AISHELL-3	VCTK	AISHELL-3
Naturalness	CN	4.47 ± 0.1	–	3.72 ± 0.12	3.27 ± 0.1	3.47 ± 0.1
	EN	3.98 ± 0.13	3.73 ± 0.1	–	3.24 ± 0.11	3.15 ± 0.1
	CS	3.95 ± 0.13	–	–	3.18 ± 0.1	3.31 ± 0.1
	ALL	4.13 ± 0.07	–	–	3.23 ± 0.06	3.31 ± 0.06
Intelligibility	CN	4.7 ± 0.08	–	4.27 ± 0.1	4.41 ± 0.09	4.42 ± 0.08
	EN	4.01 ± 0.15	3.76 ± 0.12	–	4.18 ± 0.09	4.06 ± 0.09
	CS	4.25 ± 0.12	–	–	4.07 ± 0.11	4.16 ± 0.09
	ALL	4.32 ± 0.07	–	–	4.22 ± 0.06	4.22 ± 0.05
Similarity	CN	4.26 ± 0.12	–	3.82 ± 0.11	3.54 ± 0.11	3.36 ± 0.12
	EN	3.9 ± 0.14	3.93 ± 0.1	–	3.63 ± 0.12	3.46 ± 0.11
	CS	4.19 ± 0.12	–	–	3.44 ± 0.13	3.32 ± 0.12
	ALL	4.12 ± 0.07	–	–	3.54 ± 0.07	3.38 ± 0.07

Table 5
p-values obtained by the Mann–Whitney U test comparing MOS results between CLMS (data-insufficient) and the Voice Conversion (VC) system. The test is performed under the null hypothesis that the distribution underlying MOS samples from system CLMS (data-insufficient) is the same as the distribution underlying samples from the VC system.

p-value	Naturalness	Intelligibility	Similarity
VCTK-EN	$<10^{-5}$	$<10^{-5}$	$<10^{-3}$
AISHELL-3-CN	$<10^{-3}$	0.062	$<10^{-5}$

significant level at $p < 10^{-3}$. We observe the same outcomes from synthesizing Mandarin speech with voices from AISHELL-3. This indicates that the multi-speaker synthesis model outperforms the voice conversion method, regardless of the cross-lingual scenarios and early stopping issues. However, the intelligibility performance of the voice conversion results is outstanding. Concerning the voice conversion system, all synthesis scenarios, including cross-lingual and code-switching synthesis, achieve a MOS above 4. The results are consistent with the synthesis performance of speaker DB-4 in Table 1 since converted utterances are transformed based on synthesized utterances with the DB-4 voice. The voice conversion module does not have much of an impact on intelligibility. The CLMS system fails to generate cross-lingual speech for monolingual speakers in the data-insufficient scenario. In contrast, the voice conversion system is able to synthesize highly intelligible speech from cross-lingual text with a slight decrease in speech naturalness and speaker similarity at a strong statistically significant level, according to the MOS results.

5.2.5. Ablation study

The performance on speaker similarity is essential for multi-speaker voice conversion. As we adopt the VC for the data-insufficient scenario, the performance of cross-lingual multi-speaker synthesis is affected by the performance of the VC model. Therefore we provide an ablation study on the speaker embedding consistency loss (ECL) to investigate the improvement regarding speaker similarity.

We train another voice conversion system without using the embedding consistency loss. Our proposed model enables zero-shot conversion even for unseen voices by conditioning speaker representations from embedding extractors. In our experiments, we convert utterances from test sets to utterances with voices from both seen and unseen speakers for the system trained with ECL and the one trained without ECL. The unseen voices come from the test set of VCTK and AISHLL-3. Then for both seen and unseen cases, we synthesize 2000 utterances for each scenario list below:

- Monolingual scenario: convert voices between monolingual speakers with the same language; for example, convert an English speaker voice to another English speaker's voice.
- Cross-lingual scenario: convert voices across monolingual speakers that speak different languages; for example, convert the voice of a Mandarin speaker to an English speaker's.

