university of groningen

campus fryslân

Exploring the Impact of Prosodic Styles in Datasets on Mandarin Speech Synthesis Using BERT-VITS

Yanhua Liao

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Exploring the Impact of Prosodic Styles in Datasets on Mandarin Speech Synthesis Using BERT-VITS

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 second reader being

Yanhua Liao (S5096413)

June 11, 2024

Acknowledgements

First and foremost, I would like to express my gratitude to all of my teachers. My thesis supervisor, Phat.Do, has taught me a great deal about voice technology during my year of study at University of Groningen; he is an exceptional educator. It was the accumulation of knowledge from his daily lectures that enabled me to successfully complete the experiments and writing of my thesis. Matt, head of the voice technology project, has always encouraged my various "wild ideas" from the very beginning, and yes, my graduation thesis research evolved from countless such "wild ideas." I acknowledge the Center for Information Technology of University of Groningen for technical support.

Secondly, I acknowledge my classmates. I am not a naturally gifted student, and without their daily assistance in resolving my confusions about voice technology and their emotional support during my less-than-ideal exam results, I believe it would have been very difficult for me to complete my thesis. Of course, I must also acknowledge Mr. Ma for providing his speech data, which plays a crucial part in my research. And I acknowledge all my friends and family, whose participation made it possible for me to obtain the relevant data to support my thesis research and hypotheses.

I would also like to acknowledge my husband, Mr. Xie. Thank you for supporting me in returning to school at the age of 33. During this year at the University of Groningen, he has provided me with a lot of help both emotionally and materially.

Lastly, I acknowledge myself. I thank myself for not giving up on dreaming and for healing on countless nights when I felt overwhelmed and burned out. Over the past year, I have questioned myself almost every day, often asking, "Why did I come all this way to prove I am a fool?" But when I finished my research, I told myself, "I am grateful you made that choice to come."

Abstract

As a major channel for information dissemination, broadcasting has a history spanning over a century and has always played an irreplaceable role. Over the past decade, text-to-speech (TTS) technology based on deep learning has gradually emerged and achieved numerous successes in the Mandarin domain. Many radio stations have introduced TTS models into the production of news broadcasting programs to improve production efficiency. However, the application of these technologies in more relaxed and natural entertainment programs remains relatively limited. This paper is based on the BERT-VITS model, a state-of-the-art text-to-speech synthesis system. The BERT-VITS model combines the capabilities of BERT (Bidirectional Encoder Representations from Transformers) for natural language understanding with VITS (Variational Inference Text-to-Speech) for high-quality speech synthesis. We first use a large open-source dataset with a prosodic style biased towards news broadcasting to train the TTS model. We then fine-tune the model on a smaller dataset with a more relaxed and natural prosodic style. The purpose of the experiment is to investigate the impact of basic training and fine-tuning with datasets of different prosodic styles on the model's output audio prosody after the initial training. We aim to create more diverse prosodic styles for broadcasting programs by combining the BERT-VITS model with datasets of different prosodic styles, thereby enriching the types of artificial intelligence (AI) broadcasting. Our experimental results demonstrate that the system can accurately and dynamically adjust the target voice timbre and prosodic style to match the reference speech. Even with limited data training, it can synthesize speech that matches the target speaker's prosody and style. The generated speech has good naturalness and inference speed, making it suitable for broadcasting content.

Key Words: Transfer learning, Mandarin broadcasting, BERT, VITS, prosodic styles

Contents

1 Introduction

Radio plays an important role in our daily life. It is not only one of the important channels for us to obtain information, but also a way of entertainment. The relative independence of hearing and vision attention allows people to listen to the radio while performing other activities, using it as a "background media". This also makes radio media more widely used than video media in application scenarios, for instance, when people are driving, they can listen to the music or radio but cannot pay attention to video. With the development of text-to-speech(TTS) technology based on deep learning, the broadcast industry has also been profoundly affected by artificial intelligence (AI) (Ying, 2018). In China, radio and television stations in about ten cities have established their own Artificial Intelligence Generated Content (AIGC) work studios. The use of AI not only allows broadcasters to have unlimited possibilities in content innovation, but also improves daily work efficiency.

The utilization of AI in broadcasting program production is one of the effective approaches to address the significant human resources required in traditional broadcasting. Intelligent broadcasting program production involves converting the text manuscripts of radio programs into voice content corresponding to the host's voice using speech synthesis technology, and then broadcasting it, thus achieving the automation of the production process. Compared to human-hosted broadcast programs, utilizing TTS for program production not only eliminates the announcer's need to prepare and familiarize with the manuscript, saving time and effort, but also ensures timely completion of the broadcast. Since the TTS model maintains a consistent and stable state, unlike humans who may be affected by emotional fluctuations or fatigue, intelligent broadcasting can ensure fluent language and error-free expression, enabling round-the-clock broadcasting.

At present, TTS systems are predominantly used in Mandarin broadcasting for news programs. These programs typically cover current affairs and political news. Given the political nature and seriousness of the main news items, along with the audience's expectation of accuracy, standardization, rigor, conciseness, truthfulness, and objectivity, news broadcasts are characterized by a solemn and dignified tone (Wu T., 2023). In other words, excessive tone modification is unnecessary. For instance, (Zhao W., 2023) experimented with Tacotron for Mandarin news broadcasts. However, the penetration of TTS in radio programs with the relaxed prosody and style is limited. These programs require not only accurate pronunciation but also personalized language expression from the hosts, aiming for a relaxed, friendly, and natural tone. Therefore, they emphasize colloquial language use, unrestricted form, and a tone closer to natural speech, enhancing audience engagement and connection (X. Li, 2016). Consequently, it's challenging for synthesized voices to replace human announcers in these contexts.

The key to realizing the intelligent production of radio programs lies in speech synthesis technology. Nowadays, speech synthesis technology has developed significantly and is used in various fields such as mobile assistants and robots. The evolution of Text-to-Speech technology has transitioned from complex multi-stage processes to simplified two-stage generative modeling methods(Watts, Eje Henter, Fong, & Valentini-Botinhao, 2019). Initially, TTS systems required multiple stages. For example, statistical parametric speech synthesis (SPSS) (Zen, Tokuda, & Black, 2009) includes linguistic feature modeling, acoustic feature modeling, statistical modeling, synthesis parameter generation, and waveform synthesis. This required

users of the models to have a certain background in phonetics.

With the advancement of deep neural networks, TTS system has not only significantly improved speech quality but also reduced the complexity of the architecture. Modern TTS systems primarily adopt two-stage generation modeling methods, lowering the threshold for model usage and simplifying the process of speech synthesis. In this approach, intermediate speech representations are first generated from preprocessed text, followed by the generation of the raw waveform based on these representations. For instance, in 2017, Yuxuan Wang, et al proposed Tacotron (Wang et al., 2017): Towards End-to-End Speech Synthesis, which involved training with short audio segments, utilizing a Mel-spectrogram decoder to assist in text representation learning, and designing specialized spectrogram losses. However, the two-stage workflow, similar to Tacotron, requires sequential training or fine-tuning and depends on predefined intermediate features, which limits the potential for further performance improvements. To address these issues, in 2018, Baidu's Silicon Valley AI Lab proposed ClariNet (Ping, Peng, & Chen, 2018), a model based on WaveNet for parallel generation of raw audio waveforms. ClariNet improved synthesis speed by several thousand times, achieving more than ten times real-time performance. This was also the first truly end-to-end model in the field of speech synthesis: a single neural network that directly converts text to raw audio waveforms.

In terms of speech generation strategies, the autoregressive TTS system was once mainstream (Oord et al., 2016). However, due to its sequential generation process, it was difficult to fully utilize parallel processors, leading to the emergence of non-autoregressive methods to improve synthesis speed (Ren et al., 2019). At the same time, generative adversarial networks were extensively explored in the second-stage model of TTS, to achieve high-quality raw waveform synthesis (Saito, Takamichi, & Saruwatari, 2017).

This paper relies primarily on the BERT-VITS [1] model. The core of this model lies in its integration of the strengths of BERT[2] (Devlin, Chang, Lee, & Toutanova, 2018) pre-trained models. Such models have demonstrated excellent performance in language understanding tasks, capturing rich contextual information to make synthesized speech more contextually appropriate and enhance naturalness. Additionally, it incorporates VITS[3] (Kim, Kong, & Son, 2021), which is an end-to-end variational autoencoder model that utilizes a variational auto-regressive structure to optimize latent space modeling, resulting in high-fidelity audio with smooth and coherent content .

The BERT-VITS model adopts a neural network structure comprising multiple layers, including 6 residual blocks, each with different convolutional kernel sizes and dilation factor settings. The model contains a total of 2 attention heads, with 3 convolutional layers in each head. Additionally, BERT-VITS includes an upsampling module to enhance the resolution and quality of the speech.

