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*

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 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 uses shared phonemic representations associated with language tokens. As for the data-limited scenario, we adopt a framework cascading several speech modules to achieve our goal. In particular, we proposed a parallel non-autoregressive voice conversion module to address multi-speaker synthesis for data-insufficient cases. Both approaches use limited bilingual data and demonstrate impressive performance in cross-lingual synthesis, where we can generate fluent foreign speech, even code-switching speech, for monolingual speakers. Moreover, experimental results show that our proposed voice conversion module can well maintain the voice characteristics in data-limited cases.

Keywords: Text-to-speech, Cross-lingual speech synthesis, Voice conversion, Speaker verification

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 [1, 2]. This type of TTS framework outperforms traditional frameworks like statistical parametric speech synthesis (SPSS) [3], and concatenative speech synthesis [4]. It soon becomes the state-of-the-art framework for speech synthesis and is widely applied in various TTS applications (e.g., audiobook reader, virtual assistants, navigation systems, etc.) in our daily lives.

Nonetheless, this kind of model, like vanilla Tacotron2 [2] and Fastspeech [5, 6], keeps a certain level of limitations in controllability regarding latent speech attributes when it is proposed. It renders the speech attributes in an implicit way like it learns to model those attributes during

*Corresponding Author: Ming Li
Email addresses: zexin.cai@duke.edu (Zexin Cai), ming.li369@duke.edu (Ming Li)

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 introducing global style tokens (GSTs) to the Tacotron2 in an unsupervised way [7]. This allows the model to control speaking speed or clone speaking style via GSTs. Some research works extend the Tacotron2 with conditioned features extracted from a speaker verification system to achieve speaker identity cloning for multi-speaker TTS. [8, 9].

On the other hand, language is an important attribute for multilingual speech synthesis. As bilinguists and multilinguists are commonly seen in today's world, the speech communication scenario becomes more complicated. It is essential for speech analysis tools, including speech recognition and speech synthesis, to adapt to this change to maintain their current performance [10]. The challenge is that languages, generally, have different grapheme sets and pronunciations among each other. This challenge motivates researchers to find and investigate shared representations between languages for speech analysis [11, 12, 13]. Even with appropriate representations for multiple languages, the model architecture needs to be upgraded in order to achieve multilingual processing for most speech analysis systems [13]. There are existing studies of multilingual synthesis and cross-lingual synthesis that are based on classical statistical parametric speech synthesis (SPSS) [14, 15]. Nevertheless, the synthesis performance is restricted by the relative complex pipeline and the vocoder in terms of SPSS approaches [1]. As the end-to-end TTS models can generate speech with higher fidelity compared with classical methods, extensions on the end-to-end TTS frameworks are explored for multilingual modeling as well [16, 17, 18, 19]. 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 [17]. 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 [20].

In this paper, we aim to achieve cross-lingual multi-speaker TTS from two languages, English and Mandarin. While it is costly to collect a huge dataset for multi-speaker cross-lingual synthesis, addressing the cross-lingual synthesis under low-resource data scenarios is essential. Therefore we also investigate solutions to such scenario in this paper. Two synthesis frameworks are proposed for two different scenarios, which are data-sufficient scenario and low-resource scenario, respectively. In the data-sufficient scenario, we propose a Tacotron-based model conditioned on speaker embedding and language tokens. The relevant pronunciations between languages are associated by shared phonetic inputs. The proposed model can generate high-fidelity speech for all speakers with respect to their own 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 the bilingual speaker to monolingual speakers, which enables 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 total recording of each speaker is less than half an hour. Under 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, and a multi-speaker voice conversion system. Furthermore, we propose a parallel non-autoregressive network for the multi-speaker voice conversion module. The adversarial speaker classifier [21] and the speaker embedding consistency loss [9] are employed in the conversion network to im-

prove the speaker similarity. We conduct objective evaluation and subjective evaluation on the synthesis performance. Results show that the VC system can generate high-fidelity speech with satisfactory speaker similarity. Both systems under two scenarios can tackle code-switching synthesis. Audio samples are available online for listening [1]. The contribution of our paper includes:

- We investigate multi-speaker cross-lingual speech synthesis in two multilingual data setups.
- We propose a TTS framework that uses shared phonetic representations and language tokens for cross-lingual synthesis.
- 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 Multi-Lingual 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 [8, 22], 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) [14]. They use the international pronunciation Alphabet (IPA) [23] as the input representation and applied cluster adaptive language networks for generating the language-dependent linguistic features, followed by speaker-dependent output layers for different voices. Ming et al. present a light-weighted bilingual synthesis system that adopts concatenated vectors in the linguistic-feature level to manage two languages in one model. [15] .

More recently, Li et al. propose a novel representation for all languages [13]. This representation, called Bytes, allows speech recognition models and speech synthesis models to achieve multilingual processing. The performance of using Bytes in TTS is conducted and evaluated by another group of researchers [17]. Experimental results in [17] show that using phoneme units as the input for the MLMS TTS model could achieve better synthesis performance than using Bytes. With sufficient training data (more than 500 hours), their proposed model is able to achieve cross-lingual synthesis with a high naturalness rate. The shared phoneme input is one of

[1] https://caizexin.github.io/mlms-syn-samples/index.html

the keys to the cross-lingual synthesis, which is also stated in [16]. The study reveals that similar pronunciations across languages result in close linguistic embedding vectors.