In terms of voice conversion between seen speakers, 24 utterances are randomly selected, with 12 from each scenario, for subjective evaluation. Similarly, 24 utterances, with 12 from each scenario, are chosen for evaluating unseen voice conversion. Therefore, 48 utterances from the system trained with ECL and the corresponding 48 utterances from the system trained without ECL for rating. 20 native Mandarin speakers are asked to rate the converted samples. Each listener rates every chosen utterance. The similarity MOS results are shown in Table 6. There is a slight decrease in speaker similarity for the system trained with ECL. The statistically significant level is at $p < 0.001$, as shown in the table. Both systems obtain a MOS above 3.5 on speaker similarity, which means the converted voice has a fair probability of being the same voice as the target one.

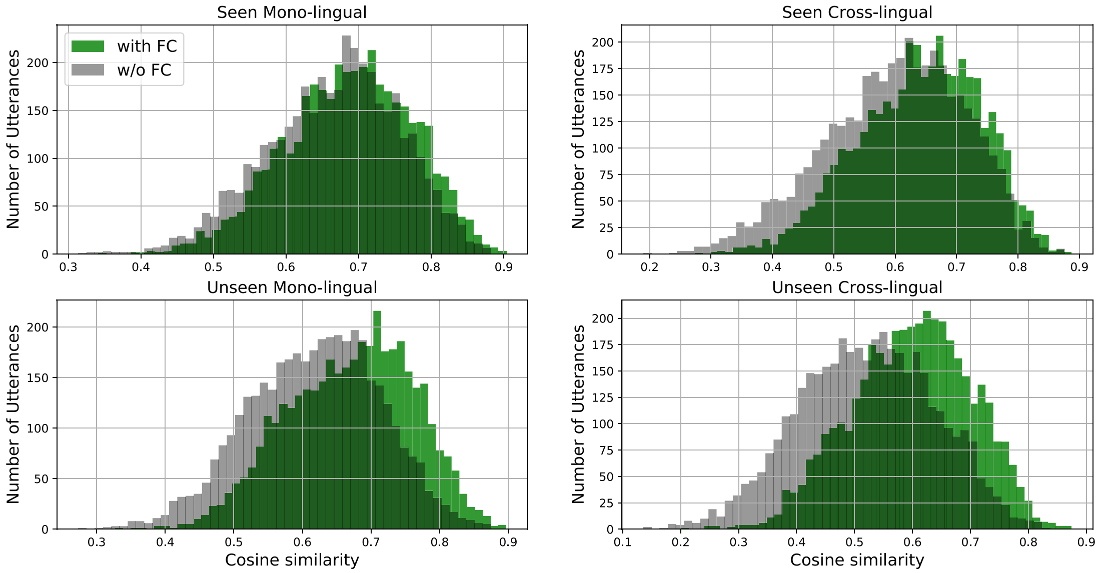


Fig. 4. The distribution of cosine similarity scores between speaker embeddings from the reference speech and the converted speech from the voice conversion system.

Table 6
Speaker similarity MOS results on systems trained with and without ECL. Both systems are trained with datasets DB-4, AISHELL-3, and VCTK under the data-insufficient scenario.

MOS	w/o ECL	With ECL	*p*-value
Seen	3.82 ± 0.07	3.56 ± 0.07	$<10^{-5}$
Unseen	3.71 ± 0.07	3.55 ± 0.07	0.001

Table 7
The speaker verification performance of systems trained with and without ECL, reported by Equal Error Rate (EER).

EER (%)	w/o ECL	With ECL
Seen Mono-lingual	18	18.98
Seen Cross-lingual	26.42	24.6
Unseen Mono-lingual	22.68	21.65
Unseen Cross-lingual	34.35	27.92

We do not observe improvement in speaker similarity from the subjective evaluation. However, the system with ECL demonstrates better spoofing capability from objective evaluation. We use a different embedding extractor to conduct our objective evaluation concerning the use of ECL. Since the ECAPA-TDNN model is incorporated in the voice conversion model during training, we surely achieve a higher similarity score between the converted speech and the reference speech when we use the same model for evaluation. Therefore, a pre-trained ResNet-based speaker verification model (Cai et al., 2018) is used to evaluate the verification performance in this experiment. The verification model is trained on the VoxCeleb2 dataset and achieves an EER with 1.94% on the AISHELL-2 test set.