To validate the influence of dataset types on model performance, we first train the model on the open-source Mandarin dataset Baker[4]. This dataset consists of recordings from a single female speaker with clear audio quality, comprising approximately 12 hours and 10,000 sentences. The speaker's voice style is characterized as intellectual and cheerful, representing a

[1]Information about the model: `https://github.com/PlayVoice/vits_chinese`

[2]Information about Bert: `https://github.com/google-research/bert`

[3]Information about VITS: `https://github.com/jaywalnut310/vits`

[4]Information about the dataset: `https://www.data-baker.com/data/index/TNtts/`

professional standard Mandarin female voice, conveying an optimistic and positive tone. This open-source recording corpus covers various domains such as news, novels, entertainment, and dialogue. The corpus design comprehensively samples the linguistic features, aiming to cover syllables, phonemes, types, tones, phonetic sequences, and prosody as comprehensively as possible within a limited corpus data volume. Subsequently, we use our self-made small size dataset as the target for transfer learning. This small dataset consists of 300 recordings from a single male speaker, recorded during the daily work of a professional broadcaster. The content mainly involves storytelling, with a conversational style and highly individualized speech characteristics. The recording environment for this corpus is a professional recording studio, with recording equipment meeting the requirements for professional broadcasting programs. Finally, we will verify our hypothesis about the model's performance through subjective listening tests and objective data analysis.

1.1 Organization

Now that a brief motivation for this research has been presented. Accordingly, the work arrangement of this paper is as follows: This paper is divided into six parts. After the introduction, a comprehensive literature review is conducted, studying the development of TTS models 2.1, the application of Bert in the field of TTS 2.1, and the development of transfer learning 2.3. The following chapter focuses on methodology 3, including the description of the establishment of the dataset, experimental strategies, parameter settings, and other specific details. Next chapter 4 presents the results, comparing the training outcomes on ordinary open-source datasets with those after fine-tuning on specially crafted datasets with a host's personal speaking style in daily conversation. In section 5, I discuss the previously-mentioned results in detail. Lastly, section 6 summarizes the thesis and presents the conclusions drawn, along with recommended future work.

1.2 Research Question and Hypothesis

In the domain of Mandarin broadcasting, the application of speech synthesis technology is predominantly focused on formal news broadcasting programs. This is because traditional news broadcasts place minimal emphasis on emotional expression, primarily prioritizing accurate pronunciation. For instance, Zhao et al. conducted experiments in Mandarin news broadcasting utilizing Tacotron (Zhao W., 2023). However, the application of such technology remains relatively limited for other types of programs, such as talk shows and variety shows. This limitation is attributed to the additional requirements of these programs, which necessitate not only accurate pronunciation but also a higher level of emotional expression, prosody, and personalized delivery—a level of performance that current model synthesis techniques struggle to replicate compared to human broadcasters.

Therefore, this study aims to explore the impact of datasets with different prosodic styles on model performance. We utilize the BERT-VITS model, pre-trained on source data characterized by clarity and fluency in broadcasting style. Subsequently, we fine-tuning the model on target data characterized by a more natural, everyday conversational tone, emphasizing a relaxed and casual style of speech. Our primary focus lies in discerning the disparities in model performance across datasets of varying styles. If successful in generating Mandarin speech that

seamlessly combines a relaxed, natural prosodic style with accurate pronunciation, it would play a crucial role in the modernization of Mandarin broadcasting institutions. Through an in-depth exploration of the BERT-VITS model application, we aim to make substantial contributions to the domain of Mandarin broadcasting content generation.

Our hypothesis is that the BERT-VITS model, when pre-trained on a standard Mandarin source dataset characterized by stable intonation, accurate pronunciation, and clear expression, and then fine-tuned on a target dataset with a more relaxed, conversational, and approachable prosodic style, can generate Mandarin speech that combines a relaxed, casual prosodic style with fundamentally accurate pronunciation. Based on a comprehensive synthesis of previous research findings in the field of speech synthesis technology, we believe this hypothesis is reasonable. These research results highlight the maturity of current techniques, with the BERT-VITS model demonstrating stability in generating extended speech, synthesizing multiple emotions, and producing smoother, more natural speech.

To validate this hypothesis, we will subjectively evaluate Mandarin speech segments generated using the pre-trained and fine-tuned BERT-VITS models, with a particular focus on whether their overall style effectively conveys a relaxed, conversational tone. By logically deducing our hypothesis from existing literature, our objective is to conduct a targeted, evidence-based exploration to assess the effectiveness of the BERT-VITS model in generating Mandarin broadcasting content characterized by a conversational rather than a news broadcasting style.

2 Literature Review

Throughout the research on speech synthesis technology, which has a history of more than 200 years, BERT also has many applications in the field of text-to-speech(TTS). For our purposes in this section we briefly review the above.

2.1 TTS

Speech synthesis technology has undergone roughly five stages: Articulatory Synthesis, formant synthesis, concatenative synthesis, statistical parametric synthesis, and neural speech synthesis (Taylor, 2009). The first three of these stages represent early traditional speech synthesis techniques. During this period, attempts were made to utilize physical models to simulate the process of human speech production. Articulation synthesis, for instance, aimed to replicate the movement of physiological components such as vocal cords and throat to generate speech signals. Formant synthesis, on the other hand, is often called synthesis-by-rule. Most formant synthesis techniques do in fact use rules of the traditional form, but data driven techniques have also been used. It adopts a modular, model-based, acoustic-phonetic approach to the synthesis problem. For instance, the Voder developed by Homer Dudly (also the developer of the Vocoder) at Bell labs created quite an impact at the 1939 World' s fair, and was an early form of electronic speech synthesiser but one that require a good deal of human speech input (Gold, Morgan, & Ellis, 2011). Additionally, in 1953, Walter Lawrence introduced the Parametric Artificial Talker (PAT), the first formant synthesizer, which comprised three electronic formant resonators connected in parallel (Story, 2019).

Concatenative synthesis, the next significant stage, represented a major shift in the approach to speech synthesisTaylor (2009). Rather than relying on theoretical models of human speech production, this method uses actual recordings of human speech. Concatenative synthesis works by stitching together small segments of recorded speech, which are stored in a database. These segments can be as short as individual phonemes or as long as entire words. The quality and naturalness of the synthesized speech depend on the size and coverage of the speech database. Larger databases can produce more natural-sounding speech but require more storage and sophisticated algorithms to select and concatenate the segments smoothly. Despite its ability to generate highly natural speech, concatenative synthesis has limitations, such as the need for extensive databases and the challenge of ensuring smooth transitions between concatenated segments. Additionally, it lacks flexibility in terms of generating new or unseen words and expressions not present in the database.

With the advancements in electronic signal processing and computer technology, coupled with the increased availability of computing resources, there has been a shift from methods grounded in acoustic principles and linguistic knowledge towards approaches based on the inherent characteristics and patterns of data for constructing more natural synthetic sounds. In this context, the SPSS has emerged (Zen et al., 2009). This model utilizes statistical parameters to characterize the acoustic attributes of speech signals, such as sound frequency spectrum and vocal tract parameters. Initially, SPSS models relied on technologies like Hidden Markov Models (HMMs) or neural networks (Zen, 2015). These models were trained to understand the relationship between speech signals and their corresponding statistical parameters, thereby facilitating the synthesis of speech signals. These systems typically comprise

front-end modules, acoustic models, and vocoders. The front-end module converts text into acoustic features, the acoustic model maps these features to acoustic characteristics, and the vocoder generates speech waveforms from the acoustic features. Compared to traditional techniques that generate waveforms directly through cascading, this approach offers the advantage of enhancing the naturalness and flexibility of synthesized speech. However, these traditional systems exhibit notable performance limitations due to their complex structures and reliance on manual feature engineering, which hinder their further advancement.

In recent years, with the rise of deep learning technology, TTS technology has ushered in new development opportunities. From traditional HMM and DNN-based models to modern end-to-end text-to-speech models, technology has shown significant progress. It not only simplifies the synthesis process but is also proven to produce better speech quality than traditional speech synthesis technology. The TTS system can use the feature learning ability of the neural network to better express the speaker's pronunciation style and rhythm (Watts et al., 2019). Neural speech synthesis models usually first extract features from text and use an upgraded acoustic model to generate mel spectrograms, and then use a vocoder to synthesize audio. The first proposed modern neural TTS model was WaveNet (Oord et al., 2016), which has a very large perceptual field based on dilated causal convolution and can generate specified waveforms directly from language features. In 2017, Arik and Gibiansky presented the DeepVoice1 and DeepVoice2 systems (Arık et al., 2017; Gibiansky et al., 2017), high-quality text-to-speech systems built entirely from deep neural networks. The model still follows the three main modules of SPSS, where the acoustic model part is upgraded with the corresponding neural network-based model to build a TTS system entirely built from the deep learning framework. The biggest advantage of this system is that it can meet the requirements of real-time conversion and is 400 times faster than WaveNet. However, because the modules are too dispersed, error accumulation is prone to occur, which increases the difficulty of training. Then Ping et al. proposes DeepVoice3 (Ping et al., 2017), that is a fully convolutional network (FCN) system that integrates an attention mechanism, simplifying the main body of the model into three modules: encoder, decoder, and converter.