We also propose a TTS framework using shared phonetic representations for cross-lingual multi-speaker speech synthesis, and it is archived at [24] [2]. After that, 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 [20], which has a similar structure as our proposed model. However, we have the language-dependent Tacotron encoder designed for allowing the TTS model to synthesized code-switching text. Zhou et al. present a novel method to merge context information between languages by adopting word embedding from a pre-trained language model. Nevertheless, The cross-lingual synthesized speech has moderate quality, as shown in the figures from [18]. Fu et al. present a code-switching speech synthesis system based on a language-dependent style token [25]. 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 [25] 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 low-resource languages recently. Marlene et al. investigate phonological features that could adapt to untrained languages with zero-shot adaptation [26]. Similarly, Korte et al. look into how different strategies work for low-resource language synthesis with data from rich-source languages [27]. In our paper, we pay attention to the low-resource scenario when each speaker contains limited data for training.

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 [28, 29, 30, 31].

The cross-lingual VC is nonparallel in nature. Since the source speaker and the target speaker 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 speak [32]. Vector quantization (VQ)-based method is used for cross-lingual VC between Japanese and English [33]. 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 polyglot speech corpus [34]. 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 [35, 36, 37, 38], which takes advantage of the linguistic

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

information from a large amount of speech data, also achieves 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 [36]. Those aforementioned methods focus on one-to-one cross-lingual VC and use the conventional vocoder WORLD [39] 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 [40]. Different from [40], 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-hour 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. [41]. The language environment in globalization inspires more and more bilinguists and multilinguists, 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 [42, 43], followed with various approaches proposed to address complicated speech analysis, including multilingual automatic speech recognition (ASR), language identification, and language diarization with respect to multilingual scenario [44, 45, 46, 47]. Likewise, TTS systems need to be improved for synthesizing natural speech for code-switching sentences [18].

One of the main challenges of code-switching TTS is that the grapheme set or the phoneme set between languages are 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, which are proposed for addressing multilingual issues in TTS, indicate that shared input representation across languages is one of the keys to realizing cross-lingual synthesis [13, 14, 16]. The shared representations include shared phoneme set, international pronunciation alphabet (IPA), and the Bytes coding [13], where the phoneme representation can obtain better performance [17].

In this paper, we choose to use a shared phoneme set from CMU dictionary [48] to investigate bilingual multi-speaker TTS and cross-lingual synthesis between Mandarin and English under a speaker-limited data-sufficient scenario. As for Mandarin, the pronunciation representation called pinyin can be converted to CMU phoneme by the pinyin-to-cmu mapping table [49]. Since Mandarin is a tone-language, digits 1 to 6 are used to denote different tones, while '0', '1', '2' are used to mark the lexical stress for English. Although the tone and stress share the same annotations in our input, which may cause ambiguity, we have language identification tokens as another input stream. Moreover, language identification tokens are used to generate language-dependent encoding features while preserving the shared information between languages, like close pronunciations. Similarly, '0', '1', '2' are used for language identification in our input

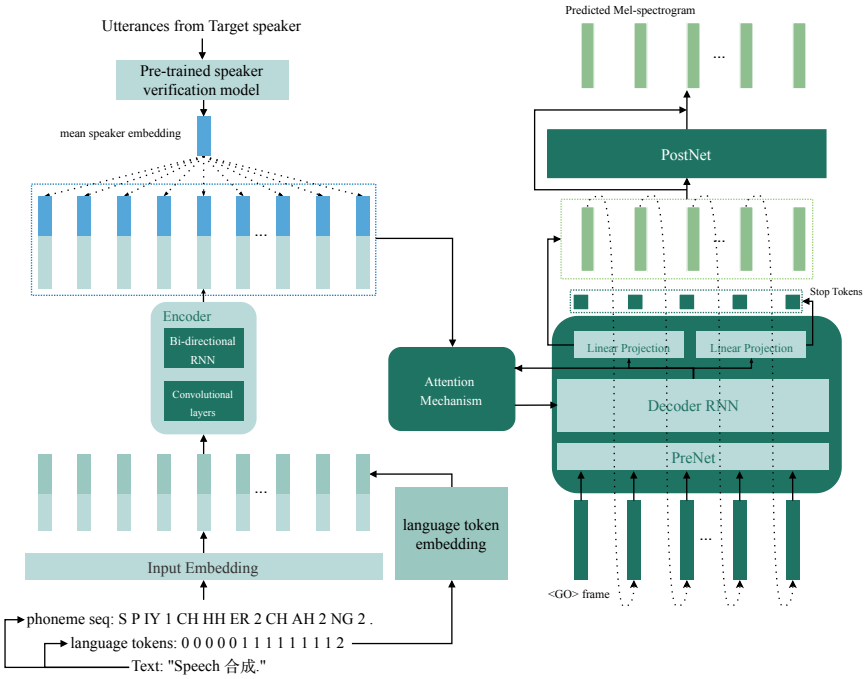


Figure 1: Proposed multilingual multispeaker TTS model

representations, where '0' represents the corresponding phoneme or stress annotation is from English, '1' is for Mandarin, and '2' 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 language identification tokens '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 figure 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 language tokens 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 layers of 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 speaker verification model

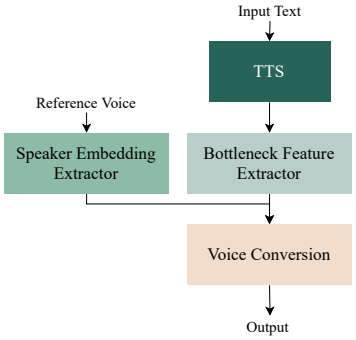


Figure 2: Synthesis pipeline for data-insufficient scenario

[50] by feeding all training utterances of each speaker. We believe that 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.

