university of groningen

campus fryslân

# Comparative Study of Low Resource Language Manchu Speech Synthesis: Transfer Learning from Spanish vs. Mandarin Chinese

Shenghuan Ding

**University of Groningen - Campus Fryslân**

# Comparative Study of Low Resource Language Manchu Speech Synthesis: Transfer Learning from Spanish vs. Mandarin Chinese

**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. Phat Do** (Voice Technology, University of Groningen)
with the external supervisor being
**Dr. Joanna Dolińska** (University of Warsaw)

**Shenghuan Ding (S5743346)**

June 11, 2024

# Acknowledgements

I acknowledge the Center for Information Technology of the University of Groningen for their technical support and for providing access to the Hábrók high-performance computing cluster.

I would like to express my sincere gratitude to Dr. Do from the Campus Frisian of the University of Groningen for serving as my supervisor, providing valuable advice, suggesting methodologies, datasets, and assisting in resolving errors encountered during training and optimizing the models.

I am thankful to Dr. Dolińska from the University of Warsaw for inspiring the research questions of my thesis and providing guidance, suggestions, and assistance in the process of data collection.

Special thanks to Xiaoxu Qi, a native speaker of Manchu from Heilongjiang University in China, for providing recordings for our research. I also extend my appreciation to Tianmai Zhang from Heilongjiang University for providing literature support in Chinese and Manchu, and to the Manchu Research Institute at Heilongjiang University for their assistance.

I would also like to thank all the faculty members of the Voice Technology group at the University of Groningen for their support and assistance.

# Abstract

This study aims to explore the effect of transfer learning from Spanish and Mandarin Chinese in Manchu speech synthesis and determine which language can achieve better synthesis results. We experimentally compare the Manchu speech synthesis effect of transfer learning from Spanish and Mandarin Chinese and analyze the impact of speech features between different languages on the synthesis results. Our hypothesis is that since Mandarin Chinese has more loanwords and possible phoneme similarities, transfer learning from Mandarin Chinese will achieve better results than Spanish.

To verify the hypothesis, we first collected speech data from Spanish and Mandarin Chinese and used them to build a speech synthesis system based on the FastSpeech 2 model. Then, we used Montreal Forced Aligner (MFA) to align speech and text to ensure the consistency of training data. Then, we used transfer learning methods to apply the trained Spanish and Mandarin Chinese models to Manchu speech synthesis. Finally, we evaluated the synthesis effect of transfer learning from different languages and analyzed its accuracy and naturalness.

The experimental results show that the Manchu speech synthesis effect of transfer learning from Mandarin Chinese is better than that of Spanish. This suggests that the language features and phoneme similarity between Mandarin Chinese and Manchu play a key role in the synthesis effect. In addition, we also found that despite the difference in the gender of the voices between the Mandarin Chinese and Spanish recordings (female for Mandarin Chinese and male for Spanish), this variation did not significantly impact the synthesis results.

The results of this study support our hypothesis that transfer learning using Mandarin Chinese will produce better Manchu speech synthesis results. This finding is of great significance for improving the quality and efficiency of speech synthesis for low resource languages and provides a useful reference for future related research.

# Contents

# 1 Introduction

With the development of science and technology, speech synthesis technology has played an important role in various fields. From intelligent assistants to automated customer service systems, speech synthesis technology makes human-machine interaction more natural and efficient. Intelligent assistants such as Siri and Alexa use speech synthesis technology to provide users with convenient voice command services, greatly improving the user experience. In the medical field, speech synthesis technology helps voiceless patients ”speak” through devices, significantly improving their quality of life. In this Internet era where self-media and short videos are popular, dubbing and explaining videos through artificially synthesized voices has become an important form of leisure and entertainment for people. These applications demonstrate the broad potential and practical value of speech synthesis technology. However, for languages with scarce resources, the development of this technology faces huge challenges. Due to the limited number of speakers and the lack of sufficient speech data, these languages lag far behind resource-rich languages in the field of speech synthesis. For example, Manchu, a language once widely used in Northeast China, is now almost extinct. The scarcity of speech data and the interruption of inheritance make it extremely difficult to develop a speech synthesis model for Manchu. This not only affects cultural inheritance, but also limits the application and development of related technologies.

Manchu is a language that is spoken in northeast China, and it used to be the official language of the Qing dynasty (AD 1644–1912). According to UNESCO, by the end of the Qing Dynasty, there were at least one million native speakers of Manchu, but now, among nearly ten million ethnic Manchus in Northeast China, there are only less than 1,000 native Manchu speakers. (Zhu et al., 2018) In addition, these native speakers are mostly old generation and the majority of their descendants are not able to speak that language. Getting high-quality Manchu language recordings with transcriptions is always difficult. As a result, it is challenging to generate a Manchu synthesis model using a small dataset. However, transfer learning methods can make it easier. We can train a model on a large language dataset first and then transfer it to the under-resourced Manchu language dataset. In this thesis, we used 6.5 hours of Spanish and 6.5 hours of Mandarin Chinese as the resource-rich languages for transfer learning. Here we emphasize that although 6.5 hours is not considered “resource-rich”, it is enough to generate a very good speech output. The reason why we only use 6.5 hours of speech is that our datasets CSS10 only has 6.5 hours of audio from a single Chinese speaker, so we cut the 24 hours of Spanish audio to 6.5 hours to complete this experiment. More importantly, we went to Heilongjiang province, China, and we recorded a 30-minute recording with 500 sentences of Qi Xiaoxu, a native Manchu speaker. By using Montreal Forced Aligner and FastSpeech 2 models, we built a simple Manchu text-to-speech model. We compared the results of the generated audios and evaluated the accuracy and naturalness of speech generation under four different conditions.

Now that a brief motivation for this research has been presented, the structure of the thesis is the following: subsection 1.1 introduces the research question posed along with a hypothesis on the outcome of the research. Section 2 introduces the linguistic-related background, a survey on text-to-speech synthesis and transfer learning on low-resource languages. In section 3, the methodology is covered and the underlying models and data used are explained. Then, section 4 describes the data processing, the experimental setup, and the evaluation process. Section 5 describes the results obtained and compares them to the baseline. In section 6, I discuss the previously-mentioned results and our limitations in detail. Lastly, section 7 summarizes the thesis and presents the conclusions drawn, along with recommended future work.

## 1.1 Research Question and Hypothesis

In light of the preceding discussion, the research question at the core of this study can be formulated as follows:

> **Which language yields the most effective transfer learning results for Manchu synthesis: Spanish (more common consonants) or Mandarin Chinese (more loan words and common vowels)?**

From which the following subquestions are derived:

- Is it possible to use just a few Manchu data to generate a decent text-to-speech model?
- The recordings in Mandarin Chinese are in a female voice, while those in Spanish are in a male voice. How does this difference affect the final synthesized speech?

Our hypothesis is that utilizing a training approach based on Mandarin Chinese, a resource-rich language, will lead to better outcomes in phoneme synthesis for Manchu than using Spanish, another resource-rich language..

# 2 Literature Review

In this literature review, we will focus on the phonetic features of Manchu, Mandarin, and Spanish, as well as the research progress of text-to-speech synthesis technology, especially the transfer learning on low resource languages method. First, we will introduce the history, current status, and phonetic features of Manchu. As an important minority language in China, Manchu once played an important role in the Qing Dynasty, but with the changes of the times and the influence of Han culture, Manchu gradually declined. We will focus on the vowel and consonant features of Manchu, as well as its history and its current status. Second, we will explore the phonetic features of Mandarin. As one of the official languages of China, Mandarin has a wide range of use and influence in China. We will introduce the characteristics of vowels, consonants, tones, and homophones in Mandarin, providing a basis for subsequent research. Subsequently , we will focus on Spanish and its phonetic features. As an important world language, Spanish has a long and complex history. We will introduce its phonetic features such as vowels, consonants, stress.

## 2.1 Research on the current state and phonemes in Manchu language

In this section, we provide a detailed introduction to the Manchu language, especially its phonological features.

China is a multi-ethnic country with 55 ethnic minorities. Most of the minority ethnic-groups have different cultures, languages and lifestyles from the Han Chinese, the main ethnic group in China. Manchu is a big major minority ethnic group in China, with approximately 10 million population. Most of the Manchu people live in the northeast part of China, Heilongjiang, Jilin, Liaoning and Inner Mongolia. In 1636, the Manchu leader Nurgachi established the Qing Dynasty. Their mother language, Manchu, became the most impactful language except Chinese language and its dialects that Han people used at that time. The Manchu language, written in Mongolian script, has left behind many important historical documents in Manchu. With the demise of the Qing Dynasty, the Manchu lifestyle and language were gradually replaced by Han culture. Rhoads proposed that the Manchus absorbed a great deal of Han culture during the long period of Qing rule, but they still maintained some unique cultural characteristics and identity. (Rhoads, 2015) However, with the outbreak of the Xinhai Revolution, the Qing government was forced to retreat, which eventually led to the end of the monarchy. After the Xinhai Revolution period, Manchu language became distinct.

In 2018, Zhao put forward that as the primary rulers of the Qing Dynasty, the Manchus designated the Manchu language as the 'national language' ('Guoyu') and 'Qingwen.' It became an important tool, medium, and bridge for internal administration and foreign affairs, playing a significant role in social development and cultural prosperity. (A. Zhao, 2018) In the early 20th century, under the significant influence of Han Chinese culture and the Chinese language, the Manchu language gradually assimilated and declined. Following the overthrow of the Qing Dynasty, ruled by the Manchus, Manchu was phased out from political, economic, and other domains, and was replaced by Chinese. As late as 2002, those who spoke purely Manchu as their native language and did not speak Chinese had passed away. Nowadays, those who speak Manchu are bilingual in both Manchu and Chinese, with Manchu inevitably influenced by Chinese. Thirdly, the enthusiasm of the Manchu people for their native language is not high. Nowadays, the number of people who are able to speak Manchu is less than 1,000.(Zhang & Peng, 2021)

After introducing the history and current status of Manchu, we will focus on the phonetic features of the language. In phonetics, vowels refer to the sounds produced when air flows through the oral cavity without obstruction. Vowels are closely related to the pharyngeal cavity, oral cavity, nasal cavity, etc. The change of tongue position in the oral cavity causes the change of the shape of the vocal cavity, forming different vowels. The Handbook of the International Phonetic Association on page 10 (Association, 1999, p. 10) states that sounds with the vocal tract closed or nearly closed are consonants, sounds with the vocal tract open are vowels, and sounds with no obstruction to the airflow through the mouth are vowels. These characterizations are relatively objective and comprehensive, reflecting the characteristics of vowels and covering some special vowels.

According to 800 Modern Manchu Sentences compiled by Ji et al. in 1989, different villages have developed different pronunciations due to the endangered status of Manchu. The book introduces that Manchu generally has 6 main vowels, 10 diphthongs and 2 triphthongs. (Ji, Zhao, & Bai, 1989) However, Qi Xiaoxu, a native speaker who participated in the recording, believes that he pronounces more than just these six vowels. Based on discussions and comparisons with IPA pronunciation, we added three vowels to the pronunciation dictionary because the speaker in the recordings pronounced them. The specific vowels for Manchu language can be seen in the figure 1.

