university of groningen / campus fryslân

An Innovative Method for Multi-Effect Speech Synthesis through Training File Modification

Yilan Wei

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

An Innovative Method for Multi-Effect Speech Synthesis through Training File Modification

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. Matt Coler (Voice Technology, University of Groningen)
with the second reader being
Dr. Shekhar Nayak (Voice Technology, University of Groningen)

Yilan Wei (s5515939)

June 11, 2024

Contents

Abstract

Human language naturally and flexibly adjusts speech rate, intonation, and voice intensity during communication. However, such dynamic changes are often inadequately modeled in current speech synthesis research. Most existing studies focus on generating audio with specific emotional tones (e.g., happy, sad, angry), but few address synthesizing audio with varied speech modifications, such as changes in speech speed and pitch adjustments within a single sentence. To address this gap, this study proposes an innovative method for multi-effect speech synthesis using the FastSpeech2 model by precisely modifying the training files and corresponding audio data. Experimental results demonstrate that this approach significantly enhances the model's ability to reproduce target speech modifications, yielding excellent performance in Chinese, English, and Spanish. Numerical analyses and manual listening assessments validate the model's sensitivity and accuracy to speech rate adjustments. Additionally, the study demonstrates the cross-linguistic generalizability and validity of the method, indicating a wide range of potential applications. This method is expected to contribute to more emotionally expressive and diverse audio synthesis, advancing speech synthesis technology.

Keyword: Speech Synthesis, Effect Control, Emotional Expression, Deep Learning, Multilingual Synthesis

1 Introduction

In the field of speech synthesis, with the development of large language models (LLMs) such as OpenAI's ChatGPT, the interaction methods and capabilities of voice assistants have been significantly improved. For example, ChatGPT-powered health assistants can adjust the tone and speed of speech based on the patient's mood and context to convey an appropriate sense of empathy and urgency when reminding patients to take their medication on time. In contrast, traditional virtual assistants such as Siri, Alexa, and Google Assistant are widely used for everyday tasks such as setting reminders, controlling smart home devices, and providing interactive entertainment, but they still fall short in handling complex dialogues, understanding context, and expressing emotions (Mahmood et al., 2023). Traditional speech synthesis methods used by these assistants typically process text at an average pitch and rate of speech, resulting in synthesized speech that lacks the variability and emotional richness of human speech. In comparison, humans can naturally and flexibly adjust their speech rate, intonation, and voice intensity when communicating.

For this reason, this paper proposes a simple and effective speech synthesis method that dynamically adjusts speech effects such as speech rate and emphasis without requiring significant modifications to existing models. It allows flexible adjustment of speech features during the synthesis process, which maintains the stability and efficiency of the text-to-speech (TTS) system while enhancing the expressiveness of the speech output to match the natural language habits of humans more closely.

Experimental validation shows that this speech synthesis method has excellent reproducibility and is suitable for speech synthesis in multiple languages. Moreover, the method demonstrates a wide range of social value and commercial application potential in several fields. In the field of education, it can help students better understand and remember learning content by emphasizing keywords or phrases. In the healthcare experience, this technology can improve the quality of interaction for voice assistants and customer service bots, enabling them to deliver urgent information or provide emotional comfort more effectively, thus making communication more human and engaging. In the field of commercial advertising, it can highlight key messages and effectively capture customers' attention. These examples of potential applications highlight not only the practicality of the technique but also its wide applicability in the real world.

To explore the specific implementation and application of this approach in more depth, the next chapters of this thesis will detail various aspects of the research. Chapter 2, Literature Review, will provide an overview of the development of speech synthesis modeling (Section 2.1), control techniques for emotional speech features (Section 2.2), and application areas of speech synthesis technology (Section 2.3), as well as identifying the problems and hypotheses of this research (Section 2.4). Chapter 3, Methodology, describes in detail the corpus selection (Section 3.1) and data pre-

processing steps (Section 3.2), covering the configuration of the preliminary and formal experiments. Chapter 4, Results and Discussion, presents the results of the experiments (Section 4.1), including numerical analyses and artificial auditory evaluations, and discusses the limitations of the current study and directions for future improvement (Section 4.2). Finally, Chapter 5 summarizes the main findings of the whole study and provides a comprehensive review of the research results.