For each synthesized utterance in this experiment, we extract the speaker embedding of the converted result and one from its corresponding reference utterance using the ResNet-based verification model. Then we evaluate the speaker similarity based on cosine similarity scores. The objective verification performance is presented in Fig. 4. The score distribution shows that the voice conversion system with ECL achieves higher similarity than the one without ECL. The mean of similarity scores of the system with ECL is larger than the mean of scores from the system without ECL except in the seen monolingual scenario. This indicates that the system trained with ECL has improvement on speaker similarity from the verification model's perspective. Thus adopting ECL may boost our proposed voice conversion system on spoofing speaker verification systems. In addition, according to the distributions shown in Fig. 4, the similarity score gap between systems trained with and without ECL is more significant in the cross-lingual case. The cosine similarity score difference in the seen monolingual case is 0.2, while it is 0.4 in the seen cross-lingual case. The score improvement increases from 0.05 to 0.08 in the unseen case. Therefore, the use of ECL in our voice conversion system yields significant verification improvement on unseen speaker synthesis and cross-lingual synthesis.

Furthermore, the improvement regarding speaker verification performance can be verified in Table 7. For each scenario, we randomly generate 8000 synthesized-real embedding pairs, where half of them are from the same speaker, and the other half are from different speakers. All speaker embeddings are extracted by the ResNet-based model. Then we calculate the equal error rate (EER) according to labels that indicate if the pair belongs to the same speaker or not. As shown in Table 7, the system trained with ECL achieves lower EERs on all scenarios except mono-lingual synthesis from the seen speakers. Particularly, the verification improvement on cross-lingual cases is significant. Compared to the system trained without ECL, the system trained with ECL yields 6.9% relative EER improvement in the Seen Cross-lingual case, and it obtains 18.7% relative EER improvement in the Unseen Cross-lingual case.

6. Discussion

While the performance of our proposed systems works well on the cross-lingual synthesis and code-switching synthesis, there are several limits that need further study. For the data-sufficient scenario, we require a bilingual dataset to accomplish the knowledge transfer between languages and speakers. Thus the shared phonemes we used for two languages can be bridged, and we are able to synthesize foreign text with monolingual speakers' voices. However, there is still a noticeable performance gap between cross-lingual and intra-lingual synthesis concerning naturalness and intelligibility. Besides, the absence of the bilingual dataset leads to unclear pronunciations and unintelligible results on cross-lingual synthesis. This phenomenon was found in many prior studies (Zhang et al., 2019; Liu and Mak, 2020), even when there is a large amount of data for training. According to a European survey, around 54% of people are bilingual, while multilingual speakers, who speak more than two languages, are only a small part of the population, especially for minority languages (European Commission, 2012). Besides, some languages do not share similar pronunciation units as we do in our experiments. Regarding those issues, developing a cross-lingual system for multiple languages is an arduous task. As future work, the study towards a universal speech synthesis system that involves more language locales is vital, which has already started (Yang and He, 2020).

Cascading several speech modules is one of the ways that achieves high-quality synthesis (Huang et al., 2020; Zhao et al., 2020). However, like other cascading approaches, the approach we used in the data-insufficient case requires more computation resources and time than end-to-end TTS approaches. In addition, the robustness of the pipeline is limited as we need per-system adaptation for almost all speech models from the pipeline to adapt to novel languages. To address those issues, universal speech recognition and zero-shot multi-speaker voice conversion are essential. However, as we have shown in Section 5.2.5, although synthetic voice may spoof machines, a human can still distinguish synthetic voices from true voices. One of the reasons is that the voice conversion model is trained in a low-resource data setup. Thus we still need future studies on improving the speaker similarity under the low-resource scenario.

7. Conclusion

We present two bilingual multi-speaker TTS approaches and investigate the cross-lingual performance with limited bilingual data for two data setups. One is a Tacotron-based model for the data-sufficient scenario. The model takes shared phonemic representations along with numeric language ID codes as input. When trained with monolingual data from Mandarin and English, the model is able to achieve high-fidelity bilingual multi-speaker TTS. In addition, by involving a bilingual dataset, the model allows monolingual voices to synthesize cross-lingual speech and even code-switching speech. The other approach is proposed for realizing cross-lingual synthesis in low-resource scenarios. Several speech modules, including a bottleneck feature extractor, a speaker embedding extractor, and a voice conversion system, are applied for this approach. In particular, we proposed a parallel non-autoregressive network for cross-lingual voice conversion. Experimental results show that our proposed conversion model can synthesize high-quality converted speech with good speaker similarity. Furthermore, we adopt embedding consistency loss during model training and evaluate its effectiveness on speaker similarity. From objective and subjective evaluations, we observe that adding embedding consistency loss does not improve speaker similarity from the human perspective, but it significantly improves speaker similarity from the speaker verification system's perspective.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

LJ Speech and DB1 are open source databases. DB4 is a commercial database that can be licensed from Data Baker company directly.