3.3. Vocoder

The vocoder is participated in TTS systems to transform the acoustic features back to audio signals in the time domain. Both Griffin-Lim [51] and neural vocoders [52, 53, 54] can be applied in our framework to reconstruct the waveform. In this work, we use MelGAN [54] as our vocoder for the proposed methods in both scenarios, since MelGan is much faster in spectrogram inversion 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. 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 cross-lingual synthesis is poor. It is not easy to generate accurate speech with foreign text for monolingual speakers. For example, the system is unable to synthesis 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 figure 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.

4.1. Bottleneck Feature Extractor

The intermediate linguistic feature 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 [55]. 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 with respect to context models like language models. The key module that 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. 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, which is also known as the bottleneck layer [56]. 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 for voice conversion. On the other hand, to extract language-independent features that work for multilingual scenarios, the acoustic model is trained with multilingual data.

4.2. Speaker Embedding Extractor

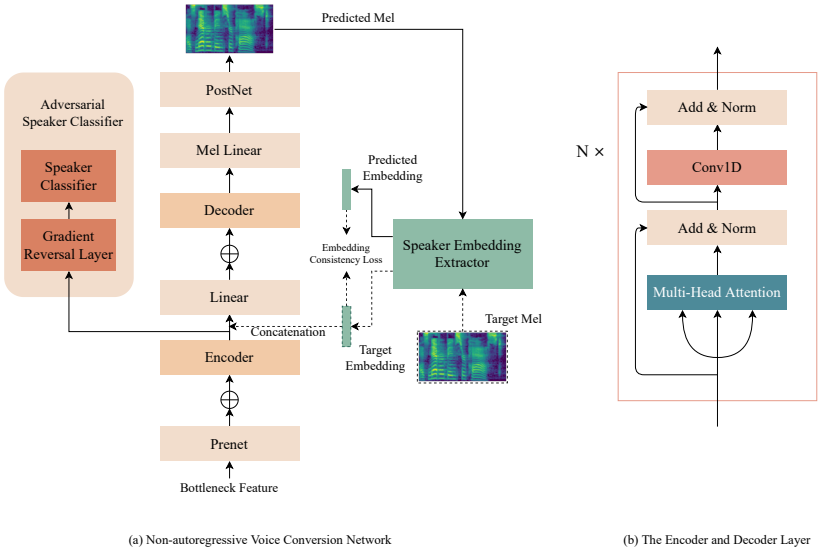
The speaker embedding extractor comes from models designed for speaker verification tasks. Speaker verification is the task of identifying persons from their voices. Recently, deep learning has revolutionized the speaker verification field. X-vector based systems [57] and its variant frameworks [58, 59] have become the most popular architectures in speaker verification. Normally, the speaker verification contains three components: a front-end pattern extractor, an encoder layer, and a back-end classifier. The fully connected layer is named speaker embedding, which is used as the 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.

4.3. Cross-lingual Voice Conversion

We propose a non-autoregressive model for voice conversion. The framework is a variant of the synthesis network FastSpeech [5]. The framework is shown in figure 3. We remove the length regulator module since the input sequence and the output sequence can share the same length in the VC task. The network is composed of a speaker encoder, encoder-decoder structure with multi-head attention mechanism and an adversarial speaker classifier.

Note that FastSpeech is first proposed for converting text-embedding sequence to acoustic features, while VC is to convert linguistic features 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 [60] 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 comparing to autoregressive models. Finally, a Post-Net module consisting of 5-layer convolution is added to obtain the residual coefficients from the predicted feature to improve the overall reconstruction quality.

The bottleneck feature, which is extracted from source speech, has been proven to contain voice characteristics from the source speaker in many-to-many VC systems [61]. To further



(a) Non-autoregressive Voice Conversion Network (b) The Encoder and Decoder Layer

Figure 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.

eliminate speaker information and prevent the converted voice from resembling the voice of the source speaker, we employ an adversarial speaker classifier in our proposed framework [21, 61]. The adversarial module contains a gradient reversal layer and a speaker classifier. Both are linear layers, while the latter one 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 [9] 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. Hyper-parameter 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 [62] 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

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

Phoneme	LJS	DB-1	DB-4	Phoneme	LJS	DB-1	DB-4	Phoneme	LJS	DB-1	DB-4
J	-	10088	12499	X	-	8050	11895	Q	-	5435	7489
IY	28587	54859	85601	EH	26397	3598	11791	AA	16976	11173	23205
L	32893	9420	23510	AY	12079	7479	15619	UW	15345	30630	44593
SH	7957	11456	17804	OW	10201	6921	13698	Y	4426	16540	27793
N	68392	33006	56359	T	65657	8698	26504	JH	4824	8994	13821
AE	21502	27640	42203	NG	7229	25895	36286	AH	102042	12558	33953
G	5901	6960	12298	AW	4248	9654	15397	Z	27845	5749	14135
M	23778	5967	14833	AO	16035	6970	14496	S	43700	5485	17965
UH	2856	7576	11253	W	20352	7151	15411	CH	4751	5118	7940
D	43601	14192	30390	ER	23525	15131	30264	B	15608	7577	15252
F	17018	4111	8890	R	40428	5025	16386	K	27866	3325	12650
HH	13785	7915	14745	EY	14695	4891	10838	P	20212	2496	8607
V	19628	-	4089	DH	29311	-	4716	IH	53904	-	11368
TH	3604	-	1250	OY	831	-	595	ZH	607	-	237
AX	156	-	418								

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

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 hours of English audio-transcript pairs recorded by a female English native speaker. The DB-1 has approximately 12 hours of Mandarin speech synthesis data recorded by a female Mandarin native speaker. The DB-4 is a bilingual dataset containing 12 hours of Chinese audio-transcript pairs, 6 hours of English pairs, and 6 hours of code-switching data from a female Mandarin speaker.