Until 2017, the release of Tacotron (Wang et al., 2017) marked the inception of the first end-to-end speech synthesis system. Tacotron primarily employs recurrent neural networks, aiding it in capturing the contextual relationships of input information, facilitating feature extraction within the model. Subsequently, it utilizes neural sequence-to-sequence models to jointly model the entire process from original text to speech. In 2018, Shen et al. introduced Tacotron2 (Elias et al., 2021), which utilizes a decoder based on autoregressive one-way long short-term memory (Hochreiter & Schmidhuber, 1997; Sak, Senior, & Beaufays, 2014) and a soft attention mechanism. On the same dataset, Tacotron2 achieved an MOS (Mean Opinion Score) of 4.53, while human speech scored 4.58, making it the highest MOS value achieved in text-to-speech history.

However, the Tacotron model still requires an acoustic post-processing step when generating speech, including the Griffin-Lim algorithm (Griffin & Lim, 1984) or a vocoder such as WaveNet. This step is used to convert the linear spectrogram into the final speech waveform. Therefore, although the Tacotron model is able to generate linear spectrograms directly from text, it does not directly generate the final speech waveform, which in a way makes it not a fully end-to-end model. Subsequently, fully end-to-end TTS systems emerged, such as ClariNet (Ping et al., 2018), which support waveform generation directly from text.

The emergence of the end-to-end TTS model is a milestone in the development of TTS technology. It simplifies the system structure and enables it to directly convert original text into final speech output. It improves the naturalness and understandability of speech and shrinks the size of synthetic speech and the difference between human voices. In the field of end-to-end TTS models, many models adopt non-autoregressive architectures. This architecture supports parallel data processing, where the model's output generation does not depend on the output of the previous time step, thus increasing processing speed. It also circumvents the error accumulation problem inherent in traditional autoregressive models, thereby significantly improving TTS performance.

FastSpeech (Ren et al., 2019) and its variants are typical representatives of non-autoregressive architectures. FastSpeech is a transformer-based feedforward network for parallel generation of TTS mel spectrograms. Specifically, Fastspeech extracts attention alignment from an encoder-decoder based teacher model for phoneme duration prediction, which is used by the length regulator to extend the source phoneme sequence to match the length of the target mel spectrogram sequence to Perform parallel mel spectrogram generation. Compared with autoregressive Transformer TTS, FastSpeech speeds up mel spectrogram generation by 270 times and speeds up end-to-end speech synthesis by 38 times.

The birth of the VITS (Kim et al., 2021) model in 2021 marks a qualitative leap in speech synthesis technology. The model adopts a parallel end-to-end TTS approach and is able to generate more natural audio than current two-stage models. The VITS model utilizes variational autoencoders (VAE) (Kingma & Welling, 2013) to connect the two modules of the TTS system through latent variables, thereby achieving efficient end-to-end learning. Its structure details are shown in the figure 1. In order to improve the expressive ability of the model in order to synthesize high-quality speech waveforms, the researchers applied the normalized flow to the conditional prior distribution on the waveform domain and conducted adversarial training.

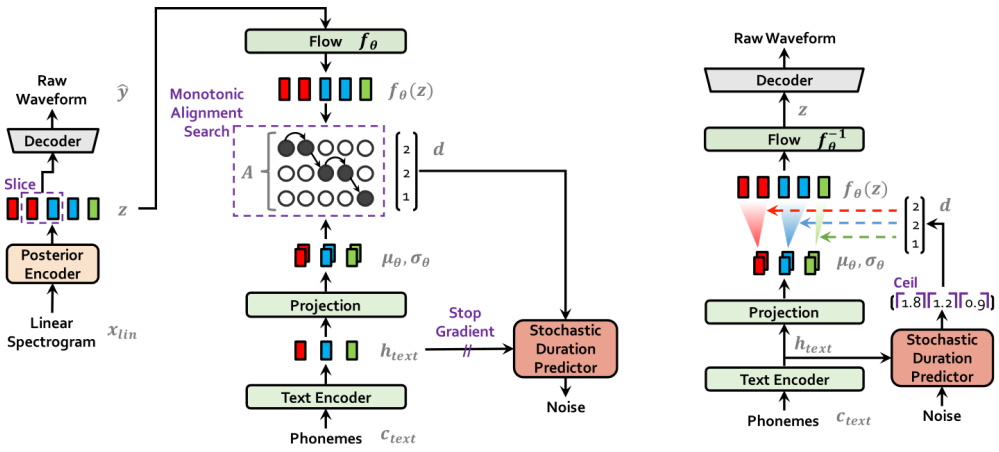

Figure 1: System diagram depicting (a) training procedure and (b) inference procedure. The proposed model can be viewed as a conditional VAE; a posterior encoder, decoder, and conditional prior (green blocks: a normalizing flow, linear projection layer, and text encoder) with a flow-based stochasticduration predictor. Kim et al. (2021).

Furthermore, they proposed a random duration predictor for synthesizing speech with different rhythms from input text. By modeling uncertainty in latent variables and the use of stochastic duration predictors, the model is able to naturally express one-to-many relationships, where text input can be pronounced in multiple ways with different tones and rhythms.3.2.

2.2 BERT

Although TTS models have made significant progress in imitating acoustic features, models may not accurately capture subtle semantic differences in the same input in different contexts because the training data is never comprehensive enough. Therefore, researchers began to use the transfer learning capabilities of the BERT model to improve the performance of TTS systems. The TTS system, which combines BERT's pre-training and fine-tuning, can better understand semantics and generate more natural speech, which is seen as an important advancement.

BERT(Devlin et al., 2018), Bidirectional Encoder Representations from Transformers, is a new language representation model that pre-trains deep bidirectional representations from unlabeled text by jointly considering the left and right context of all layers in the text. BERT utilizes a "mask language model" (MLM) pre-training objective. MLM randomly masks some tokens in the input, and its goal is to mask the original vocabulary ID of the word based only on contextual predictions. In addition to the masked language model, BERT also uses the "next sentence prediction" task to jointly pre-train text pair representations. BERT's contribution is to prove the importance of bidirectional pre-training for language representation, and to achieve pre-trained deep bi-directional representation through a mask language model. Furthermore, BERT is the first fine-tuning-based representation model that achieves state-of-the-art performance on many sentence-level and token-level tasks and surpasses many task-specific architectures. Its structure is shown in the figure 2

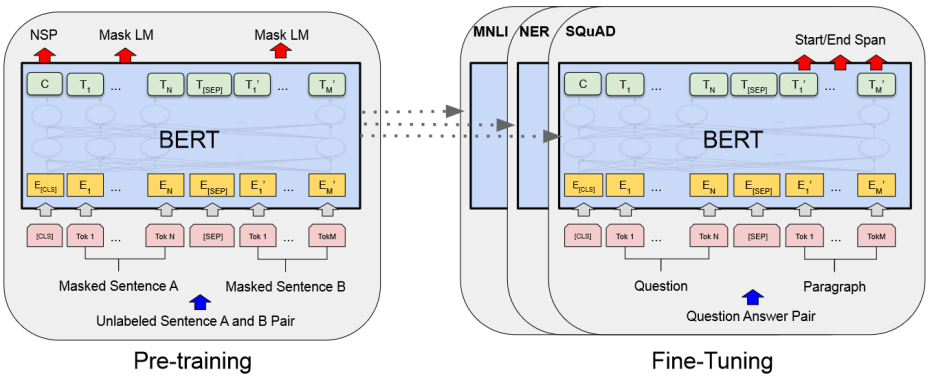

Figure 2: Overall pre-training and fine-tuning procedures for BERT Devlin et al. (2018).

(Hayashi et al., 2019) use pretrained representations (BERT) to encode input phrases as additional input to a Tacotron2-based sequence-to-sequence TTS model. Their subjective listening tests using the LJSpeech database show that the subword-level model variant slightly

but significantly improves the average opinion score compared to the baseline TTS model without pre-trained text embedding input. Similarly, Yang et al. (Yang, Zhong, & Liu, 2019) also applied a pre-trained BERT model to achieve enhanced front-end accuracy. The model's performance has been significantly improved on Mandarin multiphone disambiguation and prosodic structure prediction tasks. The model achieved an absolute improvement of 0.013 and 0.027 on F1 scores (Yacouby & Axman, 2020) for prosodic word prediction and prosodic phrase prediction, respectively, and an absolute improvement of 2.44% on multiphone disambiguation . Kenter et al. (2020) demonstrated that integrating a BERT model, pretrained on large amounts of unlabeled data and fine-tuned for speech, into an RNN-based TTS system can enhance prosody. Kenter et al. (Kenter, Sharma, & Clark, 2020) incorporated the BERT model into an RNN-based speech synthesis model, where the BERT model was pretrained on large amounts of unlabeled data and targeted Fine-tuning in the speech area. This experiment proves that BERT can improve the model's performance in prosody. Additionally, they proposed a method to use BERT to process arbitrarily long sequences, and the results showed that small BERT models performed better than large models, and that fine-tuning the BERT part of the model was crucial to obtain good results. Since prompt tuning has received widespread attention in guiding text or image generation, PromptTTS (Guo, Leng, Wu, Zhao, & Tan, 2023) takes the style and content description of the BERT model as input to generate speech with precise style control and high speech quality.