2 Literature Review

This chapter aims to provide a comprehensive overview of the current state of speech synthesis technology, focusing on the evolution of the model, the methods used to control the effects of synthesized speech, and the various application areas. I will first review the evolution of speech synthesis models, highlighting the main advances and ongoing challenges (Section 2.1). Then, a thorough discussion of the specific techniques used to control emotional and rhythmic features in synthesized speech will be presented, assessing their effectiveness and limitations (Section 2.2). Finally, I will explore the various applications of speech synthesis in different fields, emphasizing the practical implications of these techniques (Section 2.3) and clarifying the questions and hypotheses of this research (Section 2.4).

2.1 The Evolution of Speech Synthesis Models

The development of speech synthesis techniques has gone through a revolution from traditional methods to deep learning-driven, which in turn evolved into real-time and end-to-end efficient models.

Traditional speech synthesis models, such as HTS and Merlin, are mainly based on Hidden Markov Models (HMM) for forced alignment. While these systems can learn efficiently on limited datasets, the speech they generate usually lacks a sense of naturalness and fluency, sounding more mechanical and monotonous. Since this alignment is fixed and the architecture relies on manual feature engineering and acoustic modeling, they have limited learning capabilities, which limits their expressiveness and naturalness (Wu et al., 2016).

With the rise of deep learning, neural network-based systems, especially Tacotron 2, are becoming a hot research topic. These models adopt an end-to-end approach to directly convert text to speech, dramatically simplifying the speech synthesis process. Tacotron2 effectively maps complete input sequences to speech by introducing an attention mechanism. It incorporates an improved WaveNet vocoder, which significantly improves the naturalness and expressiveness of synthesized speech (Shen et al., 2018). However, although Tacotron2 enhances the naturalness of synthesized speech, it may face coherence and consistency issues when processing longer texts, and it has a high demand for computational resources. These drawbacks limit its application in resource-constrained environments.

Furthermore, the emergence of end-to-end acoustic models such as WaveNet and WaveGlow marks a further leap in the technology. The main innovations of WaveNet are the use of dilated causal convolution to increase the receptive field and the optimization of the network training through residuals and jump connections. This approach allows WaveNet to maintain high-quality audio synthesis results when

generating raw audio waveforms. However, its autoregressive nature leads to slower inference. Despite the high quality of generation, it may be limited by computational resources in real-time applications (van den Oord et al., 2016). WaveGlow, on the other hand, combines the ideas of Glow and WaveNet to generate high-quality speech from mel spectrograms using a flow network. This approach avoids the computational bottleneck of autoregressive models, making the training and inference process more efficient (Prenger et al., 2019). However, despite the excellent performance of the model in terms of generation speed and quality, training still requires significant computational resources and the ability to generalize to different datasets has not been fully validated.

In the field of real-time speech synthesis, FastSpeech2 solves the latency problem of previous models by predicting phoneme duration and pitch. FastSpeech2 not only retains the high efficiency of FastSpeech, but also introduces the prediction of key speech features, such as pitch, duration, and energy, which greatly enhances the control of synthesized speech. This results in synthesized speech that is not only highly natural but also more accurately expresses different phonetic emotions and intonation variations, which is crucial for high-quality speech synthesis (Ren et al., 2020). For example, when synthesizing sentences with different phonological features, the model can accurately control pitch and duration, which is crucial for simulating different emotional states or emphasizing specific words (Shen et al., 2018). In addition, by controlling the energy, FastSpeech2 can adjust the intensity of speech, further increasing the dynamic range and infectiousness of speech expression.

In summary, the evolution from traditional HMM models to modern deep learning models such as Tacotron2 and FastSpeech2 highlights the ongoing efforts to improve the naturalness and expressiveness of synthetic speech. These advances are crucial to address our first research question of whether precisely modifying TextGrid files and their corresponding audio affects the learning process of speech synthesis models (Section 2.4). This historical perspective lays the foundation for evaluating how our proposed modifications can further enhance these models. Thus, FastSpeech2 was chosen for this study, both as a means of exploiting its advanced performance and as an anticipation of its future potential.

2.2 Control Techniques for Synthesis of Emotional Speech Features

Speech synthesis technology is a complex multi-stage process, which mainly includes the stage of preparing files, the stage of model training, and the stage of speech synthesis.