Appendix A. Phoneme frequencies from datasets LJS, DB-1 and DB-4

Table A.8 illustrates the frequencies of all phonemes in three datasets. LJS contains only English utterances, while DB-1 contains only Chinese utterances. Three consonants, ‘J’, ‘X’, and ‘Q’ do not exist in the English dataset when using shared phoneme representations. However, these three phonemes frequently exist in the Mandarin dataset. On the other hand, 7 phonemes are not presented in the Mandarin dataset while frequently occurring in the English dataset, as shown in the table. The bilingual dataset DB-4 contains all phonemes.

Table A.8
Phonemes (without tone and stress) and their corresponding frequencies in LJ-Speech, DB-1 and DB-4.

Phoneme	LJS	DB-1	DB-4	Phoneme	LJS	DB-1	DB-4	Phoneme	LJS	DB-1	DB-4
J	–	10 088	12 499	X	–	8050	11 895	Q	–	5435	7489
IY	28 587	54 859	85 601	EH	26 397	3598	11 791	AA	16 976	11 173	23 205
L	32 893	9420	23 510	AY	12 079	7479	15 619	UW	15 345	30 630	44 593
SH	7957	11 456	17 804	OW	10 201	6921	13 698	Y	4426	16 540	27 793
N	68 392	33 006	56 359	T	65 657	8698	26 504	JH	4824	8994	13 821
AE	21 502	27 640	42 203	NG	7229	25 895	36 286	AH	102 042	12 558	33 953
G	5901	6960	12 298	AW	4248	9654	15 397	Z	27 845	5749	14 135
M	23 778	5967	14 833	AO	16 035	6970	14 496	S	43 700	5485	17 965
UH	2856	7576	11 253	W	20 352	7151	15 411	CH	4751	5118	7940
D	43 601	14 192	30 390	ER	23 525	15 131	30 264	B	15 608	7577	15 252
F	17 018	4111	8890	R	40 428	5025	16 386	K	27 866	3325	12 650
HH	13 785	7915	14 745	EY	14 695	4891	10 838	P	20 212	2496	8607
V	19 628	–	4089	DH	29 311	–	4716	IH	53 904	–	11 368
TH	3604	–	1250	OY	831	–	595	ZH	607	–	237
AX	156	–	418								

Table B.9
Hyperparameter settings of the phoneme-to-spectrogram model, including those start with ‘Feature/’ for Mel-spectrogram extraction.

Hyperparameter	
Feature/number of Mel bands	80
Feature/FFT window length	800
Feature/hop length	200
Feature/frame window size	800
Feature/preemphasis	0.97
Feature/lowest frequency	55
Feature/highest frequency	7600
Encoder/embedding dimension	512
Encoder/number of Conv layers	3
Encoder/Conv kernel size	(5,)
Encoder/Conv channel size	512
Encoder/LSTM units per direction	256
Output frames per decoding step	1
Decoder/Attention dimension	128
Decoder/Attention filters	32
Decoder/Attention kernel	(31,)
Decoder/PreNet linear layers	[256, 256]
Decoder/number of LSTM layers	2
Decoder/LSTM units	1024
Decoder/PostNet Conv layers	3
Decoder/PostNet Conv kernel size	(5,)
Decoder/PostNet Conv channel size	512

Appendix B. Hyperparameters setting for the text-to-spectrogram model

Table B.9 presents the settings for hyperparameters used in TTS model training. In the table, ‘Feature/’ refers to those parameters related to Mel-spectrogram extraction, ‘Encoder/’ refers to the network parameters for the encoder part, while ‘Decoder/’ is for the decoder part. The output frames per decoding step is set to 1 in our model training.