In particular, Mukherjee et al.(Mukherjee, Bansal, Satpal, & Mehta, 2022) proposed a novel text-aware emotion text-to-speech system that leverages a pre-trained BERT model to obtain deep representations of emotional context from text during training and inference. This method synthesizes emotional audio with emotion based on the emotional context of the input text, and the results show that this method outperforms baseline systems in terms of emotional intensity. Researchers such as (Kenter et al., 2020) apply word-level BERT to capture the semantic and syntactic structure of sentences, thereby aiding TTS synthesis. Li et al. (Y. A. Li, Han, Jiang, & Mesgarani, 2023) introduced phoneme-level BERT, in addition to conventional masked phoneme prediction, a preliminary task of predicting corresponding graphemes was also designed to enhance the naturalness of speech synthesized from out-of-distribution (OOD) text.

2.3 Transfer Learning

The traditional assumption of machine learning methods is that the training and testing data come from the same domain, ensuring consistency between the input feature space and data distribution characteristics. However, in some real-world machine learning scenarios, this assumption does not hold. In certain cases, acquiring training data may be expensive or challenging. Therefore, there is a need to develop high-performance learners that can be trained using data more easily obtained from different domains. This approach is known as transfer learning (Weiss, Khoshgoftaar, & Wang, 2016). In the field of speech synthesis, there are many applications of transfer learning. For example, cross-lingual transfer learning has been used to achieve end-to-end text-to-speech conversion for low-resource languages (Tu, Chen, Yeh, & Lee, 2019), where the method helps the model better utilize the knowledge learned from rich source data previously, resulting in more natural speech synthesis compared to models trained solely on target data. Furthermore, promising results have been achieved compared to

approaches relying solely on strong language background expertise (Tits, El Haddad, & Dutoit, 2020). Fine-tuning pre-trained deep learning-based TTS models allows for the synthesis of speech using a small dataset from another speaker.

In the field of speech synthesis, transfer learning is not only applied to cross-language TTS conversion, but also widely used in "cross-style" fine-tuning. Cross-style fine-tuning refers to the process of fine-tuning a trained speech synthesis model by using a small amount of data on different styles, so that the model can generate a variety of speech styles. This approach is of great significance in practical applications, such as enabling a single speech model to adapt to different emotional expressions, speaker styles, or the phonetic needs of specific scenes.

However, the performances of existing style transfer methods are still far behind real application needs. The root causes are mainly twofold. Firstly, the style embedding extracted from single reference speech can hardly provide fine-grained and appropriate prosody information for arbitrary text to synthesize. Secondly, in these models the content/text, prosody, and speaker timbre are usually highly entangled, it's therefore not realistic to expect a satisfied result when freely combining these components, such as to transfer speaking style between speakers. SHANG Zengqiang (2024) propose a cross-speaker style transfer TTS model with explicit prosody bottleneck. The prosody bottleneck builds up the kernels accounting for speaking style robustly, and disentangles the prosody from content and speaker timbre, therefore guarantees high quality cross-speaker style transfer.

In addition, Pan and He (2021) achieved the transition from a single speaker style to multiple speaker styles by fine-tuning existing high-quality speech synthesis models. They propose an End-to-End Multi-speaker Multi-style Multilingual speech synthesis model (E2E-3M) incorporating cross-speaker style transfer. The model uses a two-level variational autoencoder to model the generation process from text to waveform, and decouple the timbre, pronunciation and prosody information. In the inference stage, this method improves the prosody of the cross-language synthesis by transferring the prosody styles of the speakers in the target language. Experiments show that the proposed method can improve the prosodic naturalness of cross-language generation and enhance the sound quality of the generated speech.

3 Methodology

In order to prove the impact of datasets of different prosodic styles on the BERT-VITS model, this section first briefly introduces the selection and production of the open source dataset and the training process of the model. First, in subsection 3.1, We will introduce the source data selection and target data production preparation process. Next, subsection **??** will focus on the experimental settings, including training strategies, model characteristics, model settings, and training parameter settings. Subsection 3.3 will then elaborate on the evaluation method. Finally, in subsection 3.4, I will reflect on the ethical considerations inherent in this research.

3.1 Data Preparation

The quality of the speech synthesis dataset has a decisive impact on the performance of the speech synthesis model. For Mandarin broadcasting programs, training the model with clear and accurately pronounced Mandarin broadcasting datasets is a crucial step in determining the effectiveness of speech synthesis.

First, for a news broadcast style dataset with a formal focus on pronunciation clarity, we selected Baker, a large open source Mandarin standard database. The database contains 12 hours of speech data from a single female speaker, totaling 10,000 sentences. The speaker's vocal style is characterized as intelligent, sunny, friendly, and professionally standard Mandarin, conveying an optimistic and positive impression. The recordings were conducted in a professional recording studio using consistent recording software, maintaining a monaural recording format at a 48 kHz 16-bit sampling frequency in PCM WAV format. The recorded corpus covers various domains, including news, novels, technology, entertainment, and dialogues. The database was meticulously designed to comprehensively cover phonetic segments, types, tones, phonetic connections, and prosody within the limited data volume. The corpus underwent text-to-phoneme alignment, prosodic hierarchy annotation, and speech file boundary segmentation according to synthesis speech labeling standards.

For example, a particular training sample in this dataset might consist of the following components: The text format is in .txt, with one line of text and one line of pinyin. The text line starts with a sentence number, which comprises six Arabic numerals in half-width, separated by the Tab key. This is followed by the text content and ends with a carriage return and line feed. The corresponding pinyin line starts with the Tab key, followed by the text pinyin, with spaces separating the pinyin. It also ends with a carriage return and line feed.

Tone: Tones are marked with 1-5, where 1-4 correspond to the four tones of Mandarin, namely level tone, rising tone, falling-rising tone, and falling tone, and 5 represents the neutral tone.

Prosodic annotation: The prosodic structure annotation of Mandarin includes four levels of annotation: prosodic word (#1), prosodic phrase (#2), intonational phrase (#3), and sentence end (#4).However, please note that Prosodic annotations in the data set are not necessary in the actual experiment.

Example:

000384 那里的 #1 布局 #3 杂而 #1 有章 #3 乱而 #1 有序 #4。
na4 li3 de5 bu4 ju2 za2 er2 you3 zhang1 luan4 er2 you3 xu4

Syllable boundary segmentation: The segmentation of Mandarin into syllables and rimes is marked in the interval file format.

The file structure is as depicted in the Figure13.

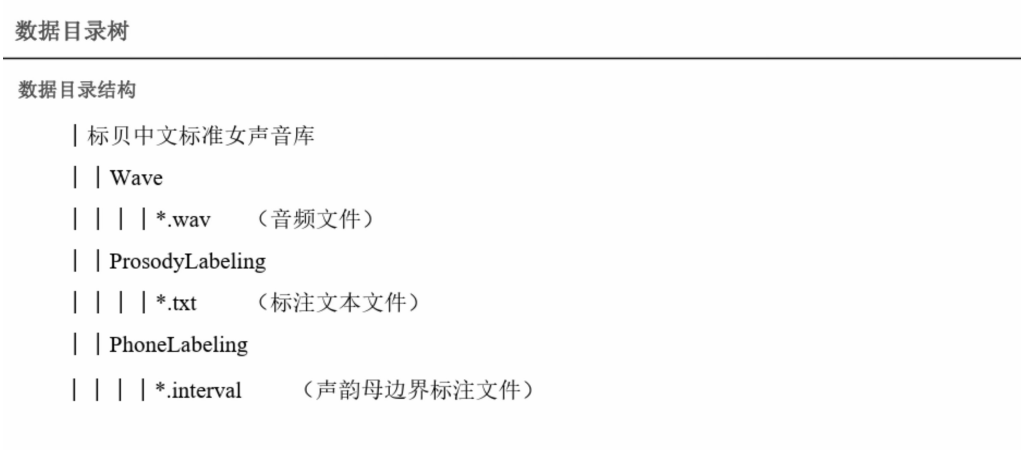

Figure 3: File structure of Baker.