In the preparation of files phase, researchers provide basic data for subsequent model training by collecting a large amount of speech data and performing careful labeling work. For example, Kayte and his colleagues, in developing a text-to-speech (TTS) system for the Marathi language, used a variety of manual and automated methods for collecting and labeling speech data. They recorded important prosodic information such as phonemes, pitch, intensity, and duration, the accurate recording of which is crucial for subsequent model training (Kayte et al., 2015). In addition, an automated tool developed by Gibbon and Bachan based on the TextGrid feature of the Praat software greatly improves the efficiency and accuracy of speech data labeling. This tool allows subsequent model training to capture subtle changes in speech data more accurately through precise temporal labeling and text alignment (Gibbon & Bachan, 2008). By using TextGrid labeled data, researchers have been able to train a variety of speech synthesis models that rely on accurate speech annotation to learn the acoustic features and linguistic patterns of speech. For example, in the Festival and Festvox systems, TextGrid-labeled data was used to train models for synthesizing speech to produce natural and fluent speech (Kayte et al., 2015).

Moving on to the model training phase, the researchers used these accurately labeled speech data to train speech synthesis models. The main task of these models is to predict and control the acoustic and linguistic features of speech to generate natural and fluent speech output. On the MaryTTS platform, Steiner and Le Maguer detail how the temporal annotation and text alignment information in TextGrid files can be utilized to train highly accurate speech models. These models are capable of automatically adjusting features such as the intensity, loudness, and rate of speech according to the input text to generate speech with diverse linguistic features (Steiner & Le Maguer, 2017). In addition, Šimko et al. introduce a novel approach to enhancing the precision of speech synthesis through meticulous modification of TextGrid and lab files during the training phase, enabling the generation of audio with distinct linguistic features such as intonation, stress, and rhythm within a single sentence. Using the FastSpeech2 model, this approach allows dynamic control of these speech effects without requiring major changes to the existing model structure. This is a substantial improvement over traditional approaches that typically focus on the overall emotional state rather than subtle speech effect changes (Šimko et al., 2023).

In the speech synthesis stage, the trained model generates the corresponding speech output based on the input text and can adjust the speech features in real time according to different application scenarios. Steiner and Le Maguer further discuss the advantages of deep neural networks in this stage. They show how trained deep neural network models can be utilized to generate high-quality speech output by adjusting various features of speech, such as intensity, loudness, and speech rate, in real time according to different text inputs. These deep learning-based models, such as Tacotron2 and WaveNet, not only better capture and reproduce the complex features of speech, but also significantly improve the naturalness and smoothness of speech synthesis (Steiner & Le Maguer, 2017). In addition, Eyben et al. introduced the application of the openSMILE tool in speech synthesis, which can achieve high-precision control of speech features by extracting multimodal features of speech and video and combining them with a machine learning model. openSMILE's modular design allows it to be flexibly integrated into a variety of speech synthesis systems, which can effectively adjust and control speech features, improving the adaptability and scalability of the system (Eyben et al., 2013).

Nonetheless, while many existing platforms and software provide convenient speech effect tuning functions, these functions usually focus on tuning the overall speech effects only, and less on the ability to generate a mixture of different effects in the same synthesized speech. For example, EmoVoice, described by Vogt et al. is a speech emotion recognition and synthesis tool that allows users to create personalized emotion recognizers that can be used in real-time emotion classification. Although EmoVoice can adjust the emotional characteristics of speech in real-time, it is primarily geared towards overall emotion adjustment and does not support the generation of a mixture of speech with different emotional characteristics in the same speech (Vogt et al., 2008). Similarly, the WISE system developed by Eskimez et al., a web-based interactive speech emotion classification system, lacks support for blending different effects in a single synthesized speech, although it allows the user to upload speech data and automatically classify emotions, as well as adjusting emotion labels based on user feedback (Eskimez et al., 2016).

In summary, although current speech synthesis techniques have made significant progress in various aspects through advanced annotation, training, and synthesis methods, they have mainly focused on tuning overall speech effects and have not yet been able to effectively handle the hybrid generation of complex effects in the same synthesized speech. My research will start from the file preparation stage, by precisely modifying the TextGrid file and its corresponding audio files. This will explore whether these adjustments can significantly affect the speech effects learned by the model during the training stage and be able to accurately synthesize emotionally expressive target speech during the synthesis stage (Section 2.4). This research is expected to address the challenge of simultaneously capturing and expressing diverse speech effects in a single synthesized speech (Steiner & Le Maguer, 2017; Šimko et al., 2023). In addition, this study will validate the cross-linguistic universality and effectiveness of

this approach to extend it to speech synthesis practices in different languages (Section 2.4). Through these explorations, I hope to contribute to the creation of more expressive and characteristically diverse speech audio, particularly in terms of providing multilingual support and fine-grained emotional expression.