References

Abe, M., Shikano, K., Kuwabara, H., 1990. Cross-language voice conversion. In: 1990 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1. pp. 345–348.

Ahmed, B.H., Tan, T.-P., 2012. Automatic speech recognition of code switching speech using 1-best rescoring. In: 2012 International Conference on Asian Language Processing. pp. 137–140.

International Phonetic Association, 1999. Handbook of the International Phonetic Association: A Guide to the use of the International Phonetic Alphabet. Cambridge University Press, Cambridge, UK.

Bernardo, A.B., 2005. Bilingual code-switching as a resource for learning and teaching: Alternative reflections on the language and education issue in the Philippines. In: Linguistics and Language Education in the Philippines and beyond: A Festschrift in Honor of Ma. Lourdes S. Bautista. pp. 151–169.

Cai, W., Chen, J., Li, M., 2018. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system. In: Proc. Odyssey 2018 the Speaker and Language Recognition Workshop. pp. 74–81.

Cai, W., Chen, J., Zhang, J., Li, M., 2020a. On-the-fly data loader and utterance-level aggregation for speaker and language recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 1038–1051.

Cai, Z., Yang, Y., Li, M., 2020b. Cross-lingual multispeaker text-to-speech under limited-data scenario. arXiv preprint arXiv:2005.10441.

Cai, Z., Zhang, C., Li, M., 2020. From speaker verification to multispeaker speech synthesis, deep transfer with feedback constraint. In: Proc. Interspeech 2020. pp. 3974–3978.

Chao Weng, 2012. https://github.com/kaldi-asr/kaldi/blob/master/egs/hkust/s5/conf/pinyin2cmu.

Chen, M., Chen, M., Liang, S., Ma, J., Chen, L., Wang, S., Xiao, J., 2019. Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding. In: Proc. Interspeech 2019. pp. 2105–2109.

Chung, J.S., Nagrani, A., Zisserman, A., 2018. VoxCeleb2: Deep speaker recognition. In: Proc. Interspeech 2018. pp. 1086–1090.

Cooper, E., Lai, C., Yasuda, Y., Fang, F., Wang, X., Chen, N., Yamagishi, J., 2020. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 6184–6188.

Desplanques, B., Thienpondt, J., Demuynck, K., 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In: Proc. Interspeech 2020. pp. 3830–3834.

Ding, S., Zhao, G., Gutierrez-Osuna, R., 2020. Improving the speaker identity of non-parallel many-to-many voice conversion with adversarial speaker recognition. In: Proc. Interspeech 2020. pp. 776–780.

Du, J., Na, X., Liu, X., Bu, H., 2018. Aishell-2: Transforming mandarin ASR research into industrial scale. arXiv preprint arXiv:1808.10583.

Special Eurobarometer European Commission, 2012. Europeans and their languages. Spec. Eurobarometer 386 (June).

Fu, R., Tao, J., Wen, Z., Yi, J., Qiang, C., Wang, T., 2020. Dynamic soft windowing and language dependent style token for code-switching end-to-end speech synthesis. In: Proc. Interspeech 2020. pp. 2937–2941.

Gales, M.J., Knill, K.M., Ragni, A., 2015. Unicode-based graphemic systems for limited resource languages. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 5186–5190.

Grézl, F., Karafiát, M., Kontár, S., Cernocky, J., 2007. Probabilistic and bottle-neck features for LVCSR of meetings. In: 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Vol. 4. pp. IV–757.

Griffin, D., Lim, J., 1984. Signal estimation from modified short-time Fourier transform. IEEE Trans. Acoust. Speech Signal Process. 32 (2), 236–243.

Huang, W.-C., Hayashi, T., Watanabe, S., Toda, T., 2020. The sequence-to-sequence baseline for the voice conversion challenge 2020: Cascading ASR and TTS. In: Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020. pp. 160–164.

Hunt, A.J., Black, A.W., 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In: 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Vol. 1. pp. 373–376.

Ito, K., 2017. The LJ speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Moreno, I.L., Wu, Y., et al., 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In: Advances in Neural Information Processing Systems. pp. 4480–4490.

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., van den Oord, A., Dieleman, S., Kavukcuoglu, K., 2018. Efficient neural audio synthesis. In: Proceedings of the 35th International Conference on Machine Learning. pp. 2410–2419.