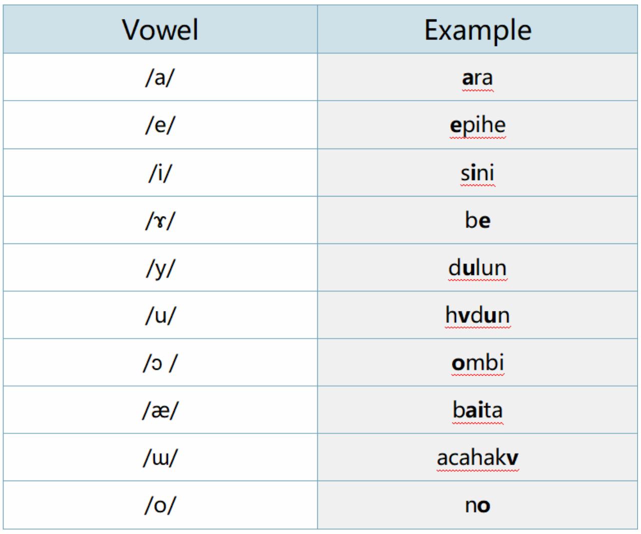
For Manchu consonants, there are many scholars who put forward different points of view. The main reasons are that there are only a few native speakers in several different villages and the language is dying. Younger generation do not have the access to learn the language. Only some of the older generation have the ability to have a conversation with Manchu. ”800 Modern Manchu Sentences” written by Ji and colleagues, (Ji et al., 1989) the book contains the spoken Manchu of Sanjiazi, which is the biggest Manchu native speakers’ village. The consonant phonemes of the language are summarized into 28 consonants. However, the standard of Manchu has always been uncertain, and different scholars have different views. Zhao Jie’s ”Phonemic Analysis of Tailai Manchu” (J. Zhao, 1987) take the Manchu language of Yibuqi Village, Tailai County, Heilongjiang Province as the research object, and summarize 24 consonant phonemes. The spoken materials present a relatively comprehensive overview of the Sanjiazi Manchu language. In our thesis, we combined all of the above information, with the main reference being the ground-truth recording provided by Qi Xiaoxu. The consonants that appear in Qi Xiaoxu’s recording are represented by the figure 2 and 3.

## 2.2 Research on Mandarin Chinese and Spanish phonology

Transitioning from our exploration of Manchu phonology, we delve into an analysis of Mandarin Chinese and Spanish phonology. These languages, with their rich histories and widespread usage, offer distinct yet fascinating insights into linguistic diversity.

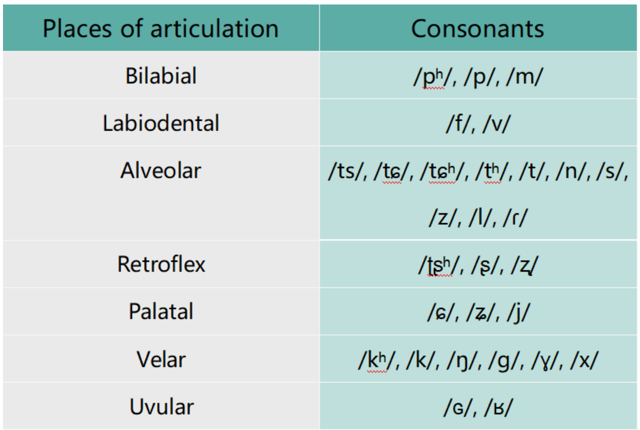
### 2.2.1 Mandarin Chinese and Mandarin Chinese Phonology

Mandarin Chinese is spoken mainly in China, with over 1.4 billion speakers. In China, there are over 55 minority ethic groups. While there exist numerous other languages (e.g., Tibetan, Miao, and Mongolian) and dialects (such as Wu, Yue, and Gan) are spoken, almost all of the speakers of these languages and dialects are still able to use Mandarin Chinese. Mandarin is written in Chinese characters (hanzi), which does not show the pronunciation directly. It is the biggest logographic language in the world. However, pinyin is commonly used in order to notate the pronunciation of Mandarin Chinese. (Wang & Andrews, 2021) Pinyin employs the 26 letters of the Latin alphabet,



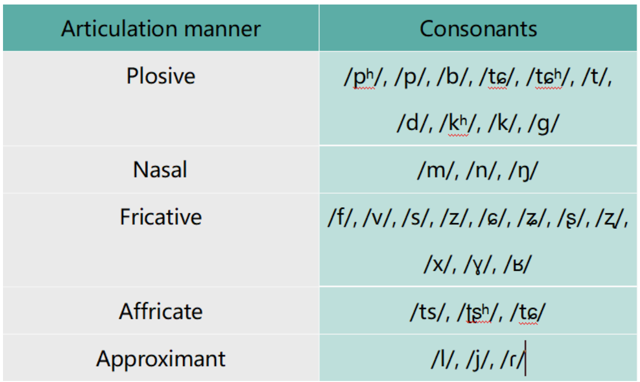
| Vowel | Example |
|---|---|
| /a/ | **a**ra |
| /e/ | **e**pihe |
| /i/ | s**i**ni |
| /ɤ/ | b**e** |
| /y/ | d**u**lun |
| /u/ | h**v**d**u**n |
| /ɔ / | **o**mbi |
| /æ/ | b**ai**ta |
| /ɯ/ | acahak**v** |
| /o/ | n**o** |

Figure 1: Manchu Vowels



| Places of articulation | Consonants |
|---|---|
| Bilabial | /pʰ/, /p/, /m/ |
| Labiodental | /f/, /v/ |
| Alveolar | /ts/, /tɕ/, /tɕʰ/, /tʰ/, /t/, /n/, /s/, /z/, /l/, /ɾ/ |
| Retroflex | /ʈʂʰ/, /ʂ/, /ʐ/ |
| Palatal | /ɕ/, /ʑ/, /j/ |
| Velar | /kʰ/, /k/, /ŋ/, /g/, /ɣ/, /x/ |
| Uvular | /ɢ/, /ʁ/ |

Figure 2: Manchu consonants classified by places of articulation



| Articulation manner | Consonants |
|---|---|
| Plosive | /pʰ/, /p/, /b/, /tɕ/, /tɕʰ/, /t/, /d/, /kʰ/, /k/, /g/ |
| Nasal | /m/, /n/, /ŋ/ |
| Fricative | /f/, /v/, /s/, /z/, /ɕ/, /ʑ/, /ʂ/, /ʐ/, /x/, /ɣ/, /ʁ/ |
| Affricate | /ts/, /ʈʂʰ/, /tɕ/ |
| Approximant | /l/, /j/, /ɾ/ |

Figure 3: Manchu consonants classified by manners of articulation

which are familiar to people in other countries and China’s own minority ethnic groups’ languages. In this part, we will introduce some literature surveys of Mandarin phonology.

First, Mandarin Chinese has six basic vowels. They are “a”, “o”, “e”, “i”, “u” and “ü” in pinyin. In addition, there are also some diphthongs or triphthongs in Mandarin Chinese, for example, “ai”, “iao” in pinyin. (Odinye, 2022)

Second, for consonants, there are six plosives, two nasals, five fricatives, six affricates, one lateral approximant in Mandarin Chinese.

Third, Mandarin Chinese is a tone-based language, and there are five basic tones. Same sounds with different tones convey different meanings. For example, the sound “ma” with the first tone means “mother”, but with the third tone means “horse”.(Jongman et al., 2006)

Fourth, Mandarin Chinese has a large number of homophones. For example, “xie3” in Mandarin can both mean “blood” and “to write”. Besides that, there are also many heteronyms in Mandarin Chinese. Heteronym means a word or a single Chinese character has more than one correct pronunciation, and different pronunciations convey different meanings. For instance, a Chinese character has two pronunciations: “xing2” and “hang2”. The first one means “to walk” or “yes”, but the second one means “row”.

### 2.2.2 Spanish and Spanish Phonology

Spanish, as an important world language, has a long and complex history. According to R.J. Penny in his book ”A History of the Spanish Language”, the evolution of Spanish can be traced back to the 3rd century BC, when the Romans conquered the Iberian Peninsula and introduced Latin. As a member of the Latin language family, Spanish is derived from Latin, specifically the spoken form of Vulgar Latin. We can still find a lot of similarities between modern Spanish and other languages from the Latin family, like French, Italian, Portuguese and Romanian. As the Roman Empire expanded, Latin spread widely across the Iberian Peninsula, gradually incorporating elements of local Celtic-Iberian languages. (Penny, 2002) This fusion resulted in early Iberian Romance languages. During the early Middle Ages, the Iberian Peninsula experienced several foreign invasions, including Gothic and Moorish rule. The Moors brought Arabic and had a significant influence on Spanish, especially in terms of vocabulary, and many Arabic words were incorporated into Spanish. As the Christian kingdom gradually regained its lost territories, Spanish began to be standardized in Castile and became the official language. At the end of the 15th century, Castilian, the predecessor of modern Spanish, became the common language of the entire Iberian Peninsula with the expansion of the Kingdom of Castile. In the 15th century, the global spread of Spanish began with colonial expansion. With the arrival of the Spanish conquistadors, Spanish was brought to parts of the Americas, Africa, and Asia, becoming the dominant language in many areas. Now, there are a large number of native Spanish speakers in the American continent and Africa. Spanish is one of the most important languages in the world, with over 400 million native speakers. Its pronunciation, vocabulary and grammar vary from region to region, resulting in multiple dialects and variants.

Having discussed the historical evolution of the Spanish language, the subsequent section will elucidate its phonetic features. First, Spanish has five basic vowels. They are represented as ”a”, ”e”, ”i”, ”o” and ”u”. In addition, there are some double or triple vowels in Spanish, such as “Ciudad” (city) and “Estudiáis” (study in second person plural). (Hualde, 2005)

Second, As for the consonants, according to Hualde, Spanish has six plosives, three nasals, five fricatives, four affricates, one trill, one tap and one rim consonant.

Third, Spanish is not a tonal language, but stress plays an important role in sentences and has the function of differentiating word meanings. Stress in Spanish usually occurs on the penultimate syllable (if the word ends in a vowel or -n, -s) or the last syllable (if the word ends in a consonant, except -n, -s).

## 2.3 Research on Text-to-Speech Synthesis

Currently, our digital life almost can not leave speech synthesis technology. We can use the automatic speech recognition systems to convey our words and information without writing and typing. We can remote control our cell phone or our intelligence house by saying some commands to the state of the art intelligence assistants such as Alexa and Siri. Those assistants generated their voice like a real human being. In addition, we can use speech synthesis technology to record videos or teaching videos, which increases people's efficiency.

In 2021, Tan and colleagues made a systematic survey of the history of speech synthesis. In that paper, Tan et al proposed that the idea of speech synthesis is not new since the days of machine learning. The development of Text to Speech (TTS) can be traced back to the 12th century, when people began trying to build machines that could synthesize human speech. In the second half of the 18th century, Hungarian scientist Wolfgang von Kempelen constructed a speaking machine that used bellows, springs, bagpipes, and resonance boxes to produce simple words and short sentences.(X. Tan et al., 2021)

In the second half of the 20th century, the first computer-based speech synthesis system appeared. Early methods included pronunciation synthesis, formant synthesis, and concatenation synthesis. Pronunciation synthesis generates speech by simulating the behavior of human vocal organs. Although it is theoretically the closest to human speech generation. However, this synthesized speech is not natural and has poor flexibility. To solve these shortcomings, Statistical Parametric Speech Synthesis (SPSS) is proposed. By generating the necessary acoustic parameters and then restoring the speech, SPSS has the advantages of being more natural, flexible and low in data cost, but the generated speech still has some artifacts. With the rapid development of neural networks and deep learning, neural networks have been gradually introduced into SPSS since the 2010s, such as models based on deep neural networks (DNN) and recurrent neural networks (RNN), modern neural TTS models (such as WaveNet ) generates waveforms directly from text or phoneme sequences, significantly improving speech quality. In recent years, end-to-end models (such as Tacotron and FastSpeech) have simplified the text analysis module, directly taking characters or phoneme sequences as input, and simplifying the processing of acoustic features, ultimately generating high-quality speech waveforms.