For datasets characterized by a more natural prosodic style resembling real-life conversations, we gathered recordings from professional radio hosts' programs. These recordings consist of 1.5 hours of content from a single male speaker, predominantly featuring traditional Mandarin mythological stories and domestic and foreign novels, including substantial dialogues between characters and commentary from the host. The overall prosodic style of these recordings is more relaxed, with a natural and smooth tone, rich emotional variations, distinct from the Baker dataset. The recordings were conducted in environments and with equipment meeting the requirements for professional broadcasting program recording, requiring no additional noise reduction processing. Following the model's requirements, we normalized the audio by resampling it to a 16 kHz 16-bit sampling frequency, converted stereo to monaural, and saved all audio files in PCM WAV format.

Labeling: With audio resources prepared, the next step involves annotating all audio with Chinese text, i.e., audio-to-text transcription. To ensure accurate transcription, we initially employed the open-source speech recognition project, whisper [5], available on GitHub, for initial transcription, followed by manual verification and correction of all errors in the output. Additionally, after practices, we found that if the average length of audio in the training dataset is too long, it directly affects whether the model can run properly in subsequent steps, as the model imposes a limit on the length of audio input. Therefore, based on the speech intervals in long audio, we segmented the speech at the sentence level. Each segmented audio clip is approximately 10 to 20 seconds in length, ensuring that each audio segment ends with a complete sentence. Ultimately, we obtained an audio dataset from a single speaker, totaling around one hour and thirty minutes, comprising 300 sentences. The dataset includes audio data in WAV format, text transcriptions from each waveform file, and phonetic annotation files.

[5]Information about the whisper: `https://github.com/openai/whisper`

As for the test sentences, in the short audio clips part, we randomly selected some Chinese content to test, such as food stories, “竹笋在 4000 多年前就是席上珍羞，至今仍是大家桌上的常备菜”, meaning that bamboo shoots were treasured on the table more than 4,000 years ago, and are still common dishes on everyone's table today. For the long audio part, in order to be closer to the practical application field of radio program, we edited a radio program on the topic of women living alone, including a complete opening and closing remarks.

3.2 Experimental Setup

We first train the model on the female voice BAKER dataset. Then, we utilize this pre-trained model to train on another dataset, namely, our self-made small male voice dataset. Subsequently, we evaluate the training results through subjective listening tests.

The hyperparameters of the model are set as follows: 192 channels for input spectrogram, 192 channels for hidden layers, and 768 channels for filter. The convolutional kernel size in the residual block is set to 3, with different dilation factor settings. We employ a Dropout probability of 0.1 to reduce the risk of overfitting. Furthermore, we do not use Spectral Normalization.

Our model is trained using the AdamW optimizer, with a learning rate initialized to 1e-4 and parameters set to $\beta 1 = 0.8$ and $\beta 2 = 0.99$. The optimizer also utilizes weight decay $\lambda = 0.01$ to control the complexity of the model. The training process consists of 20,000 epochs, with the learning rate adjusted at a decay rate of 0.999875 per epoch. The batch size during training is 8, and we do not use mixed precision training. No warm-up operation is performed during training, and a segmented training strategy is not adopted.

During the experiment, we conducted training on a GPU with the model of A100. The pre-trained model was trained for 370,000 steps, while the fine-tuned model was trained for 200,000 steps.

3.3 Evaluation

We use subjective and objective methods to evaluate the synthesized Mandarin broadcast programs with different prosodic styles datasets. For subjective evaluation, we employed the Mean Opinion Score (MOS) (Viswanathan & Viswanathan, 2005) method. The participants were divided into two groups: general listeners and professional listeners. General listeners are non-broadcasting professionals without specialized knowledge of prosody styles; they judge the audio based purely on their subjective impressions. Professional listeners, on the other hand, are industry professionals in broadcasting who inevitably use their expertise to evaluate the prosody styles of the audio.

Dividing the participants into general listeners and professional listeners has several advantages. Firstly, it allows for a multidimensional evaluation of the model's performance. General listeners represent the actual audience of broadcast programs, and their evaluations reflect how the model performs in practical applications. Professional listeners, however, can provide more detailed assessments of the synthesized speech quality from a professional standpoint. Secondly, it helps balance subjective biases. The evaluations from general listeners may be influenced by personal preferences; for instance, some might have a stereotype that a deep male voice sounds more suitable for news broadcasts, overlooking genuine prosody issues.

In contrast, professional listeners can disregard the influence of speaker timbre and quality, focusing more objectively on the prosody. Overall, by splitting the participants into these two groups, we can comprehensively evaluate the synthesized speech's performance, providing essential guidance for model improvement and optimization. This multidimensional evaluation method not only reveals the model's practical application performance but also identifies technical issues, thereby enhancing the overall quality of the synthesized speech.

The questionnaire is divided into two parts: the first part evaluates short audio clips, with each group containing two synthesized audio samples of about 10 seconds each from different models, primarily assessing two dimensions: pronunciation accuracy that is, whether the tone is accurate, whether the pronunciation is clear and understandable and whether the prosody style is closer to a relaxed, natural conversational style. The second part evaluates longer audio clips, with each group containing two synthesized audio samples of about one minute each from different models, assessing the same dimensions as the short audio clips.

The main reasons for dividing the survey questionnaire into short and long audio segments are as follows. Firstly, compared to short audio segments, long ones better showcase the variations in speech prosody. They offer a longer time frame, allowing listeners to better perceive the prosody and fluency of speech. Secondly, although short audio segments are constrained by time, they provide a convenient way for listeners to replay them repeatedly to confirm pronunciation accuracy. While short segments may have fewer variations in prosodic styles, their repeatability makes them an essential tool for assessing the accuracy of speech synthesis models. In summary, by dividing the content into short and long audio segments, we can comprehensively evaluate the performance of speech synthesis models in different contexts.

The MOS rating system uses intervals of 1 point, with 1 being the lowest score, indicating very poor pronunciation and a prosody style very much like a news broadcast rather than everyday conversation, and 5 being the highest score, indicating very accurate pronunciation and a prosody style very close to everyday conversation rather than a news broadcast. The final MOS score is the arithmetic mean of these scores.

In order to verify the experimental results comprehensively, we also use the Librosa [6] library to extract and analyze features of audio data, and use visualization methods to allow readers to more intuitively see the difference in performance between the pre-trained model and the fine-tuned model.

3.4 Ethical considerations

Although this study aims to investigate the impact of different prosodic style datasets on model performance, the technology may bring some unforeseen consequences. To mitigate these risks, we will communicate the study's results and implications in an accessible and transparent manner.

We collected voice data from a Mandarin radio program host, with his permission for its use, as documented in the consent form in the appendix. This portion of the voice data is a Mandarin male voice corpus, including storytelling, talk shows, and book reading. This data is used solely for this research. Additionally, we used a standard Mandarin female voice dataset from Beijing Data Baker Technology Co., Ltd. This is a monolingual, open corpus.

[6]Information about the Librosa: `https://github.com/librosa/librosa`

Regarding the listening test, we utilize the Qualtrics[7]platform, a powerful online survey tool that enables us to quickly create and publish questionnaires, and to collect and analyze data, as well as generate reports and charts. Prior to all participants engaging in the hearing test, they will sign an informed consent form. The informed consent form clearly states the academic purpose of the hearing test and the associated risks. After the completion of this experiment, the questionnaire will be closed and will no longer be accessible to the public. All data will be used solely for this academic research and will not be used for any other purposes.

Regarding the reproducibility of the research, the code is available on GitHub. The URLs for each model can be found in the thesis. The Baker dataset is also publicly available for download and use. The results should be more or less similar, but may not be exactly the same, due to certain factors that introduce randomness in the trained models. The hardware used may also impact the performance of the models, as the experiments were conducted on the University of Groningen's high-performance cluster, Habrók.

[7]Information about the model: `https://www.qualtrics.com/`

4 Results

Table 1: Non-professional Group Scores

	Fine-tuned		Pre-trained	
Audio	Accuracy	Prosody	Accuracy	Prosody
short1	4.44	3.72	3.00	3.11
short2	4.50	3.77	4.05	3.22
short3	4.16	4.16	3.88	2.66
short4	4.11	3.50	4.11	2.72
Average	4.30	3.78	3.76	2.92
long1	3.94	3.94	3.94	2.55
long2	3.77	3.94	3.38	2.38
Average	3.85	3.94	3.66	2.45

Table 2: Professional Group Scores

	Fine-tuned		Pre-trained	
Audio	Accuracy	Prosody	Accuracy	Prosody
short1	4.06	3.68	3.43	2.87
short2	3.87	3.61	3.75	3.31
short3	4.43	4.00	3.81	2.56
short4	4.06	3.81	3.87	3.00
Average	4.10	3.77	3.71	2.93
long1	4.18	4.12	3.81	2.93
long2	3.93	4.06	3.68	2.75
Average	4.05	4.09	3.74	2.84

In this study, we obtained detailed data through subjective listening tests and evaluated the performance of pre-trained and fine-tuned models in speech synthesis tasks by combining statistical Mean Opinion Score (MOS) ratings(1&2), box plots 5 6, and two independent-sample t -test3. The pre-trained model was trained for 370,000 steps on a large source dataset containing 12 hours of audio and 10,000 audio files, while the fine-tuned model was further trained for 200,000 steps on a relatively smaller dataset, containing only 1.5 hours and 300 audio files. In section 3.3,we mentioned dividing the subjects into two groups for the listening tests. One group comprised 16 listeners with professional broadcasting knowledge, who used their expertise to judge the synthesized results during the tests; the other group consisted of 18 ordinary listeners who relied primarily on subjective impressions to judge the prosodic style and pronunciation accuracy. The entire listening test was divided into short audio and long audio segments. The experimental results provide a more comprehensive understanding of the prosodic characteristics of the training data on the model performance.