The frequencies of all phonemes in the three datasets are shown in table 1. 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 existed in the English dataset, as shown in the table. The bilingual dataset DB-4 contains all phonemes. 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 compared to those 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 audios are downsampled to 16 kHz. The hyperparameters setting for acoustic feature extraction, network components are shown in table 2. In the table, 'Feature/' refers to those parameters related to Mel-spectrogram extraction, 'Encoder/' refers to the network parameters for the encoder part, while 'Encoder/' is for the decoder part. We set the output frames per decoding step to 1 in our model training.

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

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

Table 2: Hyperparameters of the phoneme-to-spectrogram model, including those start with 'Feature/' for Mel-spectrogram extraction.

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 concerning naturalness, similarity, and intelligibility. The naturalness is related to the quality of synthesized audios regardless of the content. The speaker similarity score measures how close the synthesized voice is to the expected speaker, while the intelligibility evaluates the clarity level of the speech content. 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.

The naturalness mean opinion scores (MOS) are shown in table 3. As shown in the table, the quality of synthesized audios varies among different systems and different speakers. Generally, the quality reaches around 4 when synthesizing audio in the target speaker's native language, while the performance degrades when generating cross-lingual speech for monolingual speakers. For example, DB-1 obtains MOS with 4.14 when synthesizing Mandarin sentences, but the score degrades to 3.12 for English sentences. In addition, as shown by the similarity scores on the table, 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 clearly observed from the table 3. Although BLMS can achieve bilingual multi-speaker synthesis, the cross-lingual synthesis performance is poor, which matches the result from [17]. The

MOS±95%CI	The data-sufficient scenario						
	BLMS			CLMS			
	DB-1	LJS	ALL	DB-1	LJS	DB-4	ALL
Naturalness	3.41±0.07	3.21±0.07	3.31±0.05	3.89±0.06	3.41±0.06	3.99±0.05	3.76±0.04
CN	3.97±0.1	2.73±0.12	3.35±0.1	4.01±0.1	3.02±0.11	3.99±0.1	3.68±0.07
EN	2.86±0.13	3.86±0.09	3.36±0.09	3.86±0.08	3.96±0.08	4.04±0.08	3.95±0.05
CS	3.4±0.11	3.05±0.11	3.22±0.08	3.81±0.1	3.24±0.1	3.95±0.1	3.66±0.06
Intelligibility	3.27±0.1	3.16±0.09	3.21±0.07	4.37±0.05	3.86±0.07	4.47±0.04	4.23±0.03
CN	4.58±0.06	2.38±0.13	3.48±0.12	4.64±0.06	**3.54±0.13**	4.65±0.06	4.28±0.06
EN	1.83±0.12	4.17±0.1	3.0±0.13	**4.17±0.09**	4.37±0.08	4.37±0.08	4.3±0.05
CS	3.4±0.1	2.92±0.13	3.16±0.08	**4.29±0.09**	**3.68±0.12**	4.41±0.07	4.13±0.06
Similarity	3.92±0.06	3.35±0.06	3.64±0.04	4.09±0.04	3.44±0.06	4.12±0.04	3.88±0.03
CN	4.25±0.08	3.16±0.1	3.7±0.08	4.21±0.07	3.18±0.1	4.16±0.08	3.85±0.06
EN	3.37±0.11	3.64±0.09	3.51±0.07	3.91±0.07	3.84±0.09	4.08±0.08	3.94±0.05
CS	4.14±0.07	3.26±0.1	3.7±0.07	4.13±0.08	3.31±0.1	4.11±0.08	3.85±0.06

Table 3: 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 contains both Mandarin and English.

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, when involving a bilingual dataset, the system CLMS is able to generate cross-lingual speech, even in code-switching cases, with intelligible pronunciations for monolingual speakers. The cross-lingual intelligibility MOS is significantly improved in this case, where the 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 the synthesized speech is exactly like a foreign speaker speak 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.

5.1.3. Alignments

In addition, the cross-lingual synthesis performance also can be seen from the attention alignments in Figure 4. The synthesized content is a code-switching sentence. For system BLMS, we can observe clear breaks when the language switches in the sentence for monolingual speakers DB-1 and DB-4 in figure 4 (a) and (b). However, the attention alignments obtained from CLMS are consistent even for monolingual speakers. This further implies that the cross-lingual knowledge and pronunciation fluency can be transferred from the bilingual speaker to monolingual speakers with our proposed model.