The key components of the TTS system are text analysis, acoustic models, and vocoders. Tan and colleagues briefly introduced these major components. Text analysis is responsible for converting the input text into a form suitable for processing, including text normalization, character-to-sound conversion, and prosody prediction; the acoustic model converts the linguistic features generated by text analysis into acoustic features, using methods such as HMM, DNN, RNN, and Transformer; the vocoder converts the acoustic features into the final audio waveform, and common vocoders include WaveNet, WaveRNN, and MelGAN. In addition, there are some fully end-to-end (E2E) TTS models that generate waveforms directly from characters or phonemes, simplifying the process of traditional systems. Now, E2E TTS models almost dominate the field of speech synthesis, because of its effectiveness.

In 2021, Mu and colleagues summarized that there are three key components of an E2E TTS model: Text front-end: processes input text and converts it into a format suitable for further processing; acoustic model: converts the language features generated by the text front-end into acoustic features; vocoder: converts the acoustic features into the final audio waveform.(Mu, Yang, & Dong, 2021)

Currently, the most widely used TTS models with good overall performance include Tacotron 2, FastSpeech2 and Glow-TTS. First, Tacotron 2 is a widely used E2E TTS system. In 2018, Shen and colleagues from University of California Berkeley and Google put forward this synthesis system. The system consists of a recursive sequence-to-sequence feature prediction network that maps character embeddings to mel-spectrograms, and then uses a modified WaveNet model as a vocoder to synthesize time-domain waveforms from these spectrograms. Tacotron 2 achieved a mean opinion score (MOS) of 4.53, which is comparable to the 4.58 for professionally recorded speech. The whole system is mainly made by two key architectures: a sequence-to-sequence (S2S) feature prediction network and a wavenet vocoder. The S2S feature prediction network can generate mel-spectrogram frames from the input character sequence. The WaveNet vocoder in the Tacotron 2 model is responsible for converting the generated Mel-spectrogram frames into a speech waveform in the time domain. In short, The Tacotron 2 model can be trained directly from the audio without relying on complex feature engineering and modifying, and synthesis natural sound with high quality close to that of natural human voice. (Shen et al., 2018)

Different from traditional autoregressive TTS models, Glow-TTS introduced by Kim and colleagues from Kakao and Seoul University in 2020 is a flow-based generation model that can perform parallel TTS generation without the need for an external aligner. The model combines the characteristics of flow and dynamic programming to automatically search for the most likely monotonic alignment between the latent representations of text and speech. By enforcing hard monotonic alignment, the model achieves robust TTS that is applicable to long sentences, and uses generative flow to achieve fast, diverse, and controllable speech synthesis. Glow-TTS can achieve an order of magnitude speed improvement compared to the autoregressive model Tacotron 2 with comparable speech quality. Its design is inspired by the sequential nature of human reading text, that is, reading text in order without skipping any words. The core goal of the model is to achieve a monotonic and non-jumpy alignment between text and speech representations to generate conditional probability distributions. By combining the features of streaming and dynamic programming, Glow-TTS achieves parallel TTS synthesis without the need for an external aligner and is able to generate speech quickly, diversely, and controllably. (J. Kim et al., 2020)

FastSpeech 2 is characterized by its autoregressive architecture, which enables faster and more accurate speech synthesis. It is also the model we chose in our experiment. Before diving into FastSpeech 2 (FS2), let us first get acquainted with its predecessor, FastSpeech. Fast Speech (Ren et al., 2019) is a feed-forward network based on Transformer to generate mel-spectrogram in parallel for TTS. Unlike previous TTS model (eg. Tacotron2), they removed the attention mechanism. Instead, they applied a length regulator or duration predictor to directly control length. All of the improvements make the model, FS2, significantly accelerate the mel-spectrogram generation process by 270 times and the end-to-end speech synthesis by 38 times compared to autoregressive Transformer TTS.

However, Ren and his colleagues pointed out three main disadvantages of Fast Speech. (Ren et al., 2020) The two-stage teacher-student training pipeline makes the training process complicated. 2) The mel-spectrograms produced by the teacher model suffer from some information loss compared to the ground-truth ones, resulting in audio quality that is generally inferior to that synthesized from the ground-truth mel-spectrograms. 3) The duration extracted from the attention map of teacher model is not very accurate. Thus, Fast Speech 2 came into being to address these problems. In FastSpeech 2, a more straightforward model architecture is applied: ground-truth mel-spectrograms are used instead of predictions from the teacher model. In addition, Fast Speech 2 has more variance information, for example, adding pitch and energy decoder input. The function and structure of the variance adaptor can achieve these. This adapter is designed to add variation information to the phoneme hidden sequence to provide enough information to predict variant speech in TTS. The duration predictor takes the phoneme hidden sequence as input and predicts the duration of each phoneme, indicating how many mel frames this phoneme corresponds to. The mean square error (MSE) loss is used for optimization, and the Montreal forced alignment (MFA) is used to extract the phoneme duration to improve the alignment accuracy. The pitch predictor uses continuous wavelet transform to predict pitch. Then, the energy predictor uses the L2 norm of the amplitude of each short-time Fourier transform (STFT) frame calculated as the energy. The original energy value is predicted using the energy predictor and optimized using the MSE loss. These improvements make it a better quality than Fast Speech. However, FastSpeech 2 is still fast, robust, and more controllable than any other former FastSpeech model. In their paper, they designed some experiments to test the performance of FastSpeech 2.

They conducted experiments on the LJSpeech dataset to evaluate FastSpeech, and used the mean opinion score (MOS) to evaluate the audio quality of FastSpeech 2. Comparisons were made with other systems, including ground-truth, Tacotron 2, Transformer TTS, and FastSpeech. The results show that FastSpeech 2 outperforms FastSpeech in terms of audio quality, with a higher MOS value. In addition, FastSpeech 2 simplifies the training process of FastSpeech and reduces the training time. In terms of inference speed, FastSpeech 2 achieves a significant speed improvement in waveform synthesis. In addition, FastSpeech 2 and 2s provide more accurate variation information, such as pitch and energy, which helps to improve the quality of generated speech. It outperformed FastSpeech, Tacotron2 and Transformer TTS in this aspect. An analysis of pitch and energy was conducted, and the results showed that FastSpeech 2 is able to generate speech with a more natural pitch.

## 2.4 Research on Transfer Leaning on Low-resourced Languages TTS

What is transfer learning? An easy to understand example is that if you learn to ride a bicycle, it will be easier for you to learn to ride a motorcycle compared to someone who hasn't learned to ride a bicycle. This is because you transfer your experience from riding a bicycle, such as how to control balance and manage speed, to motorcycle riding. In 2020, Zhuang and colleagues (Zhuang et al., 2020) summarized the concept of transfer learning, that is " Transfer learning aims at improving the performance of target learners on target domains by transferring the knowledge contained in different but related source domains. In this way, the dependence on a large number of target-domain data can be reduced for constructing target learners. " This technique is now widely used in various fields, including machine learning and multi-task learning.

In the study by Sarker 2021, (Sarker, 2021) he proposed that we are living in a data age, and the world is surrounded by data. For example, during the COVID-19 pandemic , an unprecedented amount of data has been generated and collected, ranging from infection rates and mortality statistics to social distancing compliance and vaccine distribution records. In addition, machine learning (ML) algorithms find applications in diverse domains including cybersecurity, smart cities, healthcare, e-commerce, and agriculture. In such a data-rich environment, traditional machine learning approaches may face limitations due to the scarcity of labeled data or the need for extensive computational resources. This is where transfer learning emerges as a powerful technique to leverage knowledge learned from one domain or task and apply it to another, often related, domain or task.

After discussing transfer learning, the following part will delve into its application to low-resourced languages. Although there are no rich high-quality recordings, transcriptions, acoustic models, pronunciation dictionaries and other models and contents, we can still handle these problems. This exploration will highlight how transfer learning techniques can address the challenges faced by languages with limited available data and resources.

Tan and his colleagues proposed in 2018 (C. Tan et al., 2018) that deep learning, as a new classification platform, has garnered increasing attention from researchers in recent years and has been successfully applied in many fields. In some areas, such as bioinformatics and robotics, the high cost of data collection and expensive annotations make it difficult to build a large-scale, well-annotated dataset, which limits its development. Transfer learning relaxes the assumption that training data must be independent and identically distributed with the test data, which motivates us to use transfer learning to address the issue of insufficient training data.

In the multi-speaker speech synthesis field, Jia and colleagues proposed a multi-speaker text-to-speech (TTS) modeling approach by transfer learning in 2018 (Jia & et al., 2018) . They leveraged pre-trained models or representations on large datasets to improve the performance of models on target tasks with limited data. This approach not only accelerated the training process but also enhanced the generalization capability of models across diverse domains. Their model does not require either speaker identity labels for the synthesizer training data, nor high quality clean speech or transcripts for the speaker encoder training data. The output audio is human-like and the model has learned to utilize a realistic representation of the space of speaker variation.

In 2019, Tu and colleagues proposed a way of cross-lingual transfer learning in end-to-end TTS for low-resource languages. They prepared 24 hours of English language recordings from LJSpeech as the source language. Then, the recordings of low-resourced languages, German, Mandarin and French are limited, with only 30 minutes, 30 minutes and 15 minutes respectively.In their study, the researchers standardized the symbol representations of the four languages to ensure uniformity, facilitating more effective transfer learning. For instance, they might have aligned phonetic representations or linguistic features across languages to establish a common framework for the model to learn from. The researchers used a model called the Phoneme Transformation Network (PTN) to automatically learn the mapping of source symbols to target symbols through a pre-trained automatic speech recognition (ASR) system. In this process, PTN recognizes the phonemes of the source language and finds the corresponding phonemes of the target language, thereby achieving automatic conversion and mapping of symbols. The original Tacotron architecture is used as the end-to-end TTS model, which has an encoder-decoder structure with an attention mechanism. The spectrum analysis settings are the same as those in the original paper. Since the research goal is transfer learning with small data volumes, the Griffin-Lim algorithm is used as the waveform synthesizer. In addition, their ASR system uses a pure CNN model. The results show that this approach allows the model to generate more natural speech than models trained using only the target data, and it also achieves promising results compared to techniques that rely on a strong linguistic background. (Tu & et al., 2019)

In 2020, researchers from Egypt, Fady and colleagues (Fahmy, Khalil, & Abbas, 2020) built an Arabic TTS system by applying transfer learning techniques. They applied the Tacotron 2 model, which consists of a Spectrogram Prediction Network and WaveGlow Vocoder, to generate their voice. Similar to Tu and colleagues' research, Fady's team also uses LJSpeech as the source language data. The Arabic data is Nawar Halabis Arabic Dataset, with 2.41 hours of Arabic language recordings. It is worth noting that the researchers used Arabic characters with diacritics as input. This ensures that the model can correctly capture the phonetic features represented by these diacritics when generating speech, thereby improving the accuracy and naturalness of the generated speech. The results show that the model overcomes the differences in character-level embeddings and phonetic features between the two languages and successfully apply speech synthesis from English to Arabic into practice. Additionally, Tacotron 2 with WaveGlow shows the best performance for this task.