Kameoka, H., Kaneko, T., Kou, T., Hojo, N., 2019. ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder. IEEE/ACM Trans. Audio Speech Lang. Process. PP (99), 1.

Kameoka, H., Kaneko, T., Tanaka, K., Hojo, N., 2018. StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks. In: 2018 IEEE Spoken Language Technology Workshop. pp. 266–273.

de Korte, M., Kim, J., Klabbers, E., 2020. Efficient neural speech synthesis for low-resource languages through multilingual modeling. In: Proc. Interspeech 2020. pp. 2967–2971.

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W.Z., Sotelo, J., de Brébisson, A., Bengio, Y., Courville, A.C., 2019. MelGAN: Generative adversarial networks for conditional waveform synthesis. In: Advances in Neural Information Processing Systems, Vol. 32.

Lee, S., Ko, B., Lee, K., Yoo, I., Yook, D., 2020. Many-to-many voice conversion using conditional cycle-consistent adversarial networks. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 6279–6283.

Lee, Y., Shon, S., Kim, T., 2018. Learning pronunciation from a foreign language in speech synthesis networks. arXiv preprint arXiv:1811.09364.

Li, B., Zen, H., 2016. Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis. In: Interspeech 2016. pp. 2468–2472.

Li, B., Zhang, Y., Sainath, T., Wu, Y., Chan, W., 2019. Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 5621–5625.

Liu, Z., Mak, B., 2020. Multi-lingual multi-speaker text-to-speech synthesis for voice cloning with online speaker enrollment. In: Proc. Interspeech 2020. pp. 2932–2936.

Lyu, D.-C., Chng, E.-S., Li, H., 2013. Language diarization for code-switch conversational speech. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 7314–7318.

Lyu, D.-C., Lyu, R.-Y., 2008. Language identification on code-switching utterances using multiple cues. In: Ninth Annual Conference of the International Speech Communication Association.

Lyu, D.-C., Tan, T.-P., Chng, E.S., Li, H., 2010. Seame: A mandarin-english code-switching speech corpus in south-east Asia. In: Eleventh Annual Conference of the International Speech Communication Association.

Meng, Z., Zhao, Y., Li, J., Gong, Y., 2019. Adversarial speaker verification. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 6216–6220.

Ming, H., Lu, Y., Zhang, Z., Dong, M., 2017. A light-weight method of building an LSTM-RNN-based bilingual TTS system. In: 2017 International Conference on Asian Language Processing. pp. 201–205.

Morise, M., Yokomori, F., Ozawa, K., 2016. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE Trans. Inf. Syst. 99-D (7), 1877–1884.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., 2016. WaveNet: A generative model for raw audio. In: 9th ISCA Speech Synthesis Workshop. p. 125.

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: An ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 5206–5210.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The kaldi speech recognition toolkit. In: 2011 Workshop on Automatic Speech Recognition and Understanding. pp. 1–4.

Rallabandi, S., Black, A.W., 2017. On building mixed lingual speech synthesis systems. In: Proc. Interspeech 2017. pp. 52–56.

Ramani, B., Jeeva, M.P.A., Vijayalakshmi, P., Nagarajan, T., 2014. Cross-lingual voice conversion-based polyglot speech synthesizer for Indian languages. In: Proc. Interspeech 2014. pp. 775–779.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T., 2021. FastSpeech 2: Fast and high-quality end-to-end text to speech. In: 9th International Conference on Learning Representations.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.-Y., 2019. FastSpeech: Fast, robust and controllable text to speech. In: Advances in Neural Information Processing Systems, Vol. 32.

Rosenberg, A., Ramabhadran, B., 2017. Bias and statistical significance in evaluating speech synthesis with mean opinion scores. In: Proc. Interspeech 2017. pp. 3976–3980.

Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al., 2018. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4779–4783.

Shen, H.-P., Wu, C.-H., Yang, Y.-T., Hsu, C.-S., 2011. CECOS: A Chinese-english code-switching speech database. In: 2011 International Conference on Speech Database and Assessments. pp. 120–123.

Shi, Y., Bu, H., Xu, X., Zhang, S., Li, M., 2021. AISHELL-3: A multi-speaker mandarin TTS corpus. In: Proc. Interspeech 2021. pp. 2756–2760.