5.2. Data-insufficient Utterance-limited Scenario

5.2.1. The bottleneck extractor

The English dataset Librispeech [63] and the Mandarin dataset AISHELL-2 [64] are used to train our bilingual bottleneck extractor. The receipt that is used to train Librispeech in Kaldi is used for our model training. The acoustic model, known 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

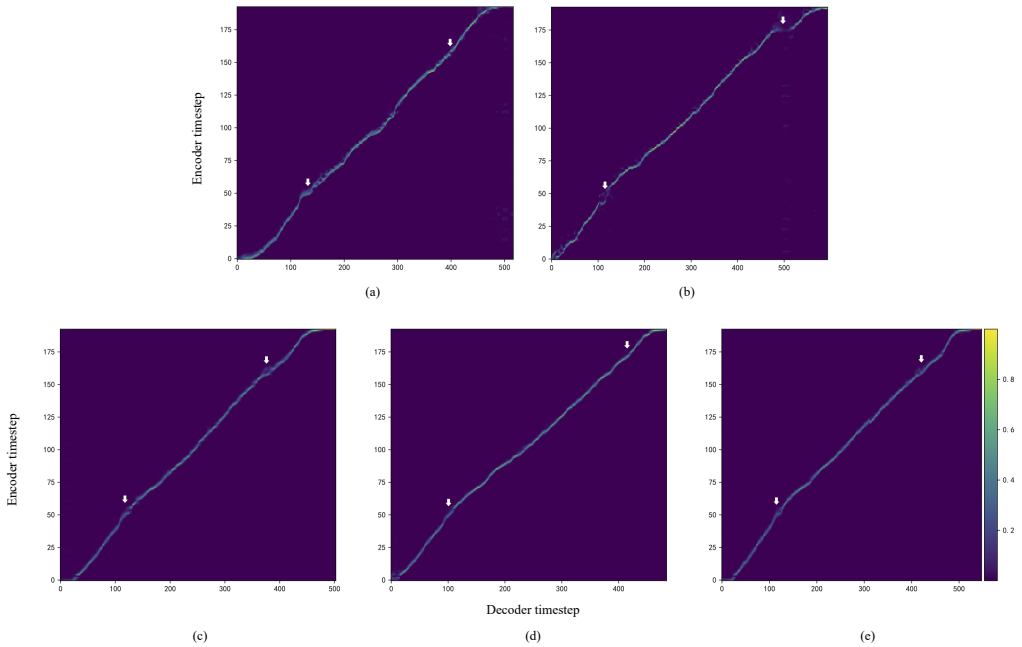


Figure 4: Attention alignments when synthesizing code-switching text '其实我很难判断 in my heart I think my Chinese is better but people tell me that my English 是比较好' (Actually, it's hard for me to tell. In my heart, I think my Chinese is better, but people tell me that my English is better): (a) The alignment from BLLS with speaker DB-1; (b) The alignment from BLLS with speaker LJS; (c) The alignment from CLMS with speaker DB-1; (d) The alignment obtained from CLMS with speaker LJS; (e) The alignment obtained from CLMS with speaker DB-4;

includes 39 English phonemes and 52 Mandarin phonemes. We use 50ms window-length and 12.5 ms frame-shift for MFCC feature extraction, which is the same setting as in TTS. In table 2, the hop length with 200 samples for input audios in 16k Hz 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 4, 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 extraction.

5.2.2. Speaker embedding extractor

The VoxCeleb2 [65] dataset is used to pre-train the speaker verification system. However, our training dataset for the voice conversion system is from a different domain (cross-lingual and cross-dataset). To obtain discriminative speaker embeddings for our cross-lingual setting, we fine-tune the pre-trained speaker verification model with data from the same domain as our voice conversion training datasets. The ECAPA-TDNN [58] model is adopted as our speaker verification model, which is also used during voice conversion training as we shown in section 4.3. The AISHELL-2 and VCTK dataset [66] are used to fine-tune the verification model. There are a total of 1901 speakers with more than 900 000 utterances from the AISHELL-2 database

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

Table 4: The ASR performance of the bottleneck extractor

MOS±95%CI	The data-insufficient scenario		
	Voice Conversion		
	VCTK	AISHELL	ALL
Naturalness	3.59±0.04	3.48±0.04	3.54±0.03
CN	3.38±0.07	3.4±0.07	3.39±0.05
EN	3.9±0.06	3.63±0.06	3.76±0.04
CS	3.5±0.07	3.42±0.07	3.46±0.05
Intelligibility	4.11±0.04	4.06±0.04	4.08±0.03
CN	4.1±0.07	4.12±0.07	4.11±0.05
EN	4.23±0.06	4.1±0.06	4.16±0.04
CS	4.0±0.07	3.95±0.07	3.98±0.05
Similarity	4.05±0.03	3.64±0.04	3.85±0.03
CN	3.95±0.06	3.6±0.07	3.78±0.04
EN	4.16±0.05	3.66±0.07	3.91±0.05
CS	4.05±0.05	3.68±0.07	3.86±0.04

Table 5: The mean opinion scores (MOS) with 95% confidence interval (CI) for all proposed systems under the data-insufficient scenario. VCTK is evaluated from voices chosen from the VCTK dataset, and AISHELL is from voices chosen from the AISHELL3 dataset. For synthesis type, CN denotes Mandarin sentences, EN denotes English sentences, and CS denotes code-switching sentences that contains both Mandarin and English.

for fine-tuning. Another subset with 100 speakers from AISHELL-2 is randomly chosen and used as the test set to evaluate the verification performance. Normally, the speaker verification performance is measured by the equal error rate (EER) and the minimum detection cost function (mDCF). The speaker verification system fine-tuned on the in-domain datasets achieves an EER with 2.18% on the test set, and the mDCF on the test set is about 0.4. The result shows that more than 99% utterance-pairs we constructed from the test set are correctly verified. Therefore, the verification system we trained is able to extract discriminative representations.