In 2020, Azizah and colleagues (Azizah, Adriani, & Jatmiko, 2020) applied the same approach into Indonesian low-resourced languages synthesizing. They proposed some advantages of end-to-end speech synthesis for low-resourced languages transfer learning: no phoneme-level alignment is required, reducing the need for engineered features; can be more easily conditioned on various attributes, such as speaker, language, or high-level features such as sentiment; more easily adaptable to new data; more powerful than multi-stage models where errors in each component accumulate. In their TTS models, there are three main components, these are Encoder, Decoder and Vocoder. Three TTS models were used in their study: T2, T2-mlms, and T2-mlms-gst. For the T2-mlms and T2-mlms-gst models, language embeddings, speaker embeddings, and style embeddings are also added to handle multi-language, multi-speaker, and transfer of style, intonation, and rhythm from reference audio. They applied hierarchical transfer learning, which means by transferring pre-trained model parameters layer by layer, cross-language, cross-speaker and style transfer multi-speech synthesis model training is achieved from high-resource languages to low-resource languages. For their experimental design, we first focus on the database. The research utilizes multiple publicly available speech corpora. The LJSpeech dataset, comprising 24 hours of English language data, serves as the source language. Additionally, the study incorporates three other datasets: TITML-IDN, an Indonesian (ID) corpus with an average duration of approximately 43 minutes per speaker; OpenSLR jv-ID, a Javanese (JV) corpus with an average duration of around 10 minutes per speaker; and OpenSLR su-ID, a Sundanese (SU) corpus with an average duration of about 7 minutes per speaker. These datasets provide diverse linguistic resources for training and evaluation purposes. In the experiment, the effects of aligned learning under two training schemes were compared: training from scratch and layer-by-layer transfer learning. At the same time, the speech synthesis produced by the TTS model trained using the transfer learning scheme was evaluated. The results showed that the hierarchical transfer learning approach significantly improved the efficiency and effectiveness of speech synthesis for low-resourced languages. By leveraging pre-trained models from high-resource languages like English and transferring their parameters layer by layer, the TTS models were able to adapt to the linguistic characteristics of Indonesian, Javanese, and Sundanese.

In 2022, Kim and colleagues applied large-scale unlabeled speech corpus for transfer learning of low-resourced languages. In other words, their TTS training model is performed using an unlabeled dataset based on a transfer learning framework. (M. Kim et al., 2022)

First, they proposed a concept called "pseudo phoneme". Pseudo phonemes are new tokens that contain information and are used for pre-training. Pseudo phonemes should have similar components to real phonemes and be obtained from a dataset containing only speech. To meet these requirements, the hidden representations of the pre-trained wav2vec2.0 model are used to classify the hidden representations through k-means clustering, and each speech frame is assigned a phoneme tag. Secondly, they applied the VITS model for training, which consisted of two components in their transfer learning architecture. The first component involved pre-training, a standard method for VITS, where pseudo-phoneme and pseudo-text encoders were utilized instead of real phoneme and text encoders. The training objective was adjusted to maximize the conditional likelihood of pseudo-phonemes given speech. The second component, fine-tuning, aimed to adapt the pre-trained model to the phoneme sequence.

Thirdly, the experiment aimed to verify the robustness and generalization ability of the proposed method under different training data scales. For single-speaker TTS, the results indicated that by leveraging large-scale unlabeled speech data for pre-training, this method not only improved generation quality and intelligibility but also significantly enhanced the model's robustness and data efficiency. In the case of zero-shot multi-speaker TTS, the proposed method consistently outperformed the baselines across various dataset sizes. Even with limited fine-tuning data, their method achieved lower character error rates and higher speaker similarity scores. In other words, despite the limited availability of data, including scenarios with only a few sentences or no data at all, the researchers demonstrated the capability to generate high-quality speech through transfer learning and fine-tuning.

In 2022, Do, Coler and colleagues also trained a MOS prediction model for low-resource languages based on wav2vec 2.0 and tested it. (Do et al., 2022) First, the researchers emphasize their evaluation method, that is Mean Opinion Score (MOS). In the field of TTS, MOS evaluation is a common method used to measure the naturalness of synthesized speech. Although MOS evaluation is widely used in TTS, it also has its challenges. These include resource consumption and evaluation subjectivity, which are very tricky. To address these challenges, automatic prediction of MOS has become a research direction that has attracted much attention. Models such as MOSNet use neural network structures to predict MOS from spectrograms, providing a more efficient method for TTS evaluation. The recent VoiceMOS challenge has promoted more related research by releasing the BVCC dataset. The datasets they applied were BVCC and SOMOS datasets. Their research aims to explore methods to optimize the MOS prediction model using the BVCC and SOMOS datasets, including determining the optimal amount of fine-tuning data and trying the fine-tuning effect of single listener data. They applied MOS data and experimented with 30 minutes of data in the LRL West Frisian (hereafter Frisian). They used FastSpeech 2 as the acoustic model and HiFi-GAN V1 as the vocoder, which are systems designed for neural TTS for low-resource languages. A total of 220 synthetic samples from 11 systems were scored (from 1 to 5). Then the researchers splitted the whole experiment into three scenarios: trained on BVCC, trained on SOMOS, and pre-trained on BVCC and then trained on SOMOS (BVCC, SOMOS, BVCC-SOMOS). Three accuracy metrics were analyzed for all scenarios: mean square error (MSE), linear correlation coefficient, and Spearman rank correlation coefficient (SRCC). The results showed that for zero shot prediction, BVCC-SOMOS dominated the test, and SOMOS outperformed BVVC. For fine tuned prediction, BVCC-SOMOS generally outperforms SOMOS and BVCC. It is worth noting that all accuracy metrics no longer improve after 30

In 2023, Liu (Liu, 2023) made a comparative analysis about Text-to-Speech synthesis by leveraging transfer learning to enhance model performance in low-resource environments. First, Wu compared the models that can be applied for low resource languages transfer learning. On pages 27-31 of the paper, he mentioned that FastSpeech2 can generate high-quality speech with less data by integrating rich voice change data. In addition, the author emphasized that Montreal Forced Aligner is a good way to automatically align pronunciations in audio with their corresponding text transcriptions on page 49-52. MFA has advantages such as high-precision triphone acoustic model and speaker adaptation, which helps researchers and developers achieve more accurate and efficient speech-to-text alignment when processing different datasets. He also emphasizes that rather than training a model from scratch, which would require a substantial amount of data, fine-tuning allows researchers to leverage existing learned features and adjust them to our specific task. This method can make significant savings in both time and computational resources while potentially achieving a higher model performance than training from scratch. Second part is the experiment. This experiment used seven pre-trained models: Tacotron2, FastSpeech2, FastPitch, Glow-TTS, VITS, SpeechT5, and OverFlow, fine-tuned on the LJSpeech dataset. Training and validation losses are the main metrics for evaluating model performance. The results show that FastSpeech2 does not generate traditional alignment maps, but predicts Mel-spectrograms, f0, and energy, which enhances the naturalness of speech. Tacotron2 has one of the longest training times, while SpeechT5 only takes 15 minutes. The FastSpeech2 model is also very effective. For the MOS evaluation, SpeechT5 performed best overall, while FastSpeech2 scored the lowest, indicating potential limitations. In the discussion part, the author concluded that transfer learning significantly improved the performance of the TTS model on small data sets. SpeechT5 performed outstandingly in fast training, low data dependence and sound quality. However, the research still had limitations in parameter selection and data set representativeness. For FastSpeech2, it is an important choice for transfer learning based on its flexibility and effectiveness.

At the end of our literature review, we will introduce our evaluation method Mean Opinion Score (MOS), which is widely utilized in assessing the quality of synthesized speech. MOS provides a quantitative measure by averaging subjective ratings assigned by human listeners based on their perception of speech quality. This method ensures a comprehensive evaluation encompassing various linguistic and acoustic aspects. (Rosenberg & Ramabhadran, 2017) MOS is a widely used method for assessing the quality of synthesized speech, providing a quantitative measure by averaging human listeners' subjective ratings of their perception of speech quality. The MOS method ensures a comprehensive assessment of various linguistic and acoustic aspects. In addition, by drawing on non-parametric statistical testing methods mentioned in the literature, we will calculate the statistical significance of MOS scores to better understand and interpret our evaluation results. However, this method also has limitations, such as participant bias and discourse bias, which means that different participants and different sentences will affect the results.

This concludes the literature review section, which provides an extensive overview of the prior research of some linguistic background of the three languages and an overview of the transfer learning on low resource languages.

While previous studies have contributed to our understanding of the linguistic diversity and phonological intricacies present in Manchu, Mandarin Chinese, and Spanish; the development of TTS synthesis model and the transfer learning method on low resource languages TTS.

Table 1: List of references for subsections 2.1-2.3, summarized

| Reference | Brief description | Sub |
|---|---|---|
| Rhoads (2015) | Manchus and Han: Ethnic relations and political power in late Qing and early republican China, 1861-1928 | 2.1 |
| A. Zhao (2018) | Research report on the current situation of Manchu, Hezhe, Xibe languages and cultures | 2.1 |
| Association (1999) | Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet | 2.1 |
| Ji et al. (1989) | 800 Sentences of Modern Manchu | 2.1 |
| J. Zhao (1987) | Phonemic analysis of Tailai Manchu | 2.1 |
| Wang and Andrews (2021) | Chinese Pinyin | 2.2 |
| Odinye (2022) | Phonology of Mandarin Chinese: A comparison of Pinyin and IPA | 2.2 |
| Jongman et al. (2006) | Perception and production of Mandarin Chinese tones | 2.2 |
| Penny (2002) | A history of the Spanish language | 2.2 |
| Hualde (2005) | The sounds of Spanish with Audio CD | 2.2 |
| Tan et al. (2021) | A survey on neural speech synthesis | 2.3 |
| Mu et al. (2021) | Review of end-to-end speech synthesis technology based on deep learning | 2.3 |
| Shen et al. (2018) | Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions | 2.4 |
| Kim et al. (2020) | Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search | 2.4 |
| Zhuang et al. (2020) | A Comprehensive Survey on Transfer Learning | 2.4 |
| Sarker (2021) | Machine Learning: Algorithms, Real-World Applications and Research Directions | 2.4 |
| Tan et al. (2018) | A Survey on Deep Transfer Learning | 2.4 |
| Jia et al. (2018) | Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis | 2.4 |
| Tu et al. (2019) | End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning | 2.4 |
| Fahmy et al. (2020) | A transfer learning end-to-end arabic text-to-speech (tts) deep architecture | 2.4 |
| Azizah et al. (2020) | Hierarchical transfer learning for multilingual, multi-speaker, and style transfer DNN-based TTS on low-resource languages | 2.4 |
| Kim et al. (2022) | Transfer learning framework for low-resource text-to-speech using a large-scale unlabeled speech corpus | 2.4 |
| Do et al. (2022) | Text-to-speech for under-resourced languages: Phoneme mapping and source language selection in transfer learning | 2.4 |
| Liu (2023) | Comparative Analysis of Transfer Learning in Deep Learning Text-to-Speech Models on a Few-Shot, Low-Resource, Customized Dataset | 2.4 |

# 3 Methodology

In this section, I will outline the methodology used to address the research question and validate the hypothesis on a high-level. First, in subsection 3.1, We will discuss the process of Manchu data collection. Next, subsection 3.2 will focus on the information to Mandarin Chinese and Spanish dataset. Subsection 3.3 will then introduce the pronunciation dictionaries we applied. Following that, in subsection 3.4, We will delve into the alignment model, Montreal Forced Aligner (MFA). Next, we will introduce our synthesis system FastSpeech 2 in 3.5. Subsection 3.6 will then elaborate on the evaluation method. Finally, in subsection 3.7, I will reflect on the ethical considerations inherent in this research.