Table 3: Comparison of Statistical Test Results

Test Type	**Statistic**	**p-value**
Accuracy t-test	t-value = 1.9999	p-value = 0.0734
Prosody t-test	t-value = 6.5612	p-value = 6.3823e-05
Accuracy Welch's t-test	t-value = 1.9999	p-value = 0.0784
Prosody Welch's t-test	t-value = 6.5612	p-value = 0.0001

4.1 Pronunciation Accuracy Results

Despite the significant difference in the sizes of the datasets used to train the two models, their performance in terms of pronunciation accuracy is surprisingly similar. According to the MOS data, the two groups of participants had similar perceptions of pronunciation accuracy, generally believing that the fine-tuned model was slightly better than the pre-trained model in this aspect. Broadcast professionals generally felt that the fine-tuned model performed significantly better in pronunciation accuracy compared to the pre-trained model. However, ordinary listeners found little difference between the two, with a median difference of only about 0.2. This suggests that professionals are indeed better at detecting subtle issues in synthesized speech compared to non-professionals.

We then conducted an independent sample t-test on the pronunciation accuracy of the two models, yielding the following results:

- The t-value was 1.9999023509019753.
- Assuming equal variances, the p-value was 0.07339997734485468.
- Using Welch's t-test (not assuming equal variances), the p-value was slightly higher at 0.07840098051748379.

Both p-values are above the commonly used significance level of 0.05, indicating that we do not have sufficient evidence to reject the null hypothesis. That is, there is no significant difference in pronunciation accuracy between the fine-tuned and pre-trained models. This similarity in pronunciation accuracy suggests that the general speech processing knowledge learned by the pre-trained model has strong generalization capabilities, enabling it to achieve similar performance across datasets of different sizes and characteristics.

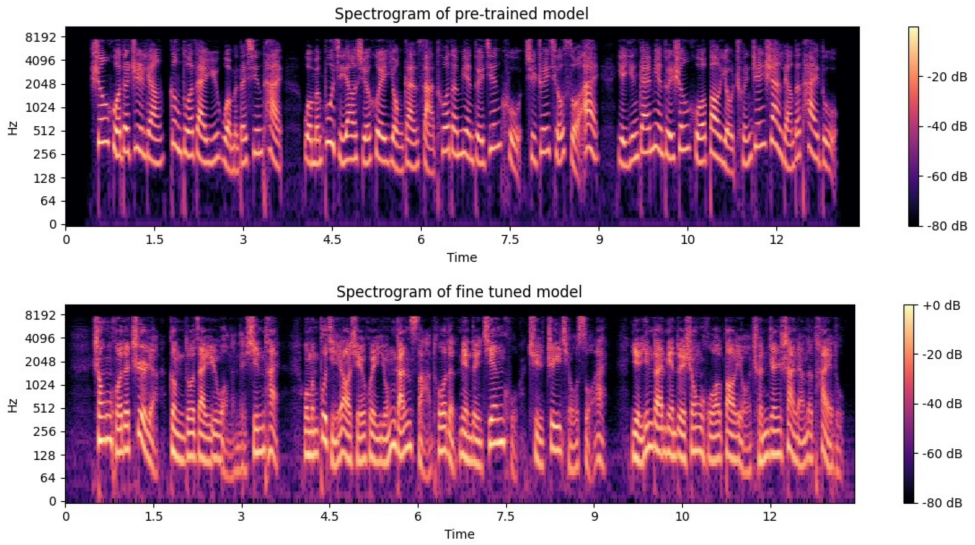

Figure 4: Spectrogram comparison of speech synthesized by the pre-trained and fine-tuned models containing the same text.

4.2 Prosodic Performance Results

In terms of rhythm performance, the fine-tuned model is more dynamic and flexible in its rhythmic expression than the pre-trained model, regardless of whether it is for long or short audio segments, and regardless of the group of participants. This means that the fine-tuned model's rhythm is closer to the relaxed and natural style of everyday conversation, rather than the more standardized and stable prosody of news broadcasting. Overall, the rhythm performance of short audio segments is slightly better than that of long audio segments. This is because short audio segments are time-limited and do not have the opportunity to exhibit more rhythmic variations, whereas long audio segments can better reflect the influence of the training data's prosodic style on the synthesis results.

To visualize the prosody performance of the two models, we used the librosa library to obtain spectrograms4 of two audio samples from different models but with the same textual content. From the spectrograms, we can see that the audio from the pre-trained model has relatively uniform intervals and a relatively smooth energy distribution, indicating that its prosody is quite stable and consistent. In contrast, the spectrogram of the fine-tuned model shows more details and a more complex prosodic pattern. The denser vertical stripes in the spectrogram indicate a higher frequency of pitch changes, resulting in a richer prosody. The energy distribution is also more varied, especially in the higher and lower frequency regions,

where the color variation is greater, indicating that the fine-tuned model's synthesized audio has more variation in volume and pitch.

To verify our hypothesis, we conducted a t-test on the rhythm performance, yielding a significant t-value of 6.561183597917967. The specific results are as follows:

- Assuming equal variances, the p-value was 6.382266132680388e-05.
- Using Welch's t-test (not assuming equal variances), the p-value was 0.00010845064834172377.

Both p-values are far below 0.05, indicating that we have very strong evidence to reject the null hypothesis. That is, the fine-tuned model's performance in prosody is indeed significantly better than that of the pre-trained model.

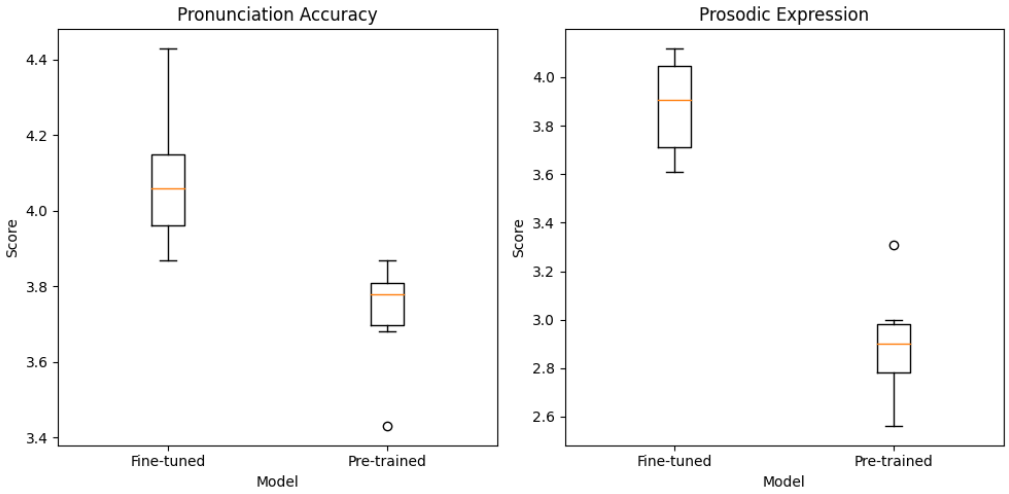

Figure 5: Ratings of pronunciation accuracy and prosodic expression by broadcasting professional participants.

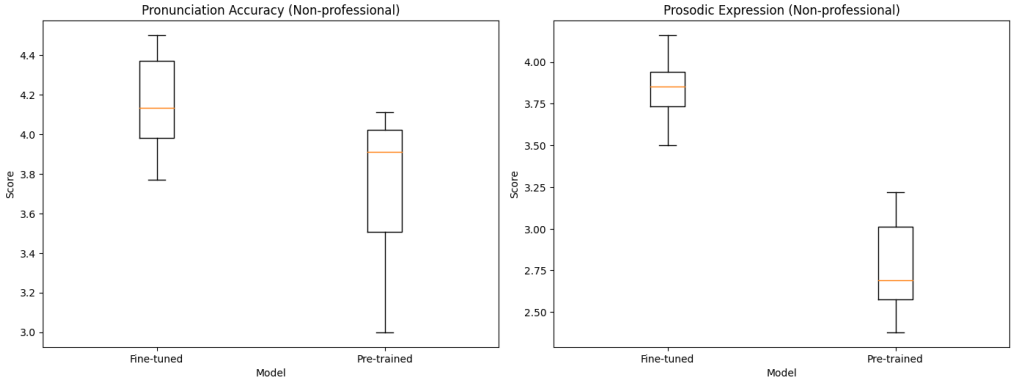

Figure 6: Ratings of pronunciation accuracy and prosodic expression by non-broadcasting professional participants.

