university of groningen

campus fryslân

Code-switching speech synthesis for Mandarin-English using FastSpeech2: A unified IPA-based approach

Wang Yinqiu

university of groningen

campus fryslân

University of Groningen - Campus Fryslân

Code-switching speech synthesis for Mandarin-English using FastSpeech2: A unified IPA-based approach

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Li Zhu (Voice Technology, University of Groningen)

Wang Yinqiu (S-5716675)

June 11, 2024

Acknowledgements

I would like to express my sincere gratitude to my advisor Li Zhu, and my external advisor Mei ZhengKun, for their invaluable advice and guidance throughout my research. Their patience and encouragement have been instrumental in my academic journey.

I also wish to extend my heartfelt thanks to my parents. Their unconditional support and understanding have been crucial to the completion of my thesis.

Lastly, I want to especially thank my friends, Zhang Shan and Jiang Weihao, for their constant support and encouragement.

Abstract

With the increasing prevalence of multilingual societies and cross-cultural interactions, the ability to synthesize natural-sounding code-switching speech has become crucial for enhancing communication and accessibility. However, the scarcity of appropriate code-switched datasets and the inherent complexity of handling multiple languages within a single utterance pose significant challenges for TTS systems. The purpose of this study is to explore Mandarin-English code-switching speech synthesis based on FastSpeech2, with the goal of synthesizing speech that is both intelligible and natural. This paper mainly explores two methods to achieve speech synthesis with code-switching between Mandarin and English: (1) directly modeling Mandarin and English phonemes; (2) unifying the input formats for both languages as phonological features(PF) based on the International Phonetic Alphabet (IPA). Additionally, considering that the current available open-source Mandarin and English code-switching datasets are designed for Automatic Speech Recognition(ASR) and have lower audio quality, this study recorded 500 high-quality Mandarin-English code-switching audio clips as a fine-tuning dataset to improve the quality of the speech synthesized by the model. The proposed method will be evaluated using subjective listening assessments. According to the MOS results, directly modeling phonemes can produce intelligible speech, while modeling PF can produce speech that is both intelligible and natural. Successful development of code-switching TTS systems as explored here can facilitate communication across languages, with applications in education, media, and assistive technologies. Here are some audio samples from the demo page: `https://wangyinqiu.github.io/Mandarin_and_English_CS_TTS/`

Keywords: Code-switching Speech Synthesis, Text-to-Speech Synthesis, Phonological Features, Multilingual Speech Technology

Contents

1 Introduction

With the advent of end-to-end Text-to-Speech (TTS) models such as Tacotron2(Shen et al., 2018) and FastSpeech2(Ren et al., 2020), the quality and naturalness of monolingual speech synthesis have been greatly improved, so that it is often difficult for ordinary listeners to distinguish whether the audio is synthesized or real. However, this is still a challenging area for code-switched speech synthesis.

Code-switching refers to the process during communication where people switch from one language to another as needed. This switching can occur between sentences or within a single sentence(Zhou et al., 2020). This phenomenon is very common in people's daily lives, especially among speakers who are fluent in multiple languages, they often switch between languages to express themselves more precisely. In addition, with the advancement of technology, many proprietary terms such as "AI", "Chatgpt", and "APP" are difficult to find appropriate equivalents for in non-English contexts, making code-switching an almost inevitable part of daily life.

In recent years, with the advancement in technology and the widespread popularity of smart home appliances and electronic products, the application of voice technology in people's daily lives has become increasingly extensive. People are using voice technology to interact with machines more frequently, such as Apple's Siri, Xiaomi's XiaoAi, and Amazon's virtual assistant Alexa, all of them provide voice interaction products, which greatly facilitate people's lives. Therefore, considering code-switching speech synthesis is crucial for providing a voice interaction product that satisfies users.

However, many current TTS systems assume that the input is a single language rather than a code-switched situation. Therefore, when synthesizing code-switched text, the TTS system usually cannot correctly process the input text. Additionally, from a linguistic resource perspective, code-switching is often considered as a low-resource language(Sitaram and Black, 2016), and the lack of code-switching datasets shows a significant challenge. To build such a code-switched dataset, the recorder needs to master at least two or more languages, making the collection of code-switching datasets costly.

Mandarin and English are both resource-rich languages, with a large number of users, open-source data, and technical support. In Chinese society, code-switching between Chinese and English is very common, as people frequently switch between the two languages to express themselves more accurately. For example:

- 我今天得在家完成assignment, 明天就是deadline了. (I have to complete the assignment at home today, and the deadline is tomorrow.)
- Happy birthday! 祝你生日快乐! (Happy birthday! Happy birthday to you!)

Although the common occurrence of code-switching between Mandarin and English, there are relatively few studies and datasets on Mandarin-English code-switching. The currently available open-source code-switched datasets for Mandarin and English are primarily designed for Automatic Speech Recognition (ASR) tasks, and these datasets often have poor audio quality. This increases the difficulty of training a TTS model that supports both Mandarin and English.

Recently, (Zhang and Lin, 2021) proposed that pre-training models using low-quality ASR speech data followed by fine-tuning with higher-quality datasets can effectively improve the ”low-quality” attributes of synthesized audio and enhance the performance of code-switching speech synthesis. At the same time, given that there are a large number of open-source datasets for both Mandarin and English, many researchers have tried to use these monolingual datasets to achieve code-switched Mandarin-English speech synthesis. (Zhao et al., 2020) proposed a cross-language voice conversion method using Tacotron2 to process monolingual corpora of Mandarin and English to build a bilingual corpus with a single speaker, enabling code-switching. Similarly, (Cao et al., 2019) explored Mandarin-English code-switching speech synthesis based on the Tacotron2 end-to-end framework. To handle the input texts in Chinese and English, the authors experimented with two different encoders: a shared multilingual encoder with explicit Language Dependent Embeddings (LDE) and separate monolingual encoders (SEP).

For code-switched speech synthesis, handling input representations in different languages is another challenge. (Li et al., 2019) proposed using Unicode bytes as the input for the model. (Zhang et al., 2019) evaluated the effect of using graphemes, phonemes, and Unicode bytes as input for a multilingual synthesis model and found that phonemes performed the best. Subsequently, (Staib et al., 2020;Wells and Richmond, 2021) proposed that phonological features are a more suitable input representation than phonemes in multilingual synthesis models.

Considering the significant differences in the phonetic structures of Mandarin and English, with almost no shared phonemes, this paper explores code-switching speech synthesis between Mandarin and English based on FastSpeech2 architecture in two ways: (1) using phonemes as input representations, directly modeling the phonemes of both Mandarin and English; (2) using phonological features (PF) as the input representation, mapping Mandarin and English to phonological features based on International Phonetic Alphabet (IPA), and modeling these PF features. This study aims to synthesize intelligible and natural code-switched speech through the above training. The final experimental results show that although directly modeling Mandarin and English phonemes can synthesize intelligible code-switching speech, the synthesized speech has heavy artificial traces, and the tone and rhythm are not natural enough. However, by modeling the phonological features of Mandarin and English, it is possible to synthesize easily understandable and natural code-switching speech.

This study is mainly divided into six parts, which are arranged as follows: The first part is the introduction, which mainly explains the prevalence of Mandarin-English code-switching scenarios and the importance of studying them. The second part is the literature review, which explains the development and advantages of neural network speech synthesis, the current state of research on code-switching speech synthesis, and research related to phonological features. The third part is the methodology, including six aspects: dataset, model used, MFA (forced alignment tool), phoneme modeling, phonological feature (PF) modeling, and ethical considerations. The fourth part is the experimental setup, detailing the experimental settings and evaluation methods. The fifth part shows the experimental results. The sixth part discusses the experimental results, draws conclusions, and looks forward to future research directions.