## 3.1 Manchu data collection

As a result of the scarcity of native Manchu speakers and their reluctance to cooperate in providing speech samples, this aspect of our work has become quite challenging. Fortunately, Qi Xiaoxu from Inner Mongolia, China, has assisted us in collecting the data. He is a native Manchu language speaker. He grew up in a Manchu-speaking family environment and is currently pursuing a Master's degree in Manchu Language and Literature at Heilongjiang University (HLJU) in China. As a multilingual, he can speak both Manchu (mainly with his families) and Mandarin Chinese (other situations). Also, he can speak English. According to his own words, his Manchu pronunciation differs from the so-called ”standard pronunciation” found in many books, and his pronunciation has its own unique style. We used Praat to record in a quiet office, with the content being the first 500 sentences from ”800 Sentences of Modern Spoken Manchu” This book describes some simple conversational scenarios in Manchu and is relatively well-regarded. The 30 minutes recordings were made at 16,000 hz. After that, we transcribe the written form of our recordings by Latin letter. Manchu is typically written using the Mongolian script. However, for the convenience of our transcription, we used the Latin alphabet. This does not affect the final phonetic outcomes. Nonetheless, I believe that not using the Mongolian script is a limitation of this study. In future research, we will seek ways to incorporate it. The content of our recordings include many aspects in everyday-speech, for example, greeting, inviting, shopping, working, schooling and requesting. It almost covers all basic and essential simple Manchu words used in everyday life.

## 3.2 Mandarin Chinese and Spanish Dataset

One of the reasons we chose Spanish as the source language for transfer learning is because they share certain consonant phonemes with Manchu, especially similarities in phonology. For example, some common consonants in Spanish and Manchu are very similar in pronunciation, such as the ”r” in ”mujer” in Spanish and the ”r” in ”ara” in Manchu.

The reason why we chose Chinese as one of the source languages for transfer learning in addition to Spanish is because there are many shared vowel phonemes between them. In fact, the six basic vowels of Chinese can all be found in Manchu. In addition, there are many loanwords from Chinese in Manchu, such as ”daifu” (doctor), which is pronounced almost the same as Chinese. Therefore, by using Chinese as the source language, we can better utilize these shared phonemes and loanwords, thereby improving the performance of the Manchu speech synthesis model.

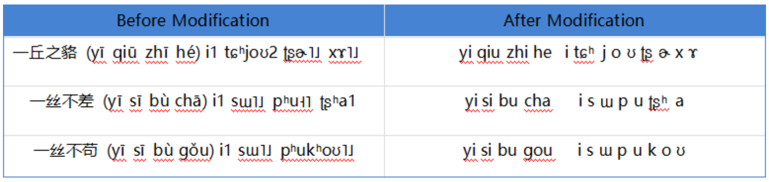
The audio datasets we used in this thesis are CSS10 datasets. In 2019, Park and Mulk created the CSS10 datasets, which included ten languages. The contribution of the datasets is that they built a single speaker datasets with aligned text for corresponding languages. It makes the users train their text to speech model more easily. (Park & Mulc, 2019) They trained the models and generated the audio from text using DCTTS and Tacotron. The results showed that the generated voice of all languages except Greek for Tacotron are excellent. In this thesis, we only use the Spanish datasets and the Mandarin Chinese datasets. However, since the Chinese dataset is only 6.5 hours long, and it is difficult for us to find a large amount of open source data from a single Chinese speaker, we use all 6.5 hours of Chinese and select a portion of the 24-hour Spanish dataset (also 6.5 hours) for training. Of course, 6.5 hours is not really considered source language, but it is still enough to generate good speech.

For the Spanish datasets, the speaker is Tux, a middle-aged man, and he read the traditional Spanish books: “Bailén”, “El 19 de Marzo y el 2 de Mayo” and “La Batalla de los Arapiles”. The lengths of each wav file are mainly from 6 to 10 seconds. There are over 15000 wav files with over 24 hours in total for Spanish. For the Mandarin Chinese datasets, the speaker is Jing Li, a middle-aged woman. Unlike the Spanish datasets, the Mandarin Chinese datasets only have 6.5 hours of recordings. The contents are the famous Chinese novels “Chao Hua Si She” and “Call to Arms”. To enhance the accuracy of the experiment, we only use 6.5 hour Spanish data and all of the Mandarin Chinese data for the purpose of transfer learning. However, in the datasets, the Mandarin Chinese recordings are of female speakers, while the Manchu and Spanish recordings are of male speakers. This would be a limitation of our research.

## 3.3 Pronunciation Dictionary

Pronunciation dictionaries are one of the necessary components of automatic speech recognition and speech synthesis models. (Nikulásdóttir, Gunason, & Rögnvaldsson, 2018) Their main purpose is to function as a look-up dictionary for the system, mapping between the grapheme and phoneme representations of an entry. Schultz in 2014 proposed that the large number of languages in the world, more than 7,100, and the need to support multiple input and output languages is a huge challenge for the speech and language community. Especially in the rapid development and deployment of speech processing systems in languages that are not yet supported. The main bottlenecks include the scarcity of speech and text data, the lack of language specifications, and the gap between technical and language expertise. (Schultz & Schlippe, 2014) In our study, we used the Spanish language pronunciation dictionary for training, but also modified the Mandarin Chinese language pronunciation dictionary for the purpose of adapting the training model. Considering that we only need to transfer learning from Mandarin Chinese to Manchu, it is not necessary to involve everything about Chinese pronunciation, such as tones. We deleted all of the tones signs of Mandarin Chinese and we used pinyin instead of Chinese characters (Hanzi). See figure 4 for some examples in this regard. This step is very important. Only after completing this step can we perform fine tuning more easily.

Obviously, as a dying language as we mentioned in the former part, there are only less than 1,000 Manchu users. Besides that, there are many different opinions on the so-called ”standard” of modern Manchu, and there is no universally recognized or definite one. We tried to find an open-sourced Manchu pronunciation dictionary In this study, but finally failed. However, we create a small size of pronunciation dictionary for Manchu language, which includes all of the words in the first 500 sentences from ”800 Sentences of Modern Spoken Manchu” 782 Manchu words, and each word has



| Before Modification | After Modification |
|---|---|
| 一丘之貉 (yī qiū zhī hé) i1 tɕʰjoʊ2 tʂɚ˥˩ xɤ˥˩ | yi qiu zhi he i tɕʰ j o ʊ tʂ ɚ x ɤ |
| 一丝不差 (yī sī bù chā) i1 sɯ˥˩ pʰu˧˥ tʂʰa1 | yi si bu cha i s ɯ p u tʂʰ a |
| 一丝不苟 (yī sī bù gǒu) i1 sɯ˥˩ pʰukʰoʊ˥˩ | yi si bu gou i s ɯ p u k o ʊ |

Figure 4: Changes of the Chinese pronunciation dictionary

a single corresponding IPA transcription. We noted down the corresponding IPA transcription by referring to “800 Sentences of Modern Spoken Manchu”, “A Research on Spoken Manchu” and Qi Xiaoxu’s ground-truth voice. The Manchu words are presented as Latin letters, instead of Mongolian letters as Manchu are written in a formal situation. The reason for doing this is that it can make our model simpler, and Manchu speakers can fully understand the transcriptions of these Latin letters. In future study, we are willing to create a new Manchu pronunciation dictionary using Mongolian letters and covering more words.

## 3.4 Aligning Model: Montreal Forced Aligner

Forced alignment shows its significant roles in linguistic sciences, including areas such as sociolinguistics, phonetics, linguistic documentation and psycholinguistics. Forced alignment means that speech and its corresponding orthographic transcription are automatically aligned at the word and phoneme level.

The Montreal Forced Aligner (McAuliffe et al., 2017) (MFA) is a command line utility for performing forced alignment of speech datasets using Kaldi. MFA adopts the standard GMM/HMM architecture and uses monophone and triphone GMMs, as well as speaker adaptation and feature transformation during training. Monophone GMMs are first iteratively trained and used to generate a basic alignment. Triphone GMMs are then trained to take surrounding phonetic context into account, along with clustering of triphones to combat sparsity. MFA uses MFCCs as acoustic features and performs Cepstral mean and variance normalization (CMVN) and Maximum Likelihood Linear Regression (fMLLR) processing. In addition, MFA includes some upgrades to Prosodylab-Aligner, which is the former version of MFA. The upgrade applications are handling of unknown words and detection of transcription bias. McGill University’s paper evaluated the Montreal Forced Aligner across three aspects: its performance compared to manual annotation, the impact of its architecture (acoustic model and speaker adaptation) on performance, and its trainability. They conducted experiments using two datasets and compared results with FAVE aligner and Prosodylab-Aligner. MFA showed good performance on both datasets and boundary types, despite some moderate to large alignment errors. The triphone acoustic model provided MFA with an advantage over Prosodylab-Aligner. MFA was updated to version 3.1 on June 4, with plans for retraining with new phone groups and rule features, as well as adding new languages such as Japanese, Arabic, and Tamil, and improving alignment methodology. In this thesis, the MFA version we used is 2.2.17.

## 3.5 Synthesis Model: Fast Speech 2

We applied Fast Speech 2 as our synthesis model in this experiment. Fast Speech 2 (FS2) was designed by Ren and his colleges from Zhejiang University and Microsoft. In the literatire review part, we have already introduced this model. In this section, we will detail the specific steps we followed to implement the FastSpeech 2 model.

The FastSpeech 2 implementation we used in this study comes from the FastSpeech 2 repository on GitHub[1]. The repository is maintained by ming024 and provides detailed implementation code and documentation. We followed the instructions in the repository to train and fine-tune the model. The specific steps include downloading the required datasets, modifying the necessary configuration files, running the preprocessing scripts and train the Spanish and Mandarin Chinese models. We then fine-tuned the Manchu model starting from the 100k checkpoint of the Spanish model. The synthesized audio was generated using specific configurations. Detailed steps and configuration files are available on our GitHub page, ensuring reproducibility and facilitating the training of custom models.

## 3.6 Evaluation

In our study, we conducted a Mean Opinion Score (MOS) test to evaluate the accuracy and naturalness of speech generation under four different conditions: ”Cn in dic,” ”Cn not in dic,” ”Es in dic,” and ”Es not in dic.” These conditions represent speech generated by transfer learning using Chinese (Cn) or Spanish (Es) datasets, where ”in dic” indicates that all words are included in the pronunciation dictionary, and ”not in dic” indicates that some words are not included in the pronunciation dictionary. Each small group has 4 audios, a total of 16 audios.

We created a questionnaire with 32 questions and 16 audios in random order. Each audio has two ratings: one is the accuracy rating, that is, whether the generated speech is correct in pronunciation, and the second is the naturalness rating, that is, whether the generated speech is natural and human-like. There are 5 options, 1-5 points, 5 points represent excellent, 1 point represents very poor

We selected a total of 8 people from the Manchu Language Research Institute of Heilongjiang University, which is also the largest research center for Manchu literature and linguistics in China, to participate in our questionnaire. They are all graduate students or teachers who are very familiar with Manchu, and some are native speakers of Manchu. The number of participants is relatively small, but as Manchu is an endangered language, it is not easy to find people who know Manchu. Due to various reasons, only 8 people were able to participate in the evaluation in this experiment, which is also a limitation of our experiment. This may lead to too much randomness in the results, thus affecting the accuracy of the experiment.

The collected ratings were then averaged for each condition to provide a clear comparison of performance. The analysis revealed nuanced insights into the effectiveness of using Chinese and Spanish datasets for transfer learning in Manchu speech synthesis, particularly highlighting the importance of a complete pronunciation dictionary. These findings are critical for guiding future research and development efforts in the field of speech synthesis for low-resource languages.