5.2.3. Multilingual multi-speaker voice conversion

Three public available datasets is used in our experiments, including the LJ Speech (LJS) dataset [62] introduced in section 5.1, the VCTK English dataset [66] and the AISHELL-3 Mandarin dataset [67]. 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 for training. All audios are downsampled to 16 kHz. Settings for extracting Mel-spectrogram are the same as the one used in the utterance-limited scenario (Table 2). Hyper-parameters λ, α are set to 1.0 and 5.0, respectively. 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 voice conversion evaluation.

MOS	w/o ECL	with ECL
seen	3.73 ± 0.075	3.75 ± 0.078
unseen	3.65 ± 0.088	3.75 ± 0.086

Table 6: Speaker similarity MOS results on systems with and without ECL

Results are shown in table 5. Regarding naturalness, the voice conversion performance is not as good as the systems from the data-sufficient scenario. Comparing to the CLMS system, the overall MOS degrades from 3.76 to 3.54. However, the intelligibility remains outstanding. Speaker similarity is the most significant criterion for voice conversion. As shown in the table, the overall speaker similarity MOS is around 3.85, which indicates that the converted voice is highly close to the original voice.

5.2.4. 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. By conditioning speaker representations from embedding extractors, our proposed model enables zero-shot conversion even for unseen voices. In our experiments, we convert utterances from test sets to speech 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, Librispeech, and LJS. Then for both seen and unseen cases, we synthesize around 10 000 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.
- Code-switching scenario: convert code-switching utterances (from DB4) to monolingual speakers' voices.

In terms of voice conversion between seen speakers, 48 utterances are randomly selected, with 16 from each scenario, for subjective evaluation. Similarly, 24 utterances, with 8 from each scenario, are chosen for evaluation regarding unseen voice conversion. The similarity MOS results are shown in table 6. The similarity performances of systems with and without ECL are close. Both systems obtain 3.7 on speaker similarity. Likewise, there are no significant differences between seen speakers and unseen speakers on voice cloning performance.

We do not observe significant improvement in speaker similarity from the subjective evaluation. However, the system with ECL demonstrates better spoofing capability from objective evaluation. 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 verification. Nevertheless, we use a different

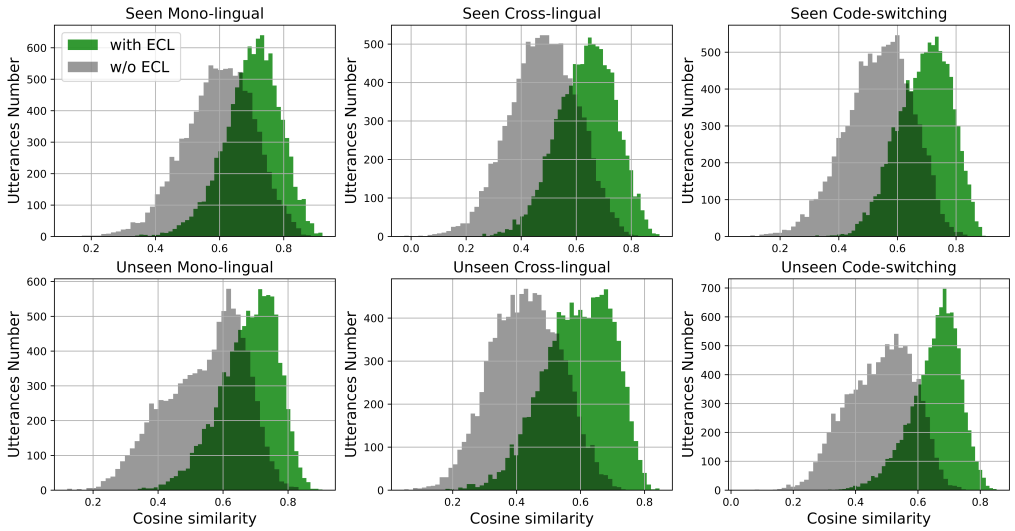


Figure 5: The distribution of cosine similarity scores between speaker embeddings from the reference speech and the converted speech in our experiments

model for objective evaluation. A ResNet-based speaker verification model [59] is trained to evaluate the verification performance between ground-truth voices and converted voices in this experiment. The verification model is trained on the VoxCeleb2 dataset and achieves an EER with 1.64% on the AISHELL-2 test set.

For each synthesized utterance in this experiment, we extract the speaker embedding of the converted result and the one of 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 figure 5. The score distribution shows that voice conversion system with ECL achieves higher similarity comparing to the one without ECL. For all scenarios, the mean of similarity scores of system with ECL is larger than the mean of scores from the system without ECL. For the system with ECL, the mean of similarity scores is larger than 0.6. This indicates that 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.

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 synthesis 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 is also

commonly existed in many prior studies [17, 20], even when there is a large amount of data for training. In today's world, multilingual speakers are only a small part of the world population, especially for minority languages. Besides, some languages do not share many similar pronunciation units as we do in our experiments. Regarding those issues, developing a cross-lingual system for those challenging languages becomes an arduous task. As future work, the study towards a universal speech synthesis system is vital, which has already started [68].

For the utterance-limited scenario, cascading several speech modules is one of the ways that achieves high-quality synthesis [69, 70]. Such a synthesis pipeline requires more computation resources and time. In addition, the robustness of the pipeline is poor as we need per-system adaptation for almost all speech models from the pipeline to adapt to novel languages or speakers. To address those issues, universal speech recognition and zero-shot multi-speaker voice conversion are essential. However, as we shown in Section 5.2.4, although synthetic voice may spoof machines, human can distinguish synthetic voices from the true voices. One of the reason 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 language tokens 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 the adding of embedding consistency loss does not achieve much improvement from the human perspective, while it significantly improves speaker similarity from the speaker verification system's perspective.