2 Literature Review

This chapter will conduct a comprehensive literature review on speech synthesis, code-switching speech synthesis, and phonological features (PF). The structure of the chapter is as follows: In Section 2.1, I will mainly discuss the development process of neural speech synthesis. In Section 2.2, I will primarily discuss research on code-switched speech synthesis. In Section 2.3, I will mainly discuss research related to modeling PF. In Section 2.4, I will discuss the research questions and hypothesis.

2.1 Neural Speech Synthesis

Text-to-speech (TTS) is the process of converting text into sound. With the continuous development and progress of the times and technology, TTS methods have continued to evolve, from articulatory synthesis, formant synthesis, and concatenative synthesis to statistical parameter speech synthesis(SPSS), and later to neural network-based speech synthesis, with increasingly better audio quality. (Coker, 1976) proposed a human articulator that can convert English text into understandable synthesized speech. (Seeviour et al., 1976) tried to generate speech by simulating the resonant frequency (i.e., formant) of the human vocal tract. Subsequently, (Yoshimura et al., 1999) described how to simultaneously model spectrum, pitch, and duration in the HMM speech synthesis system and synthesize speech based on these parameters.

With the development of deep learning, Text-to-Speech (TTS) technology based on neural networks was proposed, which mainly uses deep learning networks as the core architecture of speech synthesis(Tan et al., 2021). (Zen et al., 2013) introduced the method of using deep neural network (DNN) instead of the Hidden Markov Model (HMM) for acoustic modeling in statistical parametric speech synthesis (SPSS). (Wu et al., 2016) introduced Merlin, a model that uses various neural network architectures such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) for speech synthesis. Later, (Van Den Oord et al., 2016) introduced WaveNet, a model that can generate waveforms directly from linguistic features and is considered the first modern neural TTS model(Tan et al., 2021). Subsequently, the end-to-end model Tacotron1/2(Wang et al., 2017;Shen et al., 2018) were proposed. Tacotron2 uses characters as input to simplify the front-end processing process. At the same time, it uses Mel-Spectrogram to simplify the acoustic features, significantly improving the quality of synthesized speech. However, because Tacotron2 is an autoregressive model based on RNN structure, its training and inference speed are very slow, and it is not stable enough when dealing with long-term dependencies, often resulting in skipped or repeated words.

To address these issues, FastSpeech(Ren et al., 2019) uses a feed-forward Transformer network to generate Mel-Spectrograms in parallel, which greatly improves the inference speed. Additionally, FastSpeech introduces a Length Regulator to predict the duration of phonemes, effectively solving the problems of word skipping and repetition. One year later, (Ren et al., 2020) proposed the FastSpeech2 model, which further enhanced FastSpeeh by using ground-truth Mel-Spectrograms as training data, thereby simplifying the training process. Moreover, FastSpeech2 introduces the Variance Adapter, which not only considers the duration but also pitch and energy, to better solve the one-to-many mapping problem in TTS. The prediction of duration is achieved by using the phoneme duration obtained through the forced alignment tool Montreal Forced Aligner (MFA) (McAuliffe et al., 2017).

Based on the above discussion, this study chose to use FastSpeech2 as the experimental model, hoping to improve the inference speed while generating stable and natural speech.

2.2 Code-Switched Speech Synthesis

Code-switching refers to the conversion from one language to another within a single utterance. It is commonly seen in social media platforms, news broadcasts, and educational contexts, making it an important and under-explored mode of communication(Sitaram et al.). In daily life, bilingual individuals proficient in Mandarin and English might frequently switch between these two languages to more accurately express their thoughts. Additionally, with the development of technology, it is difficult to find a suitable Chinese alternative for proper nouns such as "AI" and "Chatgpt". Therefore, in people's daily lives, the phenomenon of code-switching is almost inevitable. However, although the current TTS models can synthesize high-quality synthetic audio, most TTS models assume the input to be a single language. If text mixed with multiple languages is input directly, the model may synthesize incorrectly pronounced audio or might even skip some words altogether(Cao et al., 2019).

Early explorations of code-switched TTS were primarily based on HMM architectures. (Latorre et al., 2005) proposed using the HMM framework combined with data from monolingual speakers to create a multilingual averaged voice. (Liang et al., 2007) introduced a bilingual Mandarin-English TTS system based on HMMs. (Zen et al., 2012) provides an HMM-based method to separate and independently process speaker features and language features in multi-language environments, which is very useful for speech synthesis that handles multiple language switching. Subsequently, (Li and Zen, 2016) introduced a multilingual SPSS speech synthesis system that uses Cluster Adaptive Training (CAT) to simulate language variations and simulate speaker changes through a speaker-dependent output layer to achieve natural Multilingual speech synthesis. (Ming et al., 2017) proposed using data from a single speaker to build a bilingual TTS system capable of handling both Mandarin and English.

Recently, (Cao et al., 2019) explored Mandarin-English code-switching speech synthesis using the Tacotron2 end-to-end framework. To process input texts in both Chinese and English, they experimented with two different encoder types: a shared multilingual encoder incorporating explicit Language Dependent Embeddings (LDE), and separate monolingual encoders (SEP). Later,(Zhao et al., 2020) proposed a cross-language voice conversion method based on Tacotron2 to process monolingual corpora of Mandarin and English to build a bilingual corpus with a single speaker, enabling code-switching. (Zhang and Lin, 2021) proposed that pre-training models with low-quality ASR speech data followed by fine-tuning with higher-quality datasets can effectively improve the "low-quality" attributes of synthesized audio and improve the performance of code-switched speech synthesis.

Some studies have explored the input representations for code-switching. (Li et al., 2019) compared the effects of modeling characters and Unicode byte sequences, proposing that using Unicode byte sequences to model text can improve the performance of multilingual TTS systems. (Zhang et al., 2019) introduced a multilingual TTS system based on Tacotron2 and evaluated the effects of modeling characters, UTF-8 encoded bytes, and phonemes, pointing out that using phoneme input representation can achieve better performance. (Cai et al., 2020) introduced a bilingual multi-speaker TTS method based on shared phoneme representation, and highlighted that training with code-switched corpora results in better code-switching effects. Additionally, (Staib et al., 2020;Wells and Richmond, 2021) suggested using phonological features as a unified representation for inputs in different languages, facilitating code-switching and cross-language speech synthesis.

Based on this, it can be seen that when dealing with the problem of code-switching speech synthesis, quite a few researchers focus on solving the input format problems between different languages. For example, they explore how to unify multiple languages into a standard format for input, or whether it is necessary to unify the format. This study also focuses on solving the problem of input formats for different languages. However, unlike previous studies, this paper mainly compares speech synthesized by phoneme modeling and phonological feature (PF) modeling. On the basis of ensuring that the model can generate code-switching speech, this study will also consider the naturalness and comprehensibility of the synthesized speech.

2.3 Phonological Features

Phonemes are a common input representation for many TTS models and have a stronger correlation with acoustic features compared to graphemes. Although phonemes are defined as the smallest contrastive sound units of language, they are not indivisible. In fact, phonemes can be considered as combinations of phonological features (PF), where the sum of multiple individual attributes forms a phoneme. These features represent the smallest units in phonetic analysis(Giegerich, 1992).

PF can differentiate between various sounds in a language. For example, the Voiced/Unvoiced (VUV) feature can distinguish between two dental sounds: [t] and [d]. Features such as Vowel frontness, Vowel openness, and Vowel roundedness determine the articulation position of vowels. (Lux and Vu, 2022) proposed that using this kind of phonetic unit can alleviate the problem of low-resource speech synthesis because most of these phonetic units are independent of language.