## 3.7 Ethical considerations

In this research, we aimed to develop a speech synthesis model for the endangered Manchu language using transfer learning techniques. To achieve this, we went to Heilongjiang University, China to find the native Manchu speaker Qi Xiaoxu. With the cooperation of both parties, we recorded the

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

conversation. Qi Xiaoxu agreed to the whole process and we also informed him of the purpose of the recording. Qi Xiaoxu fully agreed and was very interested in our project and was willing to provide all kinds of help.

In addition to Manchu recordings, we also used Chinese and Spanish data from the public dataset CSS10. These datasets contain rich speech data, which helps us conduct preliminary model training and transfer learning. The CSS10 dataset is a multilingual speech synthesis dataset that has been widely used in research in the field of speech synthesis. The datasets is licensed under [2]

Since Qi Xiaoxu is the only provider of Manchu data, the amount of data is relatively limited, which brings certain challenges to model training. However, the CSS10 dataset provides enough Chinese and Spanish speech data, allowing us to apply these resource-rich language data to the development of Manchu speech synthesis models through transfer learning methods.

In the process of using the data, we strictly follow ethical standards. First, Qi Xiaoxu was fully informed of the recording process and purpose and voluntarily participated in the recording. Second, the CSS10 dataset is public, and all recordings are collected with informed consent to ensure the legality and compliance of the data.

In addition, considering the possible bias issues in the data, we explicitly point out that the bias in the dataset may have an impact on the training and evaluation of the model, and discuss the possible impact of these biases and their mitigation measures in detail in the research report. We used subjective evaluation indicators, specifically the Mean Opinion Score (MOS) test, to assess the performance of the speech synthesis model. Although subjective evaluation methods involving human participants were used, we took measures to ensure the process was fair and unbiased, thus addressing related ethical issues. Subjective metrics, the MOS test, were employed for evaluation, which are relevant to the field.. Therefore, subjective evaluation methods involving human participants have not been used and are mostly not significant to use in the field of speech recognition. Therefore, there are no concerns regarding the ethics of involving human participants or any other issues that do not align with the ethics of the faculty. The MOS scores are all filled out by teachers and graduate students from the Manchu Language Institute of Heilongjiang University in China. Their evaluation data is objective and effective, and many of them are native Manchu speakers. They all strictly follow the requirements and the scores are anonymous.

In addition, we also considered the potential impact of this technology on the Manchu community. We asked the Manchu community about their views on this text-to-speech (TTS) model and understood their concerns and expectations. We recognize that although the TTS model has important potential in protecting and spreading the Manchu language, it may cause negative impacts on the language due to inaccurate pronunciation. In response to this, we communicated with the community and received understanding and support, especially the support of the Manchu Language Research Center of Heilongjiang University. We reached a consensus that this project can make a positive contribution to the preservation of the endangered Manchu language and has significant positive value.

As when it comes to the replicability of the research, the code is available via GitHub[3]. All steps and details on how to reproduce the experiments to be described in the thesis can be found under section 4. The dataset is publicly available to download and use. The outcomes should be more or less similar, but they may not be exactly the same due to certain elements that introduce randomness

[2] Information about the CSS10: `https://github.com/Kyubyong/css10https://github.com/Kyubyong/css10`

[3] `https://github.com/dingshenghuan3161581896/Manchu_synthesis`

in the trained models. The hardware used may also impact the performance of the models since the experiments have been conducted on the University of Groningen's high-performance cluster, Hábrók.

This concludes the methodology section which explains the methods employed during this research. In the next section, the experimental setup will be presented which will include more low-level details about the dataset used and the parameters of the models.

# 4 Experiments

We will make a systematic introduction about our experiments in this section. In order for the research to be fully reproducible, please check out the GitHub link at the bottom of this page and read the readme to verify and reproduce our experiment. The specific parameters and steps are presented there.

We will start by examining my datasets that are used in our experiment4.1. Then, we will show the detailed infomation to our pronunciation dictionaries4.2. Next, we will explain our training process.4.3. The last part is the synthesis and Evaluation.4.4

## 4.1 Data Collections

As we mentioned in the previews parts, we have prepared a 30-minutes Manchu audio, which was recorded by us and the native Manchu speaker Qi Xiaoxu and 6.5 hours of Spanish and Mandarin Chinese data from CSS10 datasets respectively. We put those three datasets into four folders, they are: ”cn” (6.5 hours of the Mandarin Chinese data), ”es” (6.5 hours of the Spanish data), ”cnm” (0.5 hour of the Manchu data) and ”esm” (0.5 hour of the Manchu data). To avoid confusion during training, we copied the Manchu audio and presented it in two folders according to the different languages used for transfer learning. Then we use a Python code to label the transcription files corresponding to the recordings in three languages to each recording. In this way, each of our recordings will have a corresponding ”.lab” file, and the content of the file is the transcription of this recording.

Prior to commencing training, we dedicated meticulous attention to fine-tuning the pronunciation dictionaries for both languages. This involved the harmonization of phonemes that shared similar sounds but were represented by different symbols. We also extended this harmonization process to the creation of the pronunciation dictionary for the Manchu language, ensuring consistency and accuracy in linguistic representations.

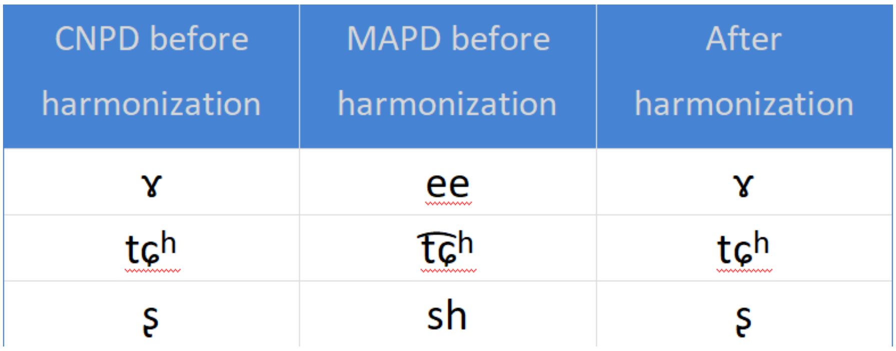
## 4.2 Pronunciation dictionaries

We created our Manchu pronunciation dictionary. We can not find any open-sourced Manchu pronunciation dictionary, and then we decided to create a new one based on our recordings. Our Manchu pronunciation dictionary involves 782 words, which almost covers all of the words appear in the 30 minutes of Manchu recordings. Each words have their corresponding transcriptions on IPA.

Prior to commencing training, we dedicated meticulous attention to fine-tuning the pronunciation dictionaries for both Spanish and Mandarin Chinese. This involved the harmonization of phonemes that shared similar sounds but were represented by different symbols. We also extended this harmonization process to the creation of the pronunciation dictionary for the Manchu language, ensuring consistency and accuracy in linguistic representations. Some examples of harmonization are shown in the figure5.

## 4.3 Training Data

Subsequently, we utilized MFA (Montreal Forced Aligner) to facilitate alignment, a crucial step before training new acoustic models for the three languages under scrutiny. This alignment process



| CNPD before harmonization | MAPD before harmonization | After harmonization |
|---|---|---|
| ɤ | ee | ɤ |
| tɕʰ | t͡ɕʰ | tɕʰ |
| ʂ | sh | ʂ |

Figure 5: Examples of the harmonization of our pronunciation dictionary (CNPD means Chinese pronunciation dictionary and MAPD means Manchu pronunciation dictionary)

ensured that the speech data and linguistic annotations matched correctly, providing a solid foundation for our subsequent training efforts. Following alignment, we proceeded with data preprocessing to optimize data quality and prepare it for learning algorithms. During the trial training phase, we found that just 100k steps sufficed to produce relatively high-quality speech in both Spanish and Mandarin Chinese. This efficiency highlighted the effectiveness of the transfer learning approach adopted in our study. For more details, please check our GitHub page. All parameters are set in the ”config” folder, and all steps required to run and verify the model can be found in the README. [4]

Encouraged by this progress, we proceeded with aligning Manchu using MFA, then resumed training from the 100k steps checkpoint for each language. We meticulously recorded regular checkpoints every 2k steps, ensuring a thorough and effective training process. Upon completion of training, thorough testing ensued, yielding promising results. We found that fine-tuning for 100k steps (culminating in a total of 200k steps) yielded commendable outcomes, indicative of the efficacy of the transfer learning methodology in the realm of Manchu synthesis. With the training phase concluded, our focus shifted to the generation of Manchu recordings based on the 200k checkpoints. We carefully crafted eight recordings, paying close attention to every detail, with durations ranging from 2 to 6 seconds. Among these recordings, four contained vocabulary sourced from the pronunciation dictionary, while the remaining four featured vocabulary not found in the dictionary.

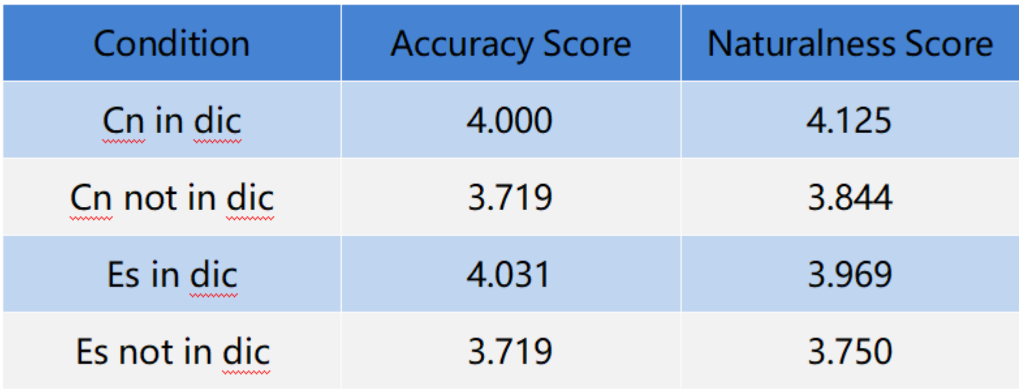
## 4.4 Synthesis and Evaluation

Finally, these recordings were provided to native Manchu speakers, as well as master’s students and faculty members at the Manchu Research Center at Heilongjiang University, for MOS ratings. As for the text to generate audio, we continue to choose ”800 sentences of modern Manchu”, which is also the material we used when recording. We used the first 500 sentences when recording, and 8 sentences from the last 300 sentences for synthesized speech. Among the 8 selected sentences, all the words in 4 sentences appear in the Manchu pronunciation dictionary we created, which covers 782 words, and half of the words in the other 4 sentences are not included in the pronunciation dictionary.