4.3 Summary

In summary, combining the results of the t-tests and the subjective listening test analysis, we can conclude that there is no significant performance difference between the fine-tuned and pre-trained models in terms of pronunciation accuracy. This indicates that the general speech processing knowledge learned by the pre-trained model has strong generalization capabilities. However, in terms of rhythm performance, the fine-tuned model significantly outperforms the pre-trained model. The fine-tuned model exhibits more complex and natural rhythmic characteristics, closer to the style of everyday conversation, whereas the pre-trained model tends to have a more stable, news-broadcast-like intonation with consistent volume and pitch.

Therefore, although the two models perform similarly in pronunciation accuracy, the fine-tuned model has a significant advantage in rhythm performance. It is better suited for speech synthesis tasks that require natural, conversational speech flow.

5 Discussion

Based on the analysis of the results presented in the previous section, it is evident that our hypothesis in Subsection 1.2, namely that the BERT-VITS model, after pre-training on a large standard Mandarin dataset emphasizing smooth intonation, pronunciation accuracy, and clear expression, followed by fine-tuning on a dataset with a more relaxed and conversational prosodic style emphasizing friendliness and naturalness, indeed produces speech synthesis results that ensure basic pronunciation accuracy while offering a more varied and relaxed prosodic style than the source dataset. This suggests that different prosodic styles in datasets significantly influence model performance. In the subsequent subsections, we delve into the reasons behind the results presented in the tables1 2.

5.1 Prosodic Expressiveness

As mentioned in the previous section, the prosodic style of the model after fine-tuning tends to lean towards a casual conversational style. In other words, listeners can perceive relatively rich prosody variations in the synthesized results, whereas the prosodic style of the pre-trained model is more neutral, with MOS (Mean Opinion Score) ranging from 2.8 to 2.9. This disparity is due to the relatively stable prosody performance of the training dataset used for the pre-trained model. It is noteworthy that almost all 15 professional broadcasters who participated in the listening test expressed a consistent opinion: *The audio sounds positive and clear but lacks prosodic variation.* This is consistent with the characteristics of the samples in the source dataset. In contrast, the target dataset used for fine-tuning the model contains a large amount of narrative content, with samples themselves exhibiting strong emotional expression capabilities. For instance, in narrative performances, besides containing relatively deep and stable narrative readings, they also include impromptu analyses and interpretations of book contents by the hosts, showcasing strong individual characteristics. Consequently, the model fine-tuned on such data exhibits more varied prosody performance.

5.2 Pronunciation Accuracy

Regarding pronunciation accuracy, the evaluations by professional broadcasters and non-professional participants of the synthesized results generally align with our expectations. Broadcasters inevitably leverage their industry knowledge to make judgments, thus being sensitive to subtle differences. Consequently, as shown in Table1, there are significant differences in pronunciation accuracy between the two models. However, for ordinary participants, they may just get a rough idea, and as long as it does not affect the overall understanding of the content, it is difficult for them to judge whether the pronunciation is standard.

5.3 Limitations

Our experiments were conducted only on small, clean datasets, and further exploration is needed to understand their impact on more complex datasets. Additionally, we did not account for homophones and tone dissimilationWenjie (2013) (a phonetic phenomenon, which refers to the pronunciation of a single word in the flow of speech, or is affected by the phonemes before

and after, and the pronunciation of the phenomenon is different from that when it is alone) in Mandarin within our system, such as the tonal adjustments that occur when two third-tones or fourth-tones appear side by side, which significantly impacts the synthesized results.

Specifically, in the audio "short2", the content is: " 生活的质量更多在于我们的心态，我们不仅要学会勇敢洒脱的面对艰苦去追求索爱，也应该面对生活保佑纯真善良的态度", which means that "The quality of life depends more on our mindset. We must not only learn to face hardships bravely and pursue love with a free and easy attitude but also maintain a pure and kind attitude towards life." In the word " 勇敢 (brave)" both " 勇" (brave) and " 敢" (daring) are characters with the third tone. According to the tone sandhi rules in Mandarin, when two third-tone characters appear consecutively, the tone of the first character changes to the second tone. In the field of speech synthesis, this subtle phonetic feature is crucial for generating natural and fluent speech. However, in the pre-trained model, we found that the character " 勇" did not follow the tone sandhi rules of Mandarin and still appeared in the third tone. This indicates that the model did not take into account the influence of tone sandhi when generating speech.

In contrast, the fine-tuned model correctly adjusted the tone of " 勇" to the second tone. Whereas, it is worth noting that this improvement did not come from specific tone training for the model, but rather because the model was exposed to a large number of data samples after adjusted during the fine-tuning process. In other words, the fine-tuned model was able to handle the tone of " 勇" correctly because it frequently encountered the combination of " 勇" and " 敢" with adjusted tone in the training data. If we can train the model with special tonal changes at an early stage, the model will be capable of not relying on a dataset, but can freely process the tonal variations in Mandarin, making the speech synthesis more natural and smooth.

Moreover, although BERT is recognized for its ability to provide detailed insights into word importance, syntax, semantics, and general knowledge, its effectiveness is limited by the specificity of the fine-tuning method (Hayashi et al. (2019); Kenter et al. (2020)). Furthermore, BERT's inherent non-generative nature may limit its ability to interpret information beyond direct sentence context, thus we did not conduct further research on the impact of prosody style on synthesized results.(Feng and Yoshimoto (2024))

6 Conclusion

The primary objective of this study was to evaluate and analyze the performance of the BERT-VITS model across datasets with varying prosodic styles and to explore its potential applications in the production of Mandarin broadcasting programs. By initially training on a source dataset characterized by a stable and formal news-broadcasting prosody, followed by fine-tuning on a target dataset with a more relaxed and conversational tone, the study aimed to diversify the prosodic styles of AI broadcasting. This diversification can effectively assist broadcasting institutions in utilizing speech synthesis technology to produce a richer array of radio programs. The research employed a combination of subjective and objective evaluation methods, including Mean Opinion Score (MOS) assessments and spectrogram analyses, to comprehensively measure the model's pronunciation accuracy and prosodic performance.

6.1 Contributions

The study revealed that the fine-tuned BERT-VITS model demonstrated significant naturalness and diversity in prosodic expression compared to the pre-trained model, more effectively mimicking the prosody of everyday conversation. This finding holds substantial practical significance for the broadcasting industry, as it indicates that with careful adjustment, speech synthesis models can be customized to suit different program styles and voice requirements. Moreover, the study confirmed that fine-tuning can effectively enhance the model's prosodic performance even with a smaller dataset, which is beneficial for both individual users and broadcasting organizations. This approach requires less time for the production of a smaller dataset, yet still yields desirable outcomes.

6.2 Practical Applications & Significance

The outcomes of this research have direct practical implications for the broadcasting industry, particularly in enhancing the efficiency of program production. Customized speech synthesis technology allows radio programs to be presented in a more natural and engaging manner to the audience, increasing the interactivity and warmth of AI-assisted broadcasting programs. Additionally, this technology may spur innovation in multilingual broadcasting content, offering a more diverse and personalized auditory experience for listeners from different linguistic and cultural backgrounds.

6.3 Future Research:

Future studies can further explore the performance of the BERT-VITS model in a broader range of speech synthesis applications, including voice generation for different languages, dialects, and specific domains. Furthermore, researchers have begun to experiment with an innovative method that combines large-scale language models, such as Llama-VITS (Feng and Yoshimoto (2024)). This approach enriches the semantic content of text for text-to-speech(TTS) synthesis by leveraging the semantic embeddings of large-scale language models like Llama2, integrated with the state-of-the-art end-to-end TTS framework, VITS. There is potential for more innovation in model integration in future research.

References

Arık, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., … others (2017). Deep voice: Real-time neural text-to-speech. In *International conference on machine learning* (pp. 195–204).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elias, I., Zen, H., Shen, J., Zhang, Y., Jia, Y., Skerry-Ryan, R., & Wu, Y. (2021). Parallel tacotron 2: A non-autoregressive neural tts model with differentiable duration modeling. *arXiv preprint arXiv:2103.14574*.

Feng, X., & Yoshimoto, A. (2024). Llama-vits: Enhancing tts synthesis with semantic awareness. *arXiv preprint arXiv:2404.06714*.