(Gutkin et al., 2018;Demirsahin et al., 2018)explored using PF or phoneme combinations as input for multilingual models, and pointed out that various automatically derived sets of speech features can be used to replace or supplement input features, thereby improving the intelligibility of synthesized speech. (McAuliffe et al., 2017) converts PF into a 62-dimensional one-hot encoding to handle unseen phonemes, thereby achieving high-quality speech synthesis. In some ways, PF is similar to the bytes discussed in (Li et al., 2019), and can be considered as a ”byte version” of phonemes, which provides a unified format input for multilingual models(Staib et al., 2020). In addition, (Staib et al., 2020) also pointed out that PF can remain valid across different languages through a shared model structure. At the same time, it is able to link to basic phonetic units and provide a clear explanation of the actual pronunciation manner. (Wells and Richmond, 2021) also emphasizes that in multilingual speech synthesis or code-switching speech synthesis, PF is a more appropriate input representation than phonemes.

Based on this, this study believes that for Mandarin and English code-switching speech synthesis, Mandarin and English can be treated as low-resource languages, and PF can be used as the model input for modeling. It is expected that this approach will generate speech that is both easily understandable and highly natural.

2.4 Research Question and Hypothesis

Based on previous research (Zhang et al., 2019; Staib et al., 2020) and interest in code-switching speech synthesis, this study raises the following question:

> **How can code-switching speech synthesis between Mandarin and English be effectively implemented based on FastSpeech2, to synthesize intelligible and natural-sounding Mandarin-English code-switched speech?**

Based on the research question, the hypotheses of the study are as follows:

- By directly modeling the phonemes of both Mandarin and English, it is possible to effectively synthesize intelligible code-switched Mandarin-English speech;
- By mapping the phonemes of Mandarin and English to phonological features (PF) based on the International Phonetic Alphabet (IPA), it can not only achieve intelligible speech output but also significantly improve the naturalness of synthesized speech.

3 Methodology

In this chapter, I will elaborate on the two methodologies proposed to address and verify the research questions and hypotheses: (1) directly modeling Mandarin and English phonemes; (2) modeling the phonological features (PF) of Mandarin and English. The organization of this chapter is as follows: In Section 3.1, I will mainly discuss the dataset used in this study. In Section 3.2, I will primarily discuss the model architecture used in this study. In Section 3.3, I will mainly address the processing of MFA phoneme alignment, involving the creation of a mixed Mandarin-English dictionary and the training of a mixed Mandarin-English acoustic model. In Section 3.4, I will discuss the methodology for modeling the phonemes of Mandarin and English directly. In Section 3.5, I will explain how to model the phonological features of Mandarin and English. In Section 3.6, I will discuss the ethical considerations of this study.

3.1 Datasets

This study utilized three open-source datasets: the Chinese Standard Mandarin Speech Corpus from Baker data[1], LJSpeech(Ito and Johnson, 2017), and TAL_CSASR[2]. The Chinese Standard Mandarin Speech Copus(12 hours) is a standard single-speaker Mandarin corpus suitable for TTS model training. LJSpeech(24 hours) is also a very classic single-speaker English corpus suitable for TTS model training. TAL_CSASR(587 hours), on the other hand, is a multi-speaker Mandarin-English code-switching dataset designed for ASR tasks. These recordings are primarily in Mandarin, with some English words occasionally mixed in the sentences. Since this dataset comes from recordings of English classes, it has relatively poor audio quality and fast speech rates, making it theoretically unsuitable for TTS model training. In addition, considering the huge amount of data in the TAL_CSASR dataset and the limitations of time and resources, this study only extracted about 22 hours of speech, a total of 12,000 speech samples, for training.

Considering the overall poor quality of the TAL_CSASR dataset, in order to improve the quality of the final synthesized audio, this study specially recorded a high-quality single-speaker Mandarin-English code-switching dataset of 500 audio clips. For convenience, this dataset will be referred as CS_500 in this study. This dataset was recorded by bilingual speakers proficient in both Mandarin and English in a quiet room using professional microphones. The recording environment and equipment remained unchanged from beginning to end. The recordings were made in mono at a 16 kHz sampling rate, and the audio format is WAV. The recording corpus is mainly designed in Chinese, but each sentence includes at least one English word, for example:

- 今天我们班有一个Quiz，我得复习一下（Today, our class has a Quiz, and I need to review it.）
- 我今天要参加一个生日party（I have to attend a birthday party today.）

[1] https://www.data-baker.com/data/index/TNtts/

[2] https://ai.100tal.com/openData/voice

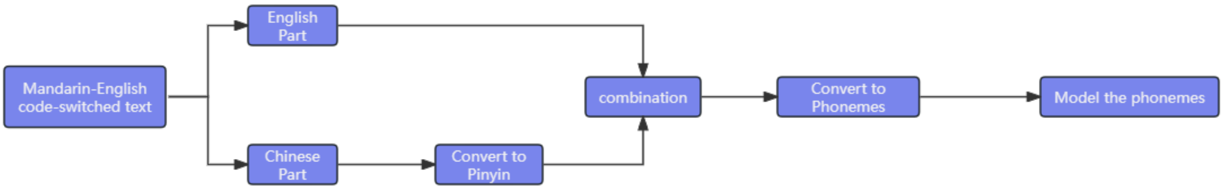


Figure 1: The process of modeling the phonemes directly.

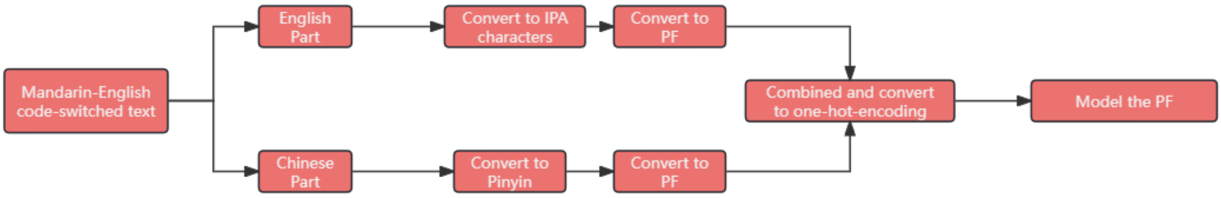

English Part
Convert to IPA characters
Convert to PF
Mandarin-English code-switched text
Combined and convert to one-hot-encoding
Model the PF
Chinese Part
Convert to Pinyin
Convert to PF

Figure 2: The process of modeling the phonological features.

3.2 FastSpeech2

This research is mainly based on the FastSpeech2 model (Ren et al., 2020), which is a fast and high-quality end-to-end non-autoregressive model. The model takes a sequence of phonemes as input and converts them into a sequence of phoneme embeddings through a phoneme embedding layer. These embedding sequences are converted into phoneme hidden sequences through the encoder. Additionally, the model has a Variance Adapter, which can add pitch, energy, duration and other information to the hidden sequence, effectively addressing the one-to-many problem in TTS. In order to achieve precise duration predictions, FastSpeech2 utilizes Montreal Forced Alignment (MFA) (McAuliffe et al., 2017) to obtain phoneme duration. Then, the Mel-Spectrogram decoder converts the hidden sequence into a Mel-Spectrogram sequence in parallel to generate high-quality speech.

This study conducted two experiments based on FastSpeech2: (1) modeling phonemes directly, as shown in Figure 1; and (2) modeling phonological features, as shown in Figure 2.

In the experiment of directly modeling Mandarin and English phonemes, considering the huge number of Chinese characters, it is first necessary to segment each corpus. The Chinese parts are converted into Pinyin while the English parts remain unchanged. Then a Chinese-English mixed dictionary is used to map the corpus, uniformly converting it into phonemes, which are subsequently modeled. The Chinese-English mixed dictionary here is a combination of Pinyin-to-phoneme and English word-to-phoneme mappings. The pinyin part takes into account four tones, which are represented by the numbers 1, 2, 3, and 4, and the light tone is represented by the number 5.

In the experiment of modeling phonological features, it is also necessary to segment the corpus into Chinese and English parts. The Chinese parts are converted into Pinyin, and the English parts are converted into IPA characters. Then the Chinese and English are converted to the pre-defined PF dictionary respectively. The PF features are then converted into one-hot vectors, which are subsequently modeled. The PF also includes the four tones in Pinyin, represented by the numbers 1, 2, 3, and 4, with the light tone represented by the number 5.