References

[1] X. Tan, T. Qin, F. Soong, T.-Y. Liu, A Survey on Neural Speech Synthesis, arXiv preprint arXiv:2106.15561.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al., Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4779–4783.

[3] H. Zen, K. Tokuda, A. W. Black, Statistical Parametric Speech Synthesis, speech communication 51 (11) (2009) 1039–1064.

[4] A. J. Hunt, A. W. Black, 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.

[5] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, T.-Y. Liu, FastSpeech: Fast, Robust and Controllable Text to Speech, in: Advances in Neural Information Processing Systems, Vol. 32, 2019.

[6] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, T. Liu, FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, in: 9th International Conference on Learning Representations, 2021.

[7] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, R. A. Saurous, Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis, in: Proceedings of the 35th International Conference on Machine Learning, 2018, pp. 5180–5189.

[8] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu, et al., Transfer Learning from Speaker Verification to Multispeaker Text-to-speech Synthesis, in: Advances in neural information processing systems, 2018, pp. 4480–4490.

[9] Z. Cai, C. Zhang, M. Li, From Speaker Verification to Multispeaker Speech Synthesis, Deep Transfer with Feedback Constraint, in: Proc. Interspeech 2020, pp. 3974–3978.

[10] S. Rallabandi, A. W. Black, On Building Mixed Lingual Speech Synthesis Systems, in: Proc. Interspeech 2017, pp. 52–56.

[11] H. of the International Phonetic Association, et al., A Guide to the Use of the International Phonetic Alphabet,(1999), The Press Syndicate of the University of Cambridge.

[12] M. J. Gales, K. M. Knill, A. Ragni, Unicode-based Graphemic Systems for Limited Resource Languages, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5186–5190.

[13] B. Li, Y. Zhang, T. Sainath, Y. Wu, W. Chan, 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.

[14] B. Li, H. Zen, Multi-Language Multi-Speaker Acoustic Modeling for LSTM-RNN Based Statistical Parametric Speech Synthesis, in: Interspeech 2016, pp. 2468–2472.

[15] H. Ming, Y. Lu, Z. Zhang, M. Dong, A Light-weight Method of Building An LSTM-RNN-based Bilingual TTS System, in: 2017 International Conference on Asian Language Processing, pp. 201–205.

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

[17] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, B. Ramabhadran, Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning, in: Proc. Interspeech 2019, pp. 2080–2084.

[18] X. Zhou, X. Tian, G. Lee, R. K. Das, H. Li, 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.

[19] M. Chen, M. Chen, S. Liang, J. Ma, L. Chen, S. Wang, J. Xiao, Cross-lingual, Multi-speaker Text-to-speech Synthesis Using Neural Speaker Embedding, Proc. Interspeech 2019 2105–2109.

[20] Z. Liu, B. Mak, Multi-Lingual Multi-Speaker Text-to-Speech Synthesis for Voice Cloning with Online Speaker Enrollment, in: Proc. Interspeech 2020, pp. 2932–2936.

[21] Z. Meng, Y. Zhao, J. Li, Y. Gong, Adversarial Speaker Verification, 2019 IEEE International Conference on Acoustics, Speech and Signal Processing 6216–6220.

[22] E. Cooper, C. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, J. Yamagishi, 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.

[23] I. P. Association, Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet., Cambridge University Press, Cambridge, U.K, 1999.

[24] Z. Cai, Y. Yang, M. Li, Cross-lingual Multispeaker Text-to-Speech under Limited-Data Scenario, arXiv preprint arXiv:2005.10441.

[25] R. Fu, J. Tao, Z. Wen, J. Yi, C. Qiang, T. Wang, Dynamic Soft Windowing and Language Dependent Style Token for Code-Switching End-to-End Speech Synthesis, in: Proc. Interspeech 2020, pp. 2937–2941.

[26] M. Staib, T. H. Teh, A. Torresquintero, D. S. R. Mohan, L. Foglianti, R. Lenain, J. Gao, Phonological Features for 0-Shot Multilingual Speech Synthesis, in: Proc. Interspeech 2020, pp. 2942–2946.

[27] M. de Korte, J. Kim, E. Klabbers, Efficient Neural Speech Synthesis for Low-Resource Languages Through Multilingual Modeling, in: Proc. Interspeech 2020, pp. 2967–2971.

[28] H. Kameoka, T. Kaneko, T. Kou, N. Hojo, ACVAE-VC: Non-Parallel Voice Conversion with Auxiliary Classifier Variational Autoencoder, IEEE/ACM Transactions on Audio, Speech, and Language Processing PP (99) (2019) 1–1.

[29] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, T. Toda, Non-Parallel Voice Conversion with Cyclic Variational Autoencoder, in: Proc. Interspeech 2019, pp. 674–678.

[30] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, StarGAN-VC: Non-parallel Many-to-many Voice Conversion Using Star Generative Adversarial Networks, in: 2018 IEEE Spoken Language Technology Workshop, pp. 266–273.

[31] S. Lee, B. Ko, K. Lee, I. Yoo, D. Yook, 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.

[32] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z.-H. Ling, T. Toda, 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.

[33] M. Abe, K. Shikano, H. Kuwabara, Cross-language Voice Conversion, in: 1990 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 345–348 vol.1.