Gibiansky, A., Arik, S., Diamos, G., Miller, J., Peng, K., Ping, W., … Zhou, Y. (2017). Deep voice 2: Multi-speaker neural text-to-speech. *Advances in neural information processing systems*, *30*.

Gold, B., Morgan, N., & Ellis, D. (2011). *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons.

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, *32*(2), 236–243.

Guo, Z., Leng, Y., Wu, Y., Zhao, S., & Tan, X. (2023). Prompttts: Controllable text-to-speech with text descriptions. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Hayashi, T., Watanabe, S., Toda, T., Takeda, K., Toshniwal, S., & Livescu, K. (2019). Pre-Trained Text Embeddings for Enhanced Text-to-Speech Synthesis. In *Proc. interspeech 2019* (pp. 4430–4434). doi: 10.21437/Interspeech.2019-3177

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Kenter, T., Sharma, M., & Clark, R. (2020). Improving the prosody of rnn-based english text-to-speech synthesis by incorporating a bert model. In *Interspeech* (Vol. 2020, pp. 4412–4416).

Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International conference on machine learning* (pp. 5530–5540).

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Li, X. (2016). *Practical oral expression and broadcast hosting*. Communication University of China Press.

Li, Y. A., Han, C., Jiang, X., & Mesgarani, N. (2023). Phoneme-level bert for enhanced prosody of text-to-speech with grapheme predictions. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Mukherjee, A., Bansal, S., Satpal, S., & Mehta, R. K. (2022). Text aware emotional text-to-speech with bert. In *Interspeech* (pp. 4601–4605).

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., … Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Pan, S., & He, L. (2021). Cross-speaker style transfer with prosody bottleneck in neural speech synthesis. *arXiv preprint arXiv:2107.12562*.

Ping, W., Peng, K., & Chen, J. (2018). Clarinet: Parallel wave generation in end-to-end text-to-speech. *arXiv preprint arXiv:1807.07281*.

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., … Miller, J. (2017). Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, *32*.

Saito, Y., Takamichi, S., & Saruwatari, H. (2017). Statistical parametric speech synthesis incorporating generative adversarial networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(1), 84–96.

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*.

SHANG Zengqiang, W. L., ZHANG Pengyuan. (2024). Multilingual text-to-waveform with cross-speaker prosody transfer. *ACTA ACUSTICA*, *49*(1), 171-180. Retrieved from `https://www.jac.ac.cn/cn/article/doi/10.12395/0371-0025.2022146` doi: 10.12395/0371-0025.2022146

Story, B. H. (2019). History of speech synthesis. In *The routledge handbook of phonetics* (pp. 9–33). Routledge.

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge university press.

Tits, N., El Haddad, K., & Dutoit, T. (2020). Exploring transfer learning for low resource emotional tts. In *Intelligent systems and applications: Proceedings of the 2019 intelligent systems conference (intellisys) volume 1* (pp. 52–60).

Tu, T., Chen, Y.-J., Yeh, C.-c., & Lee, H.-Y. (2019). End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. *arXiv preprint arXiv:1904.06508*.

Viswanathan, M., & Viswanathan, M. (2005). Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale. *Computer speech & language*, *19*(1), 55–83.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., … others (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Watts, O., Eje Henter, G., Fong, J., & Valentini-Botinhao, C. (2019). Where do the improvements come from in sequence-to-sequence neural TTS? In *Proc. 10th isca workshop on speech synthesis (ssw 10)* (pp. 217–222). doi: 10.21437/SSW.2019-39

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*, 1–40.

Wenjie, L. (2013). 语流音变–汉语有声语言中不可忽视的因素. 科学技术创新, *000*(035), 113-113.

Wu T., F. Z. (2023). Voice is power: the development of news broadcasting and hosting creative styles. *doi:CNKI:SUN:WWJH.0.2023-21-034.*.

Yacouby, R., & Axman, D. (2020). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of nlp systems* (pp. 79–91).

Yang, B., Zhong, J., & Liu, S. (2019). Pre-trained text representations for improving front-end text processing in mandarin text-to-speech synthesis. In *Interspeech* (pp. 4480–4484).

Ying, L. (2018). Application of artificial intelligence technology in the field of broadcasting and hosting. *doi:CNKI:SUN:GDXK.0.2018-11-033.*.

Zen, H. (2015). Acoustic modeling in statistical parametric speech synthesis-from hmm to lstm-rnn. *Proc. MLSLP*, *15*.

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

Zhao W., C. J. e. a., Lian Y. (2023). Multi-speaker chinese news broadcasting system based on improved tacotron2. *The Visual Computer*. doi: 10.1007/s11042-023-15279-z

Appendices

A Questionnaire Details for MOS Evaluation

Here, we will present in detail the survey questionnaire designed to obtain the Mean Opinion Score (MOS).

Before commencing the questionnaire, participants are asked to read the Informed Consent document. If they agree, they can continue; if not, they have the option to exit or click the "Disagree" button, which will direct them to the final page. Subsequently, participants are requested to read the questionnaire instructions. This section briefly introduces the structure of the survey and the key points assessed by each question. Following this, participants are provided with two samples of daily program recordings from a live radio host, one for news broadcasting and the other for a talk show, with the aim of familiarizing participants with these two distinct hosting styles.

The formal questionnaire is divided into two parts: Chapter One contains 4 sets of short audio clips around 10 seconds in duration, and Chapter Two contains 2 sets of longer audio clips around 1 minute in duration. This arrangement is designed to demonstrate whether there is a difference in the model's performance in speech synthesis tasks of varying duration. Each page presents two audio clips synthesized by different models, but with identical text content, making it easier for participants to discern the differences between the two models. The questions for each audio clip are consistent, with example questions as follows.

- **Audio Clip**

  1. How accurate do you think the pronunciation of this audio clip is? "Pronunciation accuracy" refers to the correctness of tone and the precision of consonant and vowel pronunciation.
     - 5 points: Very accurate
     - 4 points: Mostly accurate with minor errors
     - 3 points: Moderately accurate
     - 2 points: Several inaccuracies, less accurate
     - 1 point: Very inaccurate
  2. How natural and conversational do you perceive the prosodic style of this audio clip to be in the context of a broadcast program? That is, does the broadcast host's style in the audio clip lean towards an informal, natural, and vivid "spoken" news style, or is it more akin to a formal and standardized "broadcast" news style?
     - 5 points: Totally speaking, casual and storytelling prosodic style
     - 4 points: More like speaking, casual and storytelling prosodic style over formal broadcasting style
     - 3 points: Neutral delivery style
     - 2 points: More like formal newscasts style than speaking, casual and storytelling prosodic style
     - 1 point: Totally like a newscast's tone

A total of 32 participants were invited to take part in this questionnaire, 15 of whom are professionals with experience in the broadcasting industry. The remaining participants are general listeners without specialized knowledge in broadcasting hosting.

B Voice Data Use Authorization Form

The voice data use authorization form can be found on the next page; it was pushed there due to the PDF import.

C Test Samples

Test samples are available at https://poppyyanliao.github.io/tts-research.github.io/

Data Use Authorization

Project Title: The Impact of Prosodic Style Transfer Learning on Mandarin Text-to-Speech (TTS) Performance Using BERT-VITS Model

Researcher: Yanhua Liao
Institution: University of Groningen
Contact: august910821@gmail.com
Participant Name:Ran Ma
Recording Content: Mandarin male voice corpus, including storytelling, talk shows, and book readings

Description:

This authorization/consent form is intended to obtain your permission to use your recorded data for the above-mentioned research project. The research results will be used for academic research and publication and will be communicated in an accessible and transparent manner.

Authorization Details:

1. Scope of Data Use:
    - You authorize the researcher to use your recorded data for training and evaluating TTS models.
    - The recorded data will only be used for this research and will not be used for other commercial purposes.
2. Data Protection:
    - Your recorded data will be strictly protected during the research and will not be accessed by unauthorized persons.
    - Your personal information will be anonymized to ensure privacy and security.
3. Data Disclosure:
    - Your recorded data may be partially disclosed in the research results but will not include any information that can identify you personally.
4. Rights and Obligations:
    - You have the right to withdraw your consent at any time and stop the use of your recorded data.
    - The researcher is obliged to ensure that your recorded data is only used for the purposes of this research.
5. Research Reproducibility:
    - The code for this research will be available on GitHub, and the URLs for all used datasets and models will be included in the research paper.

- The disclosed datasets and code will ensure the reproducibility of the research.

Statement:

I have read and understood the above content and agree to authorize the researcher to use my recorded data for this research.

Participant Signature: _____ Date: _____

Researcher Signature: _____ Date: _____

D Others

In order to facilitate readers to better reproduce the experiments in this paper, we have added more documents to github https://github.com/PoppyYanLiao/thesis-research