3.3 Montreal Forced Aligner Models

The Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) is a widely used phoneme alignment tool that employs Kaldi to perform forced alignment on speech datasets. In the process of MFA alignment, a pronunciation dictionary and acoustic model for the specific language are required. Although MFA officially provides pronunciation dictionaries and pre-trained acoustic models for most languages, it does not offer specific resources for the special case of Mandarin-English code-switching. Fortunately, MFA supports the training of pre-trained acoustic models for various languages by users themselves, which provides the possibility to handle Mandarin-English code-switching. Therefore, for this particular study, a bilingual pronunciation dictionary and acoustic model capable of handling both Mandarin and English were independently created and trained. To ensure consistency in phoneme encoding, this bilingual pronunciation dictionary and acoustic model were used when processing the TAL_CSASR, Chinese Standard Mandarin Speech Corpus, LJSpeech datasets, and CS_500.

Considering the huge number of Chinese characters, this study used the Pinyin library to convert Chinese into Pinyin when preprocessing the dataset corpus, while English remains unchanged. Subsequently, the Chinese Pinyin dictionary officially provided by MFA was mixed with the English word dictionary to create a new dictionary, which contains the mapping of Mandarin and English phonemes. After preparing the speech corpus and pronunciation dictionary, the next challenge was addressing the acoustic model for Mandarin-English code-switching. Once the corpus and dictionary were validated and out-of-vocabulary (OOV) words were addressed, this study trained an acoustic model capable of handling Mandarin and English code-switched speech using the speech corpus and pronunciation dictionary. The corpus was then aligned, ultimately generating TextGrid files.

3.4 Modelling Phonemes Separately

Phoneme sequence is the standard input representation of the FastSpeech2 model. This model converts the phoneme embedding sequence into a phoneme hidden sequence through the encoder(Ren et al., 2020). (Zhang et al., 2019) proposed that for handling multilingual synthesis in TTS systems, phonemes from different languages can be concatenated for training, allowing for the sharing of equivalent phonemes across languages. This approach is noted to promote the ability to share models across languages. This study also uses a similar method to that described in (Zhang et al., 2019), directly training on the phoneme symbols of both Mandarin and English and sharing equivalent phonemes between the two languages.

To further verify the method and effect of directly modeling Mandarin and English phonemes, this study designed three sets of experiments:

- Group A: The model is trained entirely using code-switched Mandarin and English data. Specifically, the model is first pre-trained using the TAL_CSASR dataset and then fine-tuned using the higher-quality CS_500 dataset.

- Group B: The model's training data is primarily in Mandarin, with only a small amount of Mandarin-English code-switching data used for fine-tuning. In this group, the model is initially pre-trained using the Chinese Standard Mandarin Speech Corpus and then fine-tuned with the CS_500 dataset.

- Group C: The model does not use any code-switched Mandarin and English datasets for training. In this group, the model is also first pre-trained using the Chinese Standard Mandarin Speech Corpus, followed by fine-tuning using the LJSpeech dataset.

Since the primary goal of this study is to synthesize Mandarin-dominant Mandarin-English code-switching speech, the model was not pre-trained using a pure English dataset.

3.5 Modelling Phonological Features

The work of (Staib et al., 2020;Wells and Richmond, 2021) explored the possibility of applying phonological features (PF) in TTS systems and proposed that PF might be a more suitable input representation than phonemes for multilingual speech synthesis. Based on this, this study attempts to use PF features to train a Mandarin-English code-switching model, expecting to synthesize intelligible and natural Mandarin-English code-switched speech.

This study mainly used the IMS-Toucan toolkit (Lux et al., 2021) when modeling and training PF features. This toolkit provides monolingual speech synthesis in English, German, Spanish, Chinese, and other languages. However, the IMS-Toucan toolkit does not natively support code-switching speech synthesis. Therefore, this study made necessary modifications to the toolkit to adapt to the needs of this study.

Considering that the IMS-Toucan toolkit already supports Mandarin and English speech synthesis, the most straightforward approach was to modify the toolkit's front end to handle text with Mandarin-English code-switching and to convert it into corresponding PF. Based on this, this study treated Mandarin-English code-switched corpus as a new language. In processing each piece of text, it was divided into Mandarin and English parts, which were then processed separately according to the existing Mandarin and English logic within IMS-Toucan. Finally, the PF features are converted into 62-dimensional one-hot encoding and input into the model for modeling.

3.6 Ethical considerations

The primary objective of this study is to develop a Mandarin-English code-switching TTS system. Since the experimental process involves data collection, if the collected data is not properly processed, it may lead to serious consequences such as the leakage of personal information. Therefore, in this section, I will discuss ethical considerations, mainly focusing on two aspects: (1) audio data; (2) questionnaire data.

For audio data, this study used three open-source datasets: TAL_CSASR, LJSpeech, and Chinese Standard Mandarin Speech Copus. All three datasets are open-source, and this study strictly abides by the dataset usage terms, so they will not be discussed in detail here. Readers can download and use them from the relevant web pages if needed. In addition, this study also uses a self-made audio dataset CS_500, which consists of 500 audio recordings that I personally recorded. Due to privacy concerns, this dataset will not be made publicly available.

For the questionnaire data, all data collection has obtained the informed consent of the participants and will be properly stored after the experiment concludes. This study primarily collected two pieces of personal data: name and email address, with no requirement to provide a real name. The main purpose of collecting names and email addresses in this study is to facilitate experimental records and to be able to contact participants in case of experimental errors. After the experiment ends, all participant data will be anonymized and securely stored to prevent any leakage of personal information.

4 Experimental Setup

In this chapter, I will provide a detailed description of the experimental setup to facilitate the replication of the experiment. The organization of this chapter is as follows: In Section 4.1, I will discuss the model used and the parameters for training. In Section 4.2, I will elaborate on the evaluation tests for this study. This study encourages readers to reproduce the experiment and the following is the link to the demonstrator: `https://github.com/WANGYINQIU/Thesis_Project.git`

4.1 Training Setup

This study is mainly based on the FastSpeech2 model and employs two methods for training it, relying on the FastSpeech2[3] and IMS-Toucan[4] GitHub repositories for implementation. The experimental settings of these two methods will be described below.

This study conducted the following three sets of experiments on Mandarin and English phoneme modeling:

- Group A: The model was pre-trained on the TAL_CSASR, followed by fine-tuning with CS_500.
- Group B: The model was pre-trained on the Chinese Standard Mandarin Speech Corpus, followed by fine-tuning using the CS_500.
- Group C: The model was pre-trained on the Chinese Standard Mandarin Speech Corpus, followed by fine-tuning with LJSpeech.

In the three sets of experiments involving modeling Mandarin and English phonemes, the initial step involves using the MFA tool to process the TAL_CSASR, the Chinese Standard Mandarin Speech Corpus, LJSpeech, and CS_500 datasets to obtain TextGrid files for alignment information. Subsequently, it is necessary to preprocess the corpora and TextGrid files and then divide them into training and validation sets. Before training, it is necessary to define some phonemes in the symbols.py file that were not mentioned in the original repository. In these three sets of experiments, the pre-trained models were trained for 640,000 steps and then fine-tuned for 100,000 steps using the corresponding datasets, bringing the total number of training steps to 740,000. In the speech synthesis stage, the text to be synthesized needs to be divided into Chinese and English parts. Each part is then converted into phonemes using the Pinyin library for Mandarin and the G2p library for English respectively. For the vocoder, the HiFi-GAN universal vocoder provided in the FastSpeech library is used for speech synthesis.

Regarding the method(Group D) of modeling Phonological Features(PF), this study primarily relies on the IMS-Toucan toolkit. The following set of experiments was conducted:

- Group D: Fine-tuning the pre-trained model provided by IMS-Toucan using the CS_500 dataset for 30,000 steps.