2.3 Application Areas of Speech Synthesis

With the rapid development of artificial intelligence and computer technology, speech synthesis technology has been widely used in various fields and has demonstrated significant value in business, education, and society. As an important part of human-computer interaction, speech synthesis technology not only makes intelligent voice assistants possible but also improves the convenience and efficiency of work and life in a variety of scenarios.

In the commercial field, speech synthesis technology is especially widely used. Firstly, in the field of smart homes and smart speakers, speech synthesis technology enables users to control home appliances, play music, set reminders, etc. through voice commands, thus greatly facilitating daily life. In addition, by generating natural and attractive voice advertisements, enterprises can more effectively convey information and attract potential customers. Another important application of speech synthesis technology in the commercial field is the customer service system. Through speech synthesis technology, enterprises can provide 24-hour uninterrupted customer service, significantly improving customer satisfaction and service efficiency. For example, in the hospitality industry, the application of voice assistants has improved customer service quality and operational efficiency (Hoy, 2018).

In the field of education, speech synthesis technology has also shown significant value. Speech synthesis technology can help students in lower grades improve their writing skills and interest in learning (Plummer & Beckman, 2016). Through voice feedback, students can more intuitively understand their pronunciation and grammatical errors, thus improving their learning efficiency. In addition, speech synthesis technology has been used to develop educational software and tools to provide students with a personalized learning experience.

Applications in the social field are equally noteworthy. Speech synthesis technology plays an important role in healthcare management. For example, when helping people with disabilities, speech synthesis technology can provide audiobooks for the visually impaired and real-time subtitles for the hearing impaired, thus greatly improving their quality of life (Hoy, 2018). In addition, speech synthesis technology is widely used in the intersection of AI and healthcare. For example, through the development of intelligent health voice assistants, patients can be helped to monitor their health status, manage medication use, and provide health advice (Buhalis et al., 2022). It is also possible to make communication more empathetic and humane by adjusting the speed and tone of voice to suit the patient's emotional state when providing health guidance or psychological support dialogue, enhancing patient comfort and satisfaction.

Although speech synthesis technology has shown great potential in many fields, it still faces some problems and challenges. The main problems are the monotony of

synthetic speech, the lack of emotional expression, and the inconsistency of speech speed. Usually, the speech generated by the speech synthesis system appears to be mechanical and indifferent, lacking the emotional fluctuations of natural language, and it is difficult to establish an emotional connection with users. In addition, intonation, emotion, and pitch variation in natural language are extremely complex, which remains a huge challenge for existing speech synthesis technologies. Currently, while most speech synthesis systems can generate neutral speech, they often lack emotional richness and variety. For example, it is difficult for synthetic speech to accurately express emotions such as happiness, sadness, or anger in human speech, which limits its ability to effectively convey emotions in interpersonal communication. In addition, speech synthesis systems have trouble simulating natural intonation and prosody. In natural languages, intonation and prosody changes are complex and highly uncertain, and existing technologies have not been able to effectively mimic this complexity. As a result, synthetic speech is often monotonous and mechanical, lacking the dynamics and expressiveness of natural speech. Finally, the complexity of spontaneous language requires speech synthesis systems to be able to process unstructured inputs and produce smooth and natural outputs, but current systems generally only perform well in limited and predefined contexts (Kuligowska et al., 2018).

In summary, speech synthesis technology has demonstrated significant value in a variety of fields, including business, education, and society, and is changing the way we live by enhancing the user experience, improving learning, and improving the quality of life. However, this technology still faces many challenges (Kuligowska et al., 2018; Hoy, 2018; Plummer & Beckman, 2016). Future research could explore how diverse speech effect synthesis, such as effects like emphasis and speech rate variation, can be used to further enhance interactivity in education, humanization in healthcare, and efficiency and satisfaction in customer service.