[4] https://github.com/dingshenghuan3161581896/Manchu_synthesis

# 5 Results

In this MOS (Mean Opinion Score) test, we evaluated the accuracy and naturalness of speech generation under four different conditions: Cn in dic, Cn not in dic, Es in dic, and Es not in dic. These conditions correspond to speech generated by transfer learning using Chinese or Spanish, respectively, where ”in dic” means that all words are included in the pronunciation dictionary, and ”not in dic” means that some words are not in the pronunciation dictionary. The table shows the average MOS scores.



| Condition | Accuracy Score | Naturalness Score |
|---|---|---|
| Cn in dic | 4.000 | 4.125 |
| Cn not in dic | 3.719 | 3.844 |
| Es in dic | 4.031 | 3.969 |
| Es not in dic | 3.719 | 3.750 |

Figure 6: Average MOS Scores

In terms of accuracy score, the average score of ”Cn in dic” is 4.0, slightly lower than the 4.03125 of ”Es in dic”, which shows that the generation effect of transfer learning using Spanish is slightly better when included in the pronunciation dictionary. When not included in the pronunciation dictionary, the average scores of ”Cn not in dic” and ”Es not in dic” are both 3.71875, indicating that the two languages perform equally under this condition. It is obvious that when there are words that do not appear in the Manchu pronunciation dictionary we created, the phonetic accuracy is significantly lower than when all words are included in the dictionary, regardless of which language is used for transfer learning. In terms of naturalness score, ”Cn in dic na” has the highest average score of 4.125, which shows that the speech generated by transfer learning using Chinese is more natural when all words are included in the pronunciation dictionary. The average score of ”Es in dic na” is 3.96875, slightly lower than ”Cn in dic na”. The average scores of ”Cn not in dic na” and ”Es not in dic na” are 3.84375 and 3.75 respectively, indicating that the generation effect of using Chinese is slightly better than that of Spanish when it is not included in the pronunciation dictionary. Like the accuracy results, when there are words that do not appear in the Manchu pronunciation dictionary we produced, the naturalness of the speech is significantly lower than when all words are included in our pronunciation dictionary, regardless of which language is used for transfer learning.

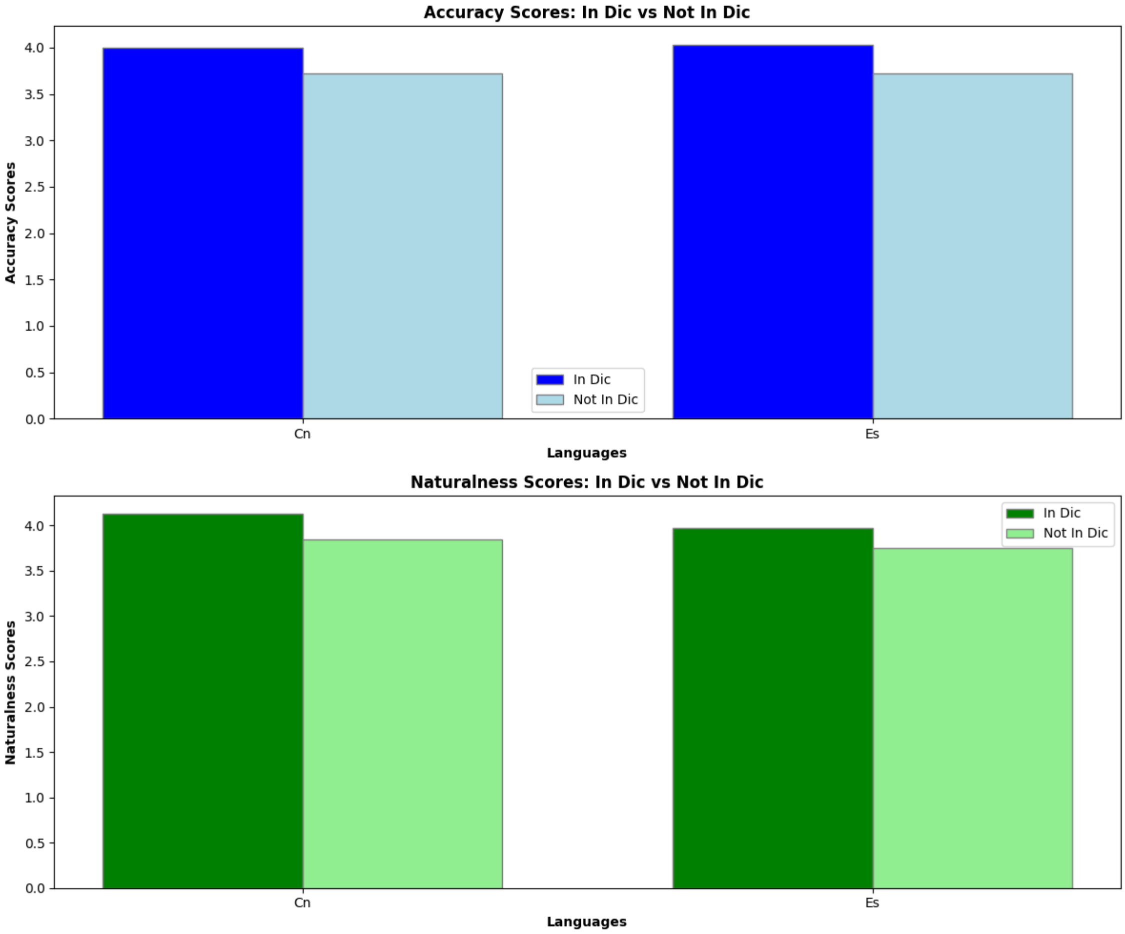


Figure 7: Comparison between in dic and not in dic

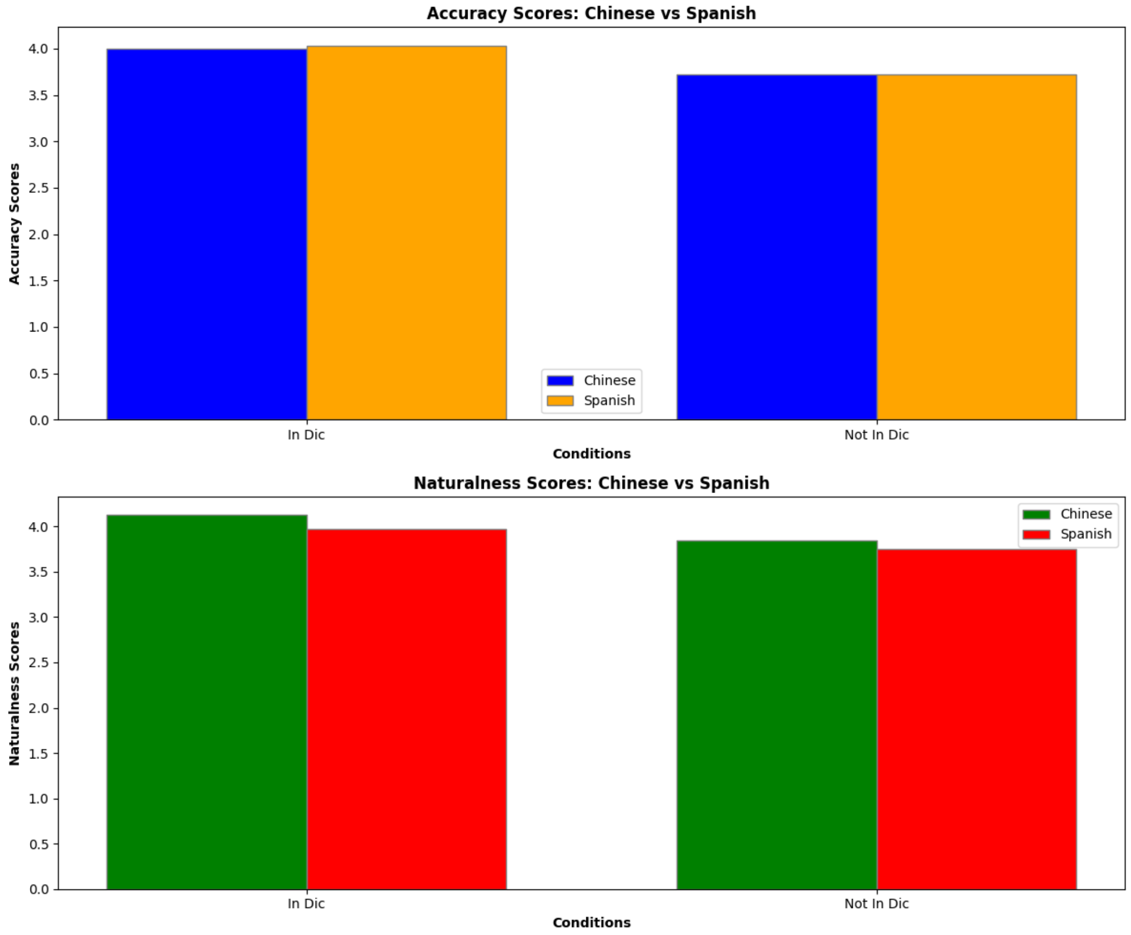


Figure 8: Comparison between cn and es

Based on the above analysis, although the two languages perform similarly under some conditions, Chinese performs particularly well when the vocabulary is fully included in the pronunciation dictionary, especially in terms of speech naturalness.

# 6 Discussion

Through transfer learning, we have generated a model that can generate Manchu speech well when all words are covered by the pronunciation dictionary. However, when the pronunciation dictionary is not covered, the performance of the model drops significantly, but it can still generate some pronunciations.

## 6.1 Validation of the Hypothesis

The primary hypothesis of this study posited that utilizing the training approach with Mandarin Chinese would yield better outcomes in phoneme synthesis for Manchu than using Spanish. Our MOS test results provide significant insights into the validation of this hypothesis.

While accuracy between Chinese and Spanish exhibits marginal differences, Chinese demonstrates superior naturalness compared to Spanish. In both ”in dic” and ”not in dic” groups, Chinese outperforms Spanish by 0.13 and nearly 0.1 points, respectively. This highlights the nuanced performance dynamics between the two languages, where naturalness emerges as a significant factor. These findings underscore the complexity of phoneme synthesis and the importance of considering not only accuracy but also naturalness in language processing tasks. Further investigations into the underlying factors influencing naturalness are warranted to optimize phoneme synthesis techniques for diverse linguistic contexts.

While our hypothesis anticipated Mandarin Chinese to yield superior outcomes in phoneme synthesis for Manchu compared to Spanish, our findings revealed a nuanced picture. Although Spanish demonstrated a slight edge in accuracy, Chinese showcased superior naturalness. Despite this, the hypothesis was not fully realized, indicating the complex interplay between training approaches and language characteristics. Further exploration into optimizing training methodologies and considering linguistic nuances is warranted for enhancing Manchu phoneme synthesis.

## 6.2 Validation of the First Subquestion

We raised three subquestions. The first one was “Is it possible to use just a few Manchu data to generate a decent text-to-speech model?” Our study indicates that it is indeed possible to generate a decent text-to-speech model with a limited amount of Manchu data. The MOS scores suggest that even with a small dataset, leveraging transfer learning from a well-represented language like Mandarin Chinese can produce relatively high-quality speech synthesis. The results demonstrate that the inclusion of all words in the pronunciation dictionary significantly enhances both accuracy and naturalness, emphasizing the importance of comprehensive linguistic resources. This finding is particularly valuable for low-resource languages, showing that effective TTS models can be developed without extensive datasets. It can be seen that modern Manchu has been greatly influenced by Chinese, especially in pronunciation. In addition, the Manchu recording provider Qi Xiaoxu’s business office is a native Mandarin speaker, which may have influenced his Manchu pronunciation.

## 6.3 Validation of the Second Subquestion

The second subquestion is that ”the recordings in Mandarin Chinese are in a female voice, while those in Spanish are in a male voice. How does this difference affect the final synthesized speech?”

While our experiment cannot completely validate the impact of voice gender differences, the results suggest that the gender of the voice used for transfer learning does not significantly affect the final synthesized speech quality. Despite using female voice recordings for Mandarin Chinese and male voice recordings for Manchu fine-tuning, the performance of the Mandarin Chinese-to-Manchu model was generally better than the Spanish-to-Manchu model, even though both Spanish and Manchu used male voices. This indicates that under the conditions of 100k source language checkpoints and 100k fine-tuning iterations, the speaker's gender did not play a decisive role in the final audio quality. Interestingly, in the naturalness tests, the Mandarin Chinese transfer to Manchu model performed better, further suggesting that voice gender was not a determining factor in this scenario.