[3] https://github.com/ming024/FastSpeech2

[4] https://github.com/DigitalPhonetics/IMS-Toucan

Considering that the PF-based model requires a large number of training languages to effectively generate speech in new languages (Staib et al., 2020), this study chose to fine-tune the pre-trained model provided by the IMS-Toucan library using the CS_500 dataset, with fine-tuning involving 30,000 steps. Since IMS-Toucan itself does not support Mandarin-English code-switching, this study treats Mandarin-English code-switching as a new language processing scenario and has developed a new training pipeline specifically for it.

The training pipeline is mainly modified for the front end, dividing each record of CS_500 into Chinese and English parts. These parts are then converted into PF according to the logic for handling Mandarin and English within the IMS-Toucan toolkit. Next, these PF are converted into a 62-dimensional one-hot encoding, which is used as input for training the model. In the speech synthesized step, the same logic is applied, converting the text to be synthesized into one-hot encoding based on PF, to facilitate speech synthesis. This process ensures that the model can effectively handle code-switching between Mandarin and English, thus producing natural and fluent bilingual speech output.

4.2 Evaluation

This study uses Mean Opinion Score (MOS) evaluations to test the effect of synthetic speech in each group of experiments. Participants are asked to rate the audio based on two criteria: intelligibility and naturalness. The evaluation includes four sets of tests, each with ten audio samples synthesized by Groups A, B, C, and D respectively. To fairly compare the quality of the four groups of samples, the content of the synthesized text is kept consistent across all groups. A total of 30 bilingual speakers, fluent in English and native in Chinese, are invited to participate in the evaluation. The texts synthesized for testing are shown in the following table 1:

Participants are asked to rate the synthesized sentences on a scale from 1 to 5, where 1 means ”bad”, 2 means ”poor”, 3 means ”fair”, 4 means ”good,” and 5 means ”excellent.” The evaluation criteria included: (1) the fluency and naturalness of the speech, which checked the effectiveness of the synthesized speech; and (2) the intelligibility of the speech, which checked whether the pronunciation of the words was accurate.

Table 1: Textual content of synthesized audio.

Numbers	Original Text	Translation
1	周末你有什么plans吗?	Do you have any plans for the weekend?
2	她对这部电影的评论是amazing。	She thinks that the movie is amazing.
3	我对这个topic很感兴趣。	I’m very interested in this topic.
4	我的朋友推荐我尝试一下yoga，说它对mental health有很好的帮助。	My friend recommended that I try yoga, saying it’s very beneficial for mental health.
5	最近工作真的很忙，但我还是找时间去gym锻炼，保持身体的fitness。	I’ve been really busy with work recently, but I still find time to go to the gym to maintain my fitness.
6	如今AI的发展太快了,我们也要学习新的knowledge,才能有更好的future。	The development of AI is progressing rapidly, we also need to learn new knowledge to have a better future.
7	我觉得他的经历很丰富，你应该多寻求他的suggestions。	I think his experience is extensive, you should seek his suggestions more.
8	Tiktok是最近非常热门的一款APP。	TikTok is a very popular APP recently.
9	我在大学学习的专业是marketing。	I studied marketing in college.
10	没关系，也许之后能有更好的chance。	It’s okay, maybe there will be a better chance later.

5 Results

5.1 MOS Results

To evaluate the effectiveness of the synthesized Mandarin-English code-switching speech, this study conducted MOS evaluations assessing the naturalness and intelligibility of the synthesized speech. The results are displayed in the following table 2:

Table 2: Naturalness and intelligibility MOS of Group A, Group B, Group C, Group D.

Experiments	Input	Intelligibility	Naturalness
Group A	phonemes	4.13	2.44
Group B	phonemes	4.05	2.73
Group C	phonemes	2.42	1.57
Group D	phonological features	4.68	3.81

From the table, it can be seen that in terms of naturalness, the performance of the four groups is: Group D > Group B > Group A > Group C. In terms of intelligibility, the performance of the four groups is: Group D > Group A > Group B > Group C. Group D (phonological features) is significantly better than the other three groups in both naturalness and intelligibility. Since Group C had never been trained using a Mandarin-English code-switching dataset, it exhibited the poorest performance in both naturalness and intelligibility among all groups. Group A and Group B performed similarly in terms of intelligibility, but in terms of naturalness, Group B performed better than Group A. This may be because the training datasets used by Group B (Chinese Standard Mandarin Speech Corpus and CS_500) are designed for TTS tasks and have higher quality. However, the dataset TAL_CSASR used by Group A in the pre-training stage is designed for ASR tasks, with poor overall quality and more noise. Although Group A used the higher-quality CS_500 in the fine-tuning stage, since CS_500 is mainly based on Chinese, the improvement effect on the English part is limited.

Specifically, the MOS results for intelligibility are as follows: Group A scored 4.13, and Group B scored 4.05, showing that the participants were relatively satisfied with the intelligibility of these two groups of experiments and were able to understand most of the content of the synthesized sentences. Group C scored significantly lower at 2.42 in intelligibility, indicating that participants had difficulty understanding most of the synthesized speech. Group D had the highest intelligibility among the four groups, with a MOS score of 4.68, indicating that participants were able to understand the synthesized sentences quite well. It can be seen that both phoneme and PF modeling can generate intelligible speech to a certain extent, but the speech generated by PF modeling is more comprehensible. In addition, the comprehensibility of Group C is the lowest among the four groups, which may be because Group C only used Mandarin and English corpora to train the model separately, making it difficult to handle code-switching between Mandarin and English.

In terms of naturalness MOS, the scores of Group A and Group B were 2.44 and 2.73 respectively. These scores indicate that participants found the audio quality of these two groups to be poor, with noticeable artificial traces, making it relatively easy to identify the speech as synthesized. Group C's score was only 1.57, indicating that participants found the audio quality of this group to be very poor, with highly unnatural intonation. In contrast, Group D had the highest naturalness MOS score of 3.81, which indicates that participants considered the audio synthesized by this group of experiments to be relatively natural. It can be seen that modeling PF can generate more natural speech, while speech generated by modeling phonemes is less natural and has obvious artificial traces. At the same time, Group B's naturalness score is slightly higher than Group A, which is mainly due to the different quality of the datasets used by the two groups in the pre-training stage. The poor quality of the dataset used by Group A may lead to its low naturalness. Group C has the lowest naturalness score, and most participants think that the audio synthesized by this group is very unnatural. Overall, directly modeling Mandarin and English phonemes without handling the model architecture makes it difficult to generate natural speech.

In summary, Group D performed the best among all groups in terms of both intelligibility and naturalness, indicating that modeling PF can generate understandable and natural speech. Group A and Group B also performed well in terms of intelligibility, indicating that directly modeling Mandarin and English phonemes can also generate understandable speech when trained with Mandarin-English code-switched corpus.

5.2 Statistical Analysis

5.2.1 Overall Comparison

This section will use ANOVA and Tukey's HSD to analyze and compare the differences in intelligibility and naturalness among Group A, Group B, Group C, and Group D. Below are the Figure 3 and Figure 4 for the MOS scores for naturalness and intelligibility of the ten synthesized sentences:

In terms of naturalness, it can be seen that the naturalness of Group D was significantly better than that of the other three groups, and the MOS naturalness score of each sentence was above 3 points. It is evident that in terms of intelligibility, Group A, Group B, and Group D all performed well overall, especially Group D, with the MOS scores for intelligibility of each sentence being above 4.