2.4 Research Question and Hypothesis

In recent years, with the development of artificial intelligence and deep learning technology, speech synthesis technology has also made significant progress. From early rule-based systems to end-to-end models utilizing deep neural networks, each technological innovation has greatly contributed to the improvement of speech synthesis quality. During this technological evolution, models such as FastSpeech2 have become a research hotspot due to their high efficiency and excellent performance. These models not only optimize the naturalness and fluency of speech but also enhance the expression of emotion and intonation.

In addition, the control techniques for speech synthesis features have been evolving. By adjusting TextGrid files and related audio data, researchers have successfully realized precise control of speech effects, which not only enhances the naturalness of speech but also permits synthesized speech to better adapt to different application scenarios and needs. For example, in the education field, by emphasizing keywords or phrases, students can be helped to better understand and memorize the learning content; in the business field, by generating attractive voice advertisements, companies can effectively convey information and attract potential customers. These applications demonstrate the extensive social value and commercial potential of speech synthesis technology.

Based on the in-depth analysis of the existing literature, I found that although the existing speech synthesis models have been improved in various aspects, there are still limitations in terms of specific speech effect control and cross-linguistic applications. Therefore, I propose the following two research questions and their corresponding hypotheses to further optimize speech synthesis techniques and explore their potential for application in different languages.

Research Question 1:

By modifying the TextGrid file and its corresponding audio in the training text, is it possible to influence the speech effects learned by the speech synthesis model in the training phase and generate audio files with different speech effects in the synthesis phase?

Hypothesis 1:

Accurate modification of TextGrid files and corresponding audio profiles can significantly influence the learning process of speech synthesis models in the training phase, enabling the models to accurately reproduce target speech effects in the synthesis phase. This hypothesis is based on the research of Steiner and Le Maguer

(2017), who noted that temporal annotation and text alignment information in TextGrid files is critical for training speech models. This approach enables the model to automatically adjust the intensity, loudness, and rate of speech to produce output with rich speech effects.

Research Question 2:

Is this method of adjusting speech synthesis characteristics by modifying TextGrid files and corresponding audio cross-linguistically universal and effective, and can it be successfully generalized to speech synthesis practices in other languages?

Hypothesis 2:

Based on the experimental validation of Chinese speech synthesis, this approach is expected to be equally applicable to other languages. The successful application of the FastSpeech2 model in a multilingual environment demonstrates the important role of tuning the training data through a high degree of control in improving the adaptability of synthesized speech. This generalizability and effectiveness are supported by the practice of FastSpeech2 modeling (Shen et al., 2018; Ren et al., 2020).

Through these research questions and hypotheses, I would like to explore new possibilities for speech synthesis techniques, with the expectation of contributing to the creation of more expressive speech audio with more diverse effects.

3 Methodology

Based on the research questions and hypotheses presented in the previous chapter (Section 2.4), I developed an innovative speech synthesis method specifically designed to generate audio with multiple effects. This method involves meticulous fine-tuning of the training files, including the addition of special symbols in front of the training text denoting specific speech effects (e.g., an asterisk (*) in front of the effect of repetition, and a percent sign (%) in front of the effect of slowing down the speech rate), and accordingly batch-tuning of the training audio to accurately control the different speech effects. The code involved in this paper is publicly available[1].

The reason I chose this method is that training files such as lab files and TextGrid files contain detailed parametric information about the speech such as pronunciation durations, timestamps of speech segments, phoneme sequences, intonation, and rhythm. These training files provide the feature extraction data required by the model during the training process of speech synthesis, and the model generates speech output based on these features. In contrast to the traditional method of training models (Section 2.2), this method changes the features learned by the model by modifying the training file, in effect changing the data used by the model to extract the features. This enables precise control over the effect of synthesizing different speech. In short, this method of modifying the training file to synthesize multiple effect audio is not only able to accurately adjust the speech parameters to achieve diversified speech effects, but also the method itself is simple and efficient, highly operable, and innovative, and can also be regarded as a kind of fine-tuning and optimization of the model's training parameters.

The experiments were implemented in several stages. Firstly, it started with the selection of datasets. To ensure the replicability of the study, publicly available datasets in three different languages were selected: the Baker dataset in Chinese, the LJSpeech dataset in English, and the CSS10 dataset in Spanish. These datasets are widely recognized in the speech synthesis field (Section 3.1).

After downloading these publicly available datasets, preliminary experiments were conducted to determine the appropriate number of training samples and target speech effects. Ultimately, 7500 audio samples from each language's dataset were selected for training. Corresponding lab files and TextGrid files were created for these audio samples, and repetition and speech slowdown were selected as the speech effects tested (Section 3.2).