Sitaram, S., Rallabandi, S.K., Rijhwani, S., Black, A.W., 2016. Experiments with cross-lingual systems for synthesis of code-mixed text. In: 9th ISCA Speech Synthesis Workshop. pp. 76–81.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S., 2018. X-vectors: Robust DNN embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 5329–5333.

Staib, M., Teh, T.H., Torresquintero, A., Mohan, D.S.R., Foglianti, L., Lenain, R., Gao, J., 2020. Phonological features for 0-shot multilingual speech synthesis. In: Proc. Interspeech 2020. pp. 2942–2946.

Sun, L., Wang, H., Kang, S., Li, K., Meng, H.M., 2016. Personalized, cross-lingual TTS using phonetic posteriorgrams. In: Proc. Interspeech 2016. pp. 322–326.

Tan, X., Qin, T., Soong, F., Liu, T.-Y., 2021a. A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561.

Tan, Z., Wei, J., Xu, J., He, Y., Lu, W., 2021b. Zero-shot voice conversion with adjusted speaker embeddings and simple acoustic features. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 5964–5968.

2015. The Carnegie Mellon Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Titus, A., Silovsky, J., Chen, N., Hsiao, R., Young, M., Ghoshal, A., 2020. Improving language identification for multilingual speakers. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 8284–8288.

Tobing, P.L., Wu, Y.-C., Hayashi, T., Kobayashi, K., Toda, T., 2019. Non-parallel voice conversion with cyclic variational autoencoder. In: Proc. Interspeech 2019. pp. 674–678.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008.

Veaux, C., Yamagishi, J., MacDonald, K., et al., 2016. Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit.

Vu, N.T., Lyu, D.-C., Weiner, J., Telaar, D., Schlippe, T., Blaicher, F., Chng, E.-S., Schultz, T., Li, H., 2012. A first speech recognition system for mandarin-english code-switch conversational speech. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4889–4892.

Wang, Y., Stanton, D., Zhang, Y., Ryan, R.-S., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., Saurous, R.A., 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In: Proceedings of the 35th International Conference on Machine Learning. pp. 5180–5189.

Xin, D., Komatsu, T., Takamichi, S., Saruwatari, H., 2021. Disentangled speaker and language representations using mutual information minimization and domain adaptation for cross-lingual TTS. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 6608–6612.

Xue, L., Song, W., Xu, G., Xie, L., Wu, Z., 2019. Building a mixed-lingual neural TTS system with only monolingual data. In: Proc. Interspeech 2019. pp. 2060–2064.

Yang, J., He, L., 2020. Towards universal text-to-speech. In: Proc. Interspeech 2020. pp. 3171–3175.

Yi, Z., Huang, W.-C., Tian, X., Yamagishi, J., Das, R.K., Kinnunen, T., Ling, Z.-H., Toda, T., 2020. Voice conversion challenge 2020 – intra-lingual semi-parallel and cross-lingual voice conversion . In: Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020. pp. 80–98.

Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Commun. 51 (11), 1039–1064.

Zhang, Y., Weiss, R.J., Zen, H., Wu, Y., Chen, Z., Skerry-Ryan, R., Jia, Y., Rosenberg, A., Ramabhadran, B., 2019. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. In: Proc. Interspeech 2019. pp. 2080–2084.

Zhao, S., Nguyen, T.H., Wang, H., Ma, B., 2020. Towards natural bilingual and code-switched speech synthesis based on mix of monolingual recordings and cross-lingual voice conversion. In: Proc. Interspeech 2020. pp. 2927–2931.

Zhao, S., Wang, H., Nguyen, T.H., Ma, B., 2021. Towards natural and controllable cross-lingual voice conversion based on neural TTS model and phonetic posteriorgram. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 5969–5973.

Zheng, H., Cai, W., Zhou, T., Zhang, S., Li, M., 2016. Text-independent voice conversion using deep neural network based phonetic level features. In: 2016 23rd International Conference on Pattern Recognition. pp. 2872–2877.

Zhou, X., Tian, X., Lee, G., Das, R.K., Li, H., 2020. End-to-end code-switching TTS with cross-lingual language model. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 7614–7618.

Zhou, Y., Tian, X., Xu, H., Das, R.K., Li, H., 2019. Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 6790–6794.