5.2.2 ANOVA and Tukey's HSD

To evaluate the significant differences in MOS scores for naturalness and intelligibility among the four groups of synthesized speech with Mandarin-English code-switching, this study conducted an ANOVA analysis. The ANOVA results for intelligibility showed an F-value of 37.32 and a p-value of 3.83e-11. The ANOVA results for naturalness showed an F-value of 50.56 and a p-value of 5.47e-13. These results show that the differences between different groups are very significant in both naturalness and intelligibility. The p-value is much less than 0.05, suggesting statistically significant differences in mean scores between groups. Therefore, this study further conducted Tukey's HSD test. Here are the Tukey's HSD results for naturalness (Table 3) and intelligibility (Table 4) MOS scores:

According to the results of Tukey's HSD test, it was found that there is no statistically significant

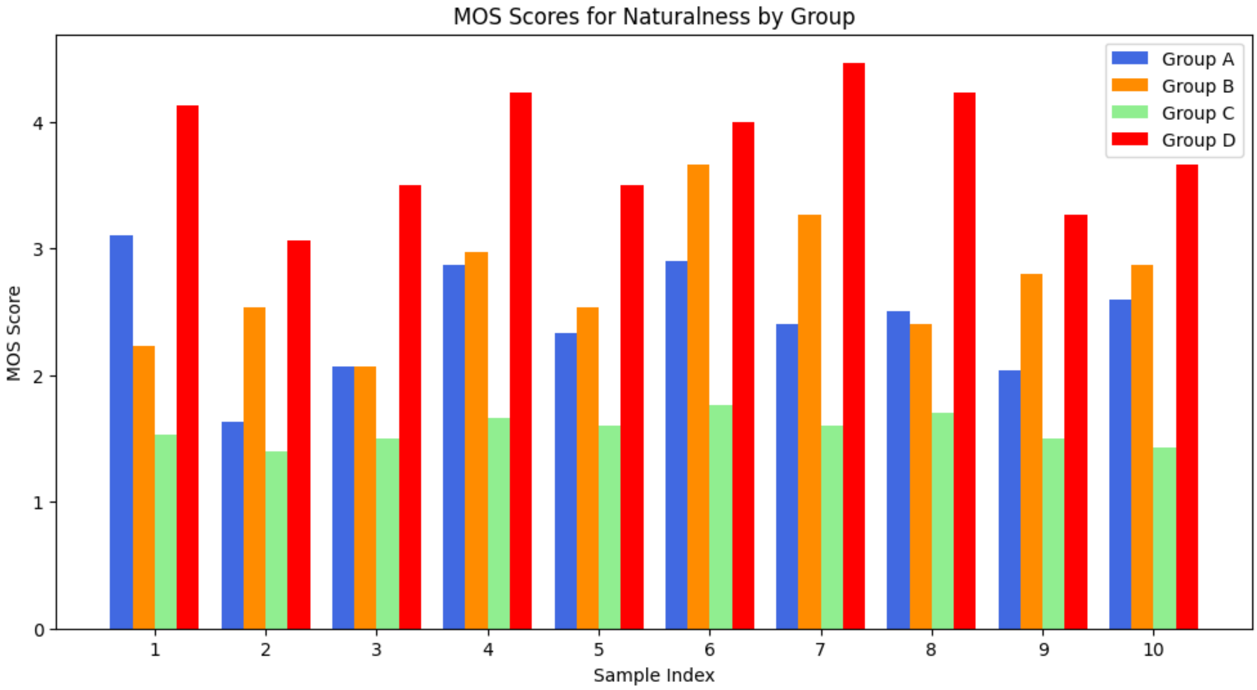


Figure 3: The MOS scores for naturalness by 10 different synthesized speech

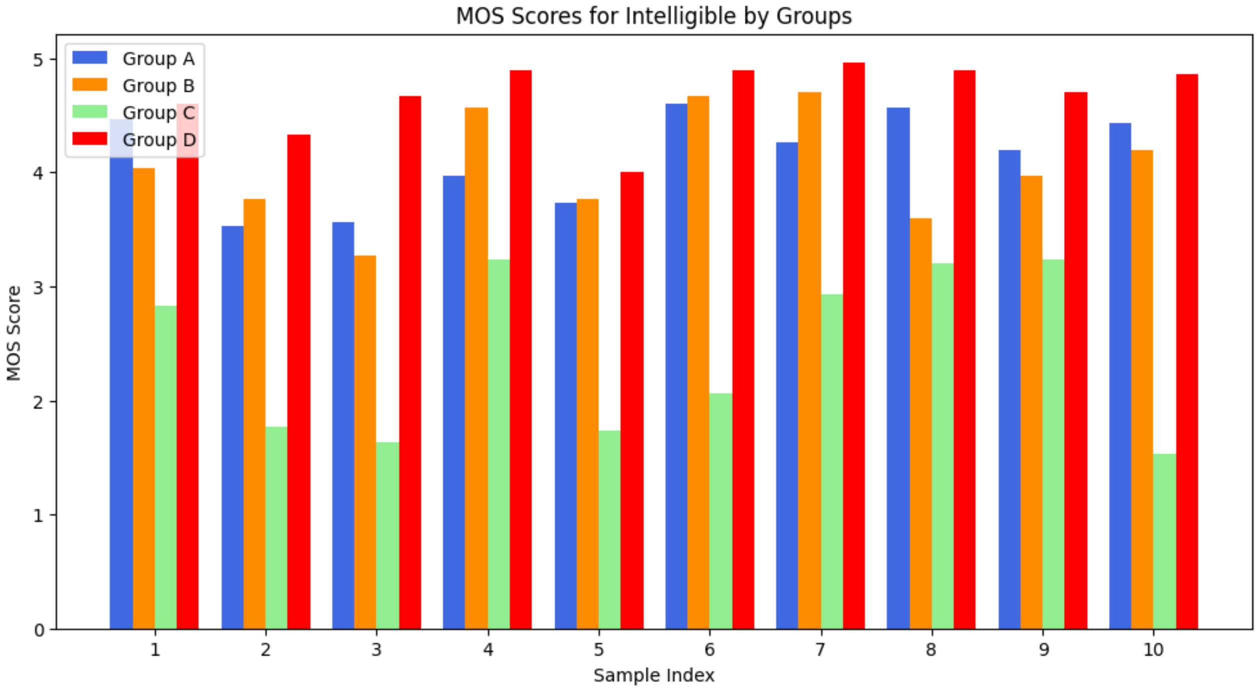


Figure 4: The MOS scores for intelligible by 10 different synthesized speech

Table 3: Tukey HSD result for Naturalness

group1	group2	meandiff	p-adj	lower	upper	reject
Group A	Group B	0.29	0.4022	-0.2042	0.7842	False
Group A	Group C	-0.8733	0.0002	-1.3675	-0.3791	True
Group A	Group D	1.3633	0.0	0.8691	1.8575	True
Group B	Group C	-1.1633	0.0	-1.6575	-0.6691	True
Group B	Group D	1.0733	0.0	0.5791	1.5675	True
Group C	Group D	2.2367	0.0	1.7425	2.7309	True

Table 4: Tukey HSD for Intelligible

group1	group2	meandiff	p-adj	lower	upper	reject
Group A	Group B	-0.08	0.9846	-0.6895	0.5295	False
Group A	Group C	-1.7167	0.0	-2.3262	-1.1071	True
Group A	Group D	0.55	0.0893	-0.0595	1.1595	False
Group B	Group C	-1.6367	0.0	-2.2462	-1.0271	True
Group B	Group D	0.63	0.0405	0.0205	1.2395	True
Group C	Group D	2.2667	0.0	1.6571	2.8762	True

difference between Group A and Group B in terms of naturalness. However, comparisons between other groups showed significant differences. Specifically, the mean differences in naturalness between Group D and the other three groups are statistically significant, indicating that modeling PF provides a significant improvement in naturalness compared to modeling phonemes.