[34] B. Ramani, M. P. A. Jeeva, P. Vijayalakshmi, T. Nagarajan, Cross-lingual Voice Conversion-based Polyglot Speech Synthesizer for Indian Languages, in: Proc. Interspeech 2014, pp. 775–779.

[35] H. Zheng, W. Cai, T. Zhou, S. Zhang, M. Li, Text-independent Voice Conversion Using Deep Neural Network Based Phonetic Level Features, in: 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 2872–2877.

[36] L. Sun, H. Wang, S. Kang, K. Li, H. M. Meng, Personalized, Cross-Lingual TTS Using Phonetic Posteriorgrams, in: Proc. Interspeech 2016, pp. 322–326.

[37] Y. Zhou, X. Tian, H. Xu, R. K. Das, H. Li, 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.

[38] S. Zhao, H. Wang, T. H. Nguyen, B. Ma, 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, 2021, pp. 5969–5973.

[39] M. Morise, F. Yokomori, K. Ozawa, WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications, IEICE Trans. Inf. Syst. 99-D (7) (2016) 1877–1884.

[40] Z. Tan, J. Wei, J. Xu, Y. He, W. Lu, 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.

[41] A. B. Bernardo, Bilingual Code-switching as A Resource for Learning and Teaching: Alternative Reflections on The Language and Education Issue in The Philippines, Linguistics and language education in the Philippines and beyond: A Festschrift in honor of Ma. Lourdes S. Bautista (2005) 151–169.

[42] D.-C. Lyu, T.-P. Tan, E. S. Chng, H. Li, Seame: A Mandarin-English Code-switching Speech Corpus in South-east Asia, in: Eleventh Annual Conference of the International Speech Communication Association, 2010.

[43] H.-P. Shen, C.-H. Wu, Y.-T. Yang, C.-S. Hsu, CECOS: A Chinese-English Code-switching Speech Database, in: 2011 International Conference on Speech Database and Assessments, pp. 120–123.

[44] B. H. Ahmed, T.-P. Tan, Automatic Speech Recognition of Code Switching Speech Using 1-best Rescoring, in: 2012 International Conference on Asian Language Processing, pp. 137–140.

[45] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, H. Li, 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.

[46] D.-C. Lyu, R.-Y. Lyu, Language Identification on Code-switching Utterances Using Multiple Cues, in: Ninth Annual Conference of the International Speech Communication Association, 2008.

[47] D.-C. Lyu, E.-S. Chng, H. Li, Language Diarization for Code-switch Conversational Speech, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7314–7318.

[48] The Carnegie Mellon Pronouncing Dictionary, `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

[49] `https://github.com/kaldi-asr/kaldi/blob/master/egs/\hkust/s5/conf/pinyin2cmu`.

[50] W. Cai, J. Chen, J. Zhang, M. Li, On-the-Fly Data Loader and Utterance-Level Aggregation for Speaker and Language Recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 1038–1051.

[51] D. Griffin, J. Lim, Signal Estimation from Modified Short-time Fourier Transform, IEEE Transactions on acoustics, speech, and signal processing 32 (2) (1984) 236–243.

[52] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, WaveNet: A Generative Model for Raw Audio, in: 9th ISCA Speech Synthesis Workshop, 2016, pp. 125–125.

[53] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, K. Kavukcuoglu, Efficient Neural Audio Synthesis, in: Proceedings of the 35th International Conference on Machine Learning, 2018, pp. 2410–2419.

[54] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, A. C. Courville, MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis, in: Advances in Neural Information Processing Systems, Vol. 32, 2019.

[55] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., The Kaldi Speech Recognition Toolkit, in: 2011 workshop on automatic speech recognition and understanding, pp. 1–4.

[56] F. Grézl, M. Karafiát, S. Kontár, J. Cernocky, 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.

[57] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust DNN Embeddings for Speaker Recognition, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5329–5333.

[58] B. Desplanques, J. Thienpondt, K. Demuynck, ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, in: Proc. Interspeech 2020, pp. 3830–3834.

[59] W. Cai, J. Chen, M. Li, 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.

[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[61] S. Ding, G. Zhao, R. Gutierrez-Osuna, Improving the Speaker Identity of Non-Parallel Many-to-Many Voice Conversion with Adversarial Speaker Recognition, in: Proc. Interspeech 2020, pp. 776–780.

[62] K. Ito, The LJ Speech Dataset, `https://keithito.com/LJ-Speech-Dataset/` (2017).

[63] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: An ASR Corpus Based on Public Domain Audio Books, in: 2015 IEEE international conference on acoustics, speech and signal processing, pp. 5206–5210.

[64] J. Du, X. Na, X. Liu, H. Bu, Aishell-2: Transforming Mandarin ASR Research into Industrial Scale, arXiv preprint arXiv:1808.10583.

[65] J. S. Chung, A. Nagrani, A. Zisserman, VoxCeleb2: Deep Speaker Recognition, in: Proc. Interspeech 2018, pp. 1086–1090.

[66] C. Veaux, J. Yamagishi, K. MacDonald, et al., Superseded-CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit.

[67] Y. Shi, H. Bu, X. Xu, S. Zhang, M. Li, AISHELL-3: A Multi-Speaker Mandarin TTS Corpus, in: Proc. Interspeech 2021, pp. 2756–2760.

[68] J. Yang, L. He, Towards Universal Text-to-Speech, in: Proc. Interspeech 2020, pp. 3171–3175.

[69] W.-C. Huang, T. Hayashi, S. Watanabe, T. Toda, 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.

[70] S. Zhao, T. H. Nguyen, H. Wang, B. Ma, 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.