Then, the FastSpeech2 speech synthesis model was used for training. The validity of the hypothesis was verified by observing the log files to determine the training

[1] https://github.com/weiyilan9/master_thesis

checkpoints for early stops, and synthesizing audio samples demonstrating different speech characteristics (Section 3.3).

3.1 Corpus

Baker[2] was chosen for the Chinese corpus, a publicly available Chinese Mandarin Female Corpus designed for non-commercial use. The database contains 10,000 sentences recorded by young women in standard Mandarin. The voices are clear, conveying warmth and a positive sense of feeling, greatly enhancing the naturalness and expressiveness of the speech. These recordings were made in a professional environment, using high-standard recording equipment to ensure that the signal-to-noise ratio of the recording files is no less than 35dB, with a sampling rate of 48KHz and 16-bit format. These technical parameters ensure the high quality and reliability of the recordings and provide a solid foundation for the research of speech synthesis technology. The corpus of the database is designed to cover a wide range of data types such as news, technology, entertainment, and conversations to ensure the diversity and usefulness of the data. The database not only contains comprehensive information on syllable consonants, tones, and rhymes but also provides detailed labeling of acoustic-rhythmic boundaries, which is crucial for accurate speech synthesis.

LJSpeech[3] was chosen for the English corpus. this was based on its wide range of applications and proven utility in the field of speech synthesis and recognition. the LJSpeech database was preferred due to its high quality, wide range of applications, and proven validity in several research papers. This database consists of public-domain audio containing a total of 13,100 short audio clips of a female pronouncer reading seven public-domain non-fiction books. The audio and corresponding text transcriptions cover a diverse range of textual material, ensuring the richness and variety of the corpus. The audio files are in single-channel 16-bit PCM WAV format with a sampling rate of 22,050 Hz, and the technical specifications ensure audio clarity and the ability to adapt to multiple processing techniques. The length of each audio clip varies from 1 to 10 seconds and contains about 17.23 words on average, making it suitable for the rapid processing of short sentences and long passage speech synthesis studies. The audio segments are automatically segmented based on silence and are usually aligned to sentence or phrase boundaries, facilitating the training of more accurate speech recognition and synthesis models. The LJSpeech database also provides detailed metadata files, including the identification of each audio file, the original transcription, and the normalized transcription, the latter of which provides a full lexical unfolding of numerals, ordinal numbers, and monetary units, among

[2] https://www.data-baker.com/en/datasets/freeDatasets/
[3] https://keithito.com/LJ-Speech-Dataset/

others. These features make LJSpeech suitable not only for basic training in speech synthesis but also for facilitating complex natural language processing tasks.

CSS10[4] was chosen for the Spanish corpus. The Spanish portion of the CSS10 database excels in automated speech synthesis applications due to its quality and consistency. CSS10 was developed by Kyubyong Park and Thomas Mulc to provide a single pronouncer speech dataset for use in speech synthesis, covering ten different languages, including Spanish. CSS10 Spanish dataset consists of audio clips of audiobooks and their aligned texts from LibriVox, which are public domain resources. The database is an invaluable resource for research because of the precise alignment of audio and text, its extensiveness, and its high quality. Recordings were done in a tightly controlled environment, and all recordings were sampled at 22kHz to ensure the clarity and usability of the audio data. CSS10 focuses on text processing and accurate alignment, as well as normalization of the text, such as expanding acronyms and transcribing numerals, which is especially critical for text-to-speech conversion. CSS10's processing of Spanish audio includes the use of Audacity audio editing software to automate the process of finding segmentation points and adjusting them to ensure the appropriate length of audio segments, which is important for accurate text-to-speech synthesis. The dataset provides high-quality speech samples and demonstrates its effectiveness by training and evaluating it on two well-known neural network text-to-speech models, Tacotron and DCTTS. Through the testing of these models and the Mean Opinion Score (MOS) scores, CSS10 demonstrates its excellent performance in natural speech synthesis, especially its high-performance scores in speech naturalness and pronunciation accuracy.

Overall, the databases for all three languages have single pronouncers and high-quality recordings with stringent requirements in processing and text alignment, and are of excellent quality, making them ideal for researching speech synthesis techniques.

[4