For intelligibility, the results show that there is still no statistically significant difference between Group A and Group B, which shows that in Mandarin-English code-switched speech synthesis primarily in Mandarin, using code-switching datasets and Mandarin datasets for pre-training yields similar effects. Therefore, if there is no code-switching dataset, an open-source Mandarin dataset can be used as the pre-training model dataset. In addition, although the results for Group A and Group D show that the mean difference is not significant, the p-value is 0.0893, which is very close to 0.05, indicating that there are still some potential differences. Additionally, other groups are significantly different from Group D, which further validating that modeling PF provides a significant improvement in intelligibility compared to modeling phonemes.

6 Discussion

This chapter will provide a detailed discussion of the study's results, validation of the research hypotheses, limitations, and potential future research directions. The basic framework of this chapter is as follows: In Section 6.1, I will discuss the results of the study and the verification of the research hypotheses. In Section 6.2, I will discuss the limitations of the study as well as future research directions.

6.1 Result and Validation of Hypothesis

The research question of this study is: How can code-switching speech synthesis between Mandarin and English be effectively implemented based on FastSpeech2(Ren et al., 2020), to synthesize intelligible and natural-sounding Mandarin-English code-switched speech? Based on the research by (Zhang et al., 2019) and (Staib et al., 2020), this paper hypothesizes that direct modeling of Mandarin and English phonemes can synthesize intelligible speech; modeling the phonological features of Mandarin and English can produce both intelligible and natural speech.

In response to hypothesis 1, that is, directly modeling Mandarin and English phonemes can effectively synthesize intelligible Mandarin-English code-switched speech, this study designed three groups of experiments: Group A, Group B, and Group C. The results show that the understandability MOS of Group A and Group B both exceed 4 points, indicating that most participants could understand and comprehend the meanings of the sentences. This is consistent with the study of (Zhang et al., 2019) who pointed out that modeling phonemes helps promote the sharing of cross-lingual models.

On the contrary, Group C had a lower MOS score of 2.42 for intelligibility, indicating that participants found it difficult to understand the meaning of most sentences. This may be because Group C had never been trained using a dataset with Mandarin-English code-switching, resulting in a lack of cross-language consistency in the model when dealing with Mandarin-English code-switching. Specifically, Mandarin and English have different phonological systems(Liao and Shen), and when the model separately models the phonemes of these two languages, it learns their respective characteristics independently and tends to forget the features learned during pre-training. This ultimately results in synthesized speech that is unclear or has poor intelligibility. In addition, since the dataset used for the fine-tuning part is English, the synthesized speech has a very obvious English speaker's accent, leading to poor intonation in the Mandarin parts.

Therefore, regarding hypothesis 1, this study concludes that it is generally valid, but a restriction needs to be added: Modeling the phonemes of Mandarin and English can effectively synthesize intelligible code-switched Mandarin-English speech, provided that the model is appropriately trained using a code-switching dataset. This ensures that the model can correctly handle cross-language scenarios.

In response to hypothesis 2, that is, intelligible and natural Mandarin-English code-switched speech can be synthesized by modeling phonological features (PF), this study designed an experiment with Group D. The results indicate that Group D achieved an intelligibility MOS score of 4.68 and a naturalness MOS score of 3.81, which shows that the participants can understand most of the content of the synthesized speech well and are relatively satisfied with the naturalness. Therefore, this study concludes that hypothesis 2 has been validated.

The study in (Staib et al., 2020) mainly trained on languages with similar phonological systems and explored the model's ability to handle completely unseen languages in synthetic training. However, this study mainly explored the feasibility of using PF to model languages with large phonological differences such as Mandarin and English. For distant languages such as Mandarin and English, the model can be trained to model PF to generate intelligible and natural code-switching speech. Then for some more similar languages, such as English and Spanish, or Russian and Finnish, the effectiveness of PF should be even more significant. (Demirsahin et al., 2018) also suggested that for various Southeast Asian languages, using phonological features (PF) to model the system can support multilingual text-to-speech synthesis.

Additionally, although Group A and Group B are similar in terms of intelligibility MOS scores, they differ significantly in naturalness MOS results. The main difference between the two groups in the training phase is the pre-training dataset used: Group A used the TAL_CSASR dataset, which is designed for ASR tasks and has lower quality, while Group B used the Chinese Standard Mandarin Speech Corpus, which is designed for TTS tasks and has higher quality. Aside from this, the number of training steps, training parameters, and the dataset used for fine-tuning during the training phase were consistent between the two groups. Therefore, the difference in naturalness MOS results between the two groups may be due to Group A using a lower-quality dataset during the pre-training phase.

By repeatedly comparing the synthesized audio, it was found that the synthesized audio of Group A had obvious noise mainly in the English part, which made the audio sound more artificial. This may be because the dataset CS_500 used in the fine-tuning stage is mainly Chinese, with fewer English words, thus offering limited improvement to the quality of the English parts. At the same time, since this study primarily focuses on synthesizing Mandarin-English code-switching speech with an emphasis on Mandarin, the synthesized audio is mainly in Mandarin. Therefore, even though Group B did not use a large amount of English data for training, its final output still sounded relatively natural.

Although the quality of TAL_CSASR is poor and the synthesized speech contains more noise without fine-tuning, the meaning of the generated sentences is still understandable. By comparing the synthesized audio, it was found that after fine-tuning the pre-trained model with CS_500, the low-quality properties of the synthesized speech are significantly improved. (Zhang and Lin, 2021) also pointed out that fine-tuning ASR datasets with high-quality datasets can significantly improve the quality of synthesized audio. Therefore, this study believes that when dealing with low-resource languages in TTS tasks, it is possible to consider using lower-quality ASR data to pre-train the model, and then fine-tune it with a small amount of high-quality datasets to improve the quality of the final synthesized audio.

6.2 Limitation and Future Work

This study conducted experiments and discussions on how to effectively implement Mandarin-English code-switched speech synthesis based on the FastSpeech2, to synthesize understandable and natural Mandarin-English code-switched speech. Although this study largely validated the two proposed hypotheses, it still has certain limitations.

Although modeling phonological features has generally enabled the production of natural and understandable speech in this study, the naturalness MOS results show that there is still room for improvement in naturalness. There are two main reasons for reflection. First, due to time constraints, this study did not specify some special Mandarin pronunciation rules in detail. Therefore, the tones on some words are not pronounced accurately, causing the overall audio sound less natural. Secondly, the definition of naturalness in the design of the MOS evaluation was relatively vague. Although the participants' evaluation of the audio synthesized by Group D was basically consistent, it would be better if a clearer explanation of naturalness could be given, such as pointing out that naturalness refers to the fluency, rhythm, and pitch of the entire sentence. Then the results may be more accurate. In addition, due to time and resource constraints, this study failed to perform noise reduction on the TAL_CSASR dataset. If noise reduction tools can be used to process this dataset, it could potentially improve the overall performance of Group A.

Based on the results and limitations of this study, future research directions include:

- Performing noise reduction on the TAL_CSASR dataset and comparing the results with Group A to explore whether noise reduction tools can significantly improve the quality of low-quality datasets and thus enhance the quality of synthesized audio.

- Exploring methods for modeling PF to determine if it is possible to generate intelligible and natural Mandarin-English code-switching speech using only monolingual datasets in Mandarin and English.

- Further study the method of PF modeling, explore the impact of the number of languages in the pre-trained model on its effectiveness, and identify which low-resource languages can benefit from these methods.

7 Conclusion

This study is primarily based on FastSpeech2 and has focused on Mandarin-English code-switching speech synthesis, validating the following two hypotheses:

- Modeling Mandarin and English phonemes directly can effectively synthesize understandable Mandarin-English code-switched speech, provided that the model needs to be properly trained using a Mandarin-English code-switched dataset.
- Modeling Phonological Features (PF) can synthesize understandable and natural Mandarin-English code-switching speech.

In addition to the above two hypotheses, this study also believes that when dealing with code-switching situations or low-resource languages, it is possible to consider using low-quality ASR data to pre-train the model in the TTS tasks. This can then be combined with a small amount of high-quality data for fine-tuning, to mitigate the low-quality attributes of the dataset and thereby improve the quality of synthesized speech.