## 6.4 Limitations

Our study provides valuable insights into the effectiveness of using transfer learning for Manchu text-to-speech synthesis. However, several limitations must be acknowledged:

First, to generate better Manchu speech, we should use more source language data for training. In this experiment, we only used 6.5 hours of Chinese and Spanish recordings. This limited training time may not be enough to fully exploit the potential of transfer learning. Using more source language data allows the model to learn the phonemes and speech features of the language in a wider range, thereby improving the quality of synthesized speech. By increasing the amount and diversity of training data, the model can better capture the complexity of the language and generate more natural and accurate Manchu speech. In addition, extending the training time and combining more language data sources can also improve the stability and robustness of the model, making it more reliable when processing various speech inputs.

Second, the Manchu pronunciation dictionary covers too few words, including only 782 words. This coverage is not enough to cover all the vocabulary in the recordings, which severely limits the ability to generate high-quality Manchu speech synthesis. Expanding the coverage of the dictionary to include more words is crucial to improving the accuracy and naturalness of the model. A more comprehensive pronunciation dictionary can help the model better handle unknown vocabulary, thereby generating more consistent and reliable speech output. Especially when dealing with new or rare vocabulary, having a widely covered dictionary can significantly improve the performance of the model. In addition, incorporating more detailed phonetic annotations and pronunciation rules can further enhance the model's performance in speech synthesis.

Finally, since Manchu speakers are very scarce, it is difficult to find enough native speakers to participate in our MOS evaluation. Therefore, only 8 participants evaluated the synthesized speech. This small sample size may introduce a certain degree of randomness and subjectivity, which may affect the overall reliability and objectivity of the results. Future research should increase the number of Manchu native speakers participating in the evaluation as much as possible to ensure that the evaluation results are more comprehensive and objective. Larger-scale evaluations can not only provide more precise feedback, but also help identify differences in the performance of the model in different language contexts. In addition, the use of multiple evaluation methods, such as the combination of automated evaluation tools and subjective listening tests, can also improve the comprehensiveness and accuracy of the evaluation, thereby providing a more solid foundation for further optimization of the model.

In summary, the research question has been addressed and my initial hypothesis has been validated. As the research objectives have been met and the study's contributions have been established, the subsequent section will serve as the concluding chapter, encapsulating the key findings and their implications.

# 7 Conclusion

A brief summary of the contributions made with this research will be presented alongside future plans and possibilities, ending with a subsection on the impact and relevance of my thesis in the field of low-resource TTS and in the Manchu language community.

## 7.1 Summary of the Main Contributions

This study aims to explore the effectiveness of using transfer learning for Manchu speech synthesis and verify the hypothesis by comparing the training methods for Spanish and Mandarin. Through the analysis of experimental results, we verified our research hypothesis that 30 minutes of Manchu and 6.5 hours of Spanish or Mandarin as training data can generate a relatively decent Manchu speech synthesis system. Due to the scarcity and extinction of Manchu recordings, this model is also the first Manchu speech synthesis model in the world. In addition, we can achieve better results in phoneme synthesis in Manchu by training with Mandarin. Our study found that despite the limited amount of Manchu data, the transfer learning method using Mandarin can produce relatively high-quality speech synthesis. In addition, our study also found that comprehensive coverage of the pronunciation dictionary is crucial to obtain highly accurate and natural synthesized speech. Our study fills the gap in the field of Manchu speech synthesis and provides an effective solution for speech synthesis of low-resource languages. Future research can further explore more transfer learning methods, expand the coverage of the pronunciation dictionary, and increase the number of native Manchu speakers participating in the evaluation to further improve the performance and robustness of the model.

## 7.2 Future Work

Future work can be divided into several stages. First, we will expand the scale of recording, improve the quality of recording equipment, and increase the scope of recording content, from simple daily spoken language to various fields, such as science, technology, literature, and art. Improving the quality of recording and covering more fields will help capture the diversity and nuances of the language, thereby improving the performance of the speech synthesis model. Second, we plan to try to use other languages for transfer learning, such as Mongolian and Korean, which are close to Manchu. The phonetic features of these languages may have more similarities with Manchu, and transfer learning may lead to better synthesis results.

In addition, based on the results of this experiment (Chinese performs better), we will continue to use more Chinese male voice datasets for transfer training to further improve the performance of the model, because the provider of my Manchu recording is male, while the Chinese dataset in this experiment is female voice. Adding different genders and more diverse voice data will help improve the universality and accuracy of the model. Furthermore, we will work on expanding and optimizing the Manchu pronunciation dictionary and adding more vocabulary to improve the accuracy and naturalness of the model. A more comprehensive dictionary will help the model better handle unseen vocabulary, thereby improving the overall synthesis quality.

We also plan to explore more data enhancement techniques to improve the performance of the model under low-resource conditions. For example, through techniques such as audio data transformation, noise addition, and speech speed change, more training data can be generated to improve the robustness and performance of the model. In addition, we will explore more advanced speech synthesis technologies, such as the latest models based on neural networks, to further improve the quality and naturalness of synthesized speech.

Finally, expand the scope of participants in the MOS test, especially increase the number of native Manchu speakers, to ensure the objectivity and reliability of the evaluation results. More native speakers participating in the evaluation can provide more accurate and detailed feedback, which will help us better optimize the model. Through these measures, we hope to further improve the quality and practicality of Manchu speech synthesis and contribute to the protection and promotion of this low-resource language.

## 7.3 Impact & Relevance

This study is of great significance in the preservation and revitalization of Manchu. As the first research project focused on Manchu speech synthesis, it opens up new avenues for the digital preservation and dissemination of this endangered language. By applying transfer learning techniques, this study shows how to use resources from other languages to make up for the lack of Manchu corpus, thereby developing a system that can generate high-quality Manchu speech. This achievement not only provides valuable data and methods for the academic community, but also brings practical application prospects to the Manchu community. For example, the Manchu speech synthesis system can be used in educational software to help learners master Manchu pronunciation more easily; at the same time, it can also be applied to cultural preservation projects to produce Manchu audiobooks and digital archives, making it easier for Manchu to be preserved and disseminated.

In addition, this study emphasizes the importance of comprehensive pronunciation dictionaries and diverse corpora in speech synthesis, which provides guidance for future research and practice. In the process of continuing to expand and optimize the Manchu speech synthesis system, these insights will help improve the accuracy and naturalness of the system and further promote the protection and inheritance of Manchu. In short, this study not only fills the gap in the field of Manchu speech synthesis, but also provides practical technical means for the protection and revitalization of Manchu. It provides a solid foundation for future research and demonstrates the potential of technology in language conservation.

# References

Association, I. P. (1999). *Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet.* Cambridge University Press. (p. 10)

Azizah, K., Adriani, M., & Jatmiko, W. (2020). Hierarchical transfer learning for multilingual, multi-speaker, and style transfer dnn-based tts on low-resource languages. *IEEE Access*, *8*, 179798–179812.

Do, P., et al. (2022). Text-to-speech for under-resourced languages: Phoneme mapping and source language selection in transfer learning. In *Proceedings of the 1st annual meeting of the elra/isca special interest group on under-resourced languages.*

Fahmy, F. K., Khalil, M. I., & Abbas, H. M. (2020). A transfer learning end-to-end arabic text-to-speech (tts) deep architecture. In *Iapr workshop on artificial neural networks in pattern recognition.* Cham.

Hualde, J. I. (2005). *The sounds of spanish with audio cd.* Cambridge University Press.

Ji, Y., Zhao, Z., & Bai, L. (1989). *800 sentences of modern manchu.* Beijing: Central University for Nationalities Press.

Jia, Y., & et al. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in neural information processing systems 31.*

Jongman, A., et al. (2006). *Perception and production of mandarin chinese tones* (Tech. Rep.).

Kim, J., et al. (2020). Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In *Advances in neural information processing systems 33* (pp. 8067–8077).

Kim, M., et al. (2022). Transfer learning framework for low-resource text-to-speech using a large-scale unlabeled speech corpus. *arXiv preprint arXiv:2203.15447*.

Liu, Z. (2023). Comparative analysis of transfer learning in deep learning text-to-speech models on a few-shot, low-resource, customized dataset. *arXiv preprint arXiv:2310.04982*.

McAuliffe, M., et al. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech* (Vol. 2017).

Mu, Z., Yang, X., & Dong, Y. (2021). Review of end-to-end speech synthesis technology based on deep learning. *arXiv preprint arXiv:2104.09995*.

Nikulásdóttir, A. B., Gunason, J., & Rögnvaldsson, E. (2018). An icelandic pronunciation dictionary for tts. In *2018 ieee spoken language technology workshop (slt).*

Odinye, I. S. (2022). Phonology of mandarin chinese: A comparison of pinyin and ipa. *Journal.*

Park, K., & Mulc, T. (2019). Css10: A collection of single speaker speech datasets for 10 languages. *arXiv preprint arXiv:1903.11269*.

Penny, R. J. (2002). *A history of the spanish language*. Cambridge University Press.

Ren, Y., et al. (2019). Fastspeech: Fast, robust and controllable text to speech. In *Advances in neural information processing systems 32.*

Ren, Y., et al. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Rhoads, E. J. M. (2015). *Manchus and han: Ethnic relations and political power in late qing and early republican china, 1861-1928*. University of Washington Press.

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

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 160.

Schultz, T., & Schlippe, T. (2014). Globalphone: Pronunciation dictionaries in 20 languages. In *Lrec.*

Shen, J., et al. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 ieee international conference on acoustics, speech and signal processing (icassp).*

Tan, C., et al. (2018). A survey on deep transfer learning. In *Artificial neural networks and machine learning–icann 2018: 27th international conference on artificial neural networks, rhodes, greece, october 4-7, 2018, proceedings, part iii* (Vol. 27). Springer International Publishing.

Tan, X., et al. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561.*

Tu, T., & et al. (2019). End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. *arXiv preprint arXiv:1904.06508.*

Wang, Q., & Andrews, J. F. (2021). Chinese pinyin. *American Annals of the Deaf*, *166*(4), 446–461. Retrieved from `https://www.jstor.org/stable/pdf/27113321.pdf?casa_token=qKQ8ekZipoAAAAAA:Vw_4wRwcyIXf19qBKfgzd90p8iX6HD_VZYnTFW6U5yoagKJmGr5AzzkWtcuTkcwqp52mrzFyic7rzDHW-cOtOj_siAWSYKEv6chzsvBNiEAqOo11sHEYYA`

Zhang, S., & Peng, S. (2021). Research on the current situation and development trend of manchu language inheritance. *Forum on Chinese Culture*, *21.* (In Chinese)

Zhao, A. (2018). Research report on the current situation of manchu, hezhe, xibe languages and cultures. *Northwest China Ethnic Studies*(3), 9. Retrieved from `https://www.cssn.cn/mzx/xksy_lswh/202208/W020220803373684051352.pdf` doi: CNKI:SUN:SAGA.0.2018-03-021

Zhao, J. (1987). Phonemic analysis of tailai manchu. *Manchu Studies*, *1*, 16. doi: CNKI:SUN:MYYJ.0.1987-01-002

Zhu, Z., Zhang, H., Zhao, J., Guo, X., Zhang, Z., Ding, Y., & Xiong, T. (2018). Using toponyms to analyze the endangered manchu language in northeast china. *Sustainability*, *10*(2), 563.

Zhuang, F., et al. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43–76.