Since the primary goal of this study is to synthesize Mandarin-dominant Mandarin-English code-switching speech, the model was not pre-trained using a pure English dataset.

Bibliography

Zexin Cai, Yaogen Yang, and Ming Li. Cross-lingual multispeaker text-to-speech under limited-data scenario. *arXiv preprint arXiv:2005.10441*, 2020.

Yuewen Cao, Xixin Wu, Songxiang Liu, Jianwei Yu, Xu Li, Zhiyong Wu, Xunying Liu, and Helen Meng. End-to-end code-switched tts with mix of monolingual recordings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6935–6939. IEEE, 2019.

Cecil H Coker. A model of articulatory dynamics and control. *Proceedings of the IEEE*, 64(4): 452–460, 1976.

Isin Demirsahin, Martin Jansche, and Alexander Gutkin. A unified phonological representation of south asian languages for multilingual text-to-speech. In *SLTU*, pages 80–84, 2018.

Heinz J. Giegerich. *Phonological features, part 1: the classification of English vowel phonemes*, page 89–111. Cambridge Textbooks in Linguistics. Cambridge University Press, 1992.

Alexander Gutkin, Martin Jansche, and Tatiana Merkulova. Fonbund: A library for combining cross-lingual phonological segment data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Keith Ito and Linda Johnson. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

Javier Latorre, Koji Iwano, and Sadaoki Furui. Polyglot synthesis using a mixture of monolingual corpora. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–1. IEEE, 2005.

Bo Li and Heiga Zen. Multi-language multi-speaker acoustic modeling for lstm-rnn based statistical parametric speech synthesis. In *Interspeech*, pages 2468–2472, 2016.

Bo Li, Yu Zhang, Tara Sainath, Yonghui Wu, and William Chan. Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5621–5625. IEEE, 2019.

Hui Liang, Yao Qian, and Frank K Soong. An hmm-based bilingual (mandarin-english) tts. *Proceedings of SSW6*, 2007.

Maozhen Liao and Nanyan Shen. A contrastive study of phonetic variations in english and chinese.

Florian Lux and Ngoc Thang Vu. Language-agnostic meta-learning for low-resource text-to-speech with articulatory features. *arXiv preprint arXiv:2203.03191*, 2022.

Florian Lux, Julia Koch, Antje Schweitzer, and Ngoc Thang Vu. The IMS Toucan system for the Blizzard Challenge 2021. In *Proc. Blizzard Challenge Workshop*, volume 2021. Speech Synthesis SIG, 2021.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502, 2017.

Huaiping Ming, Yanfeng Lu, Zhengchen Zhang, and Minghui Dong. A light-weight method of building an lstm-rnn-based bilingual tts system. In *2017 International Conference on Asian Language Processing (IALP)*, pages 201–205. IEEE, 2017.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32, 2019.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

P Seeviour, J Holmes, and M Judd. Automatic generation of control signals for a parallel formant speech synthesizer. In *ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 690–693. IEEE, 1976.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj 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 (ICASSP)*, pages 4779–4783. IEEE, 2018.

S Sitaram, KR Chandu, SK Rallabandi, and AW Black. A survey of code-switched speech and language processing. arxiv 2019. *arXiv preprint arXiv:1904.00784*.

Sunayana Sitaram and Alan W Black. Speech synthesis of code-mixed text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3422–3428, 2016.

Marlene Staib, Tian Huey Teh, Alexandra Torresquintero, Devang S Ram Mohan, Lorenzo Foglianti, Raphael Lenain, and Jiameng Gao. Phonological features for 0-shot multilingual speech synthesis. *arXiv preprint arXiv:2008.04107*, 2020.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*, 2021.

Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12, 2016.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

Dan Wells and Korin Richmond. Cross-lingual transfer of phonological features for low-resource speech synthesis. In *Proceedings of the 11th Speech Synthesis Workshop, Budapest, Hungary*, pages 160–165, 2021.

Zhizheng Wu, Oliver Watts, and Simon King. Merlin: An open source neural network speech synthesis system. In *9th ISCA Speech Synthesis Workshop*, pages 202–207, 2016.

Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. In *Sixth European conference on speech communication and technology*, 1999.

Heiga Zen, Norbert Braunschweiler, Sabine Buchholz, Mark JF Gales, Kate Knill, Sacha Krstulovic, and Javier Latorre. Statistical parametric speech synthesis based on speaker and language factorization. *IEEE transactions on audio, speech, and language processing*, 20(6):1713–1724, 2012.

Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing*, pages 7962–7966. IEEE, 2013.

Haitong Zhang and Yue Lin. Improve cross-lingual voice cloning using low-quality code-switched data. *arXiv preprint arXiv:2110.07210*, 2021.

Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. *arXiv preprint arXiv:1907.04448*, 2019.

Shengkui Zhao, Trung Hieu Nguyen, Hao Wang, and Bin Ma. Towards natural bilingual and code-switched speech synthesis based on mix of monolingual recordings and cross-lingual voice conversion. *arXiv preprint arXiv:2010.08136*, 2020.

Xuehao Zhou, Xiaohai Tian, Grandee Lee, Rohan Kumar Das, and Haizhou Li. End-to-end code-switching tts with cross-lingual language model. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7614–7618. IEEE, 2020.

Appendices

A Questionnaire Survey

The questionnaire used in this study consists of 40 sets of questions, each set containing two questions regarding intelligibility and naturalness. Considering that the participants in this study are native Chinese speakers who are proficient in English, the questionnaire is provided in both Chinese and English to facilitate better understanding by the participants. The specific content of the questionnaire can be found at the provided link[5]. Below is an introduction and a sample question from the questionnaire:

Introduction

Thank you for participating in this experiment! Your involvement will make a significant contribution to the field of multilingual synthetic speech. This experiment will take about 15-20 minutes of your time.

In each phase of this experiment, you will listen to a synthetic voice mixed with English and Chinese. Please rate the voice based on the following criteria:

1. Does the voice sound human-like to you?

2. Can you clearly hear and understand the sentences spoken by the voice?

Remember, there are no right/wrong answers. Just do your best and follow your intuition.

说明

感谢您参加这个实验！您的参与将对多语言语音合成领域做出重大贡献。本次实验将占用您大约15-20分钟的时间。

在实验的每个阶段，您将听到一个中英文混合的语音。请根据以下标准对语音进行评分：

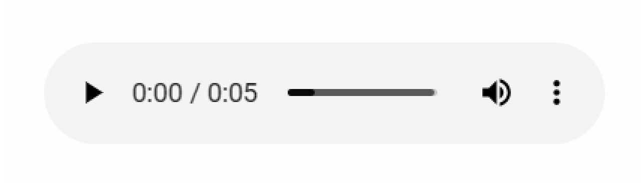
1. 您是否能够清晰听到并理解语音所说的句子?

2. 您认为这个语音听起来像人类的声音吗?

请记住，没有正确或错误的答案，只需按照您的直觉作答。

[5] https://rug.eu.qualtrics.com/jfe/preview/previewId/0c4e143c-6a55-48e2-96a7-9fe0163c5f12/SV_4V3mqLakihgcQ6y?Q_CHL=preview&Q_SurveyVersionID=current

Please listen to the audio and rate it based on your perception 请听下方的音频,并根据您的直觉对其进行评分



How natural does this audio sound to you? 你认为这个音频中的声音听起来自然吗?

- Bad 差
- Poor 一般
- Fair 良好
- Good 非常好
- Excellent 优秀

Can you clearly hear and understand the sentences spoken by the voice? 你能清楚地听到和理解音频中的句子吗?

- Completely Unintelligible 完全不能理解
- Mostly Unintelligible 大多难以理解
- Partially Intelligible 部分可理解
- Mostly Intelligible 大部分可理解
- Fully Intelligible 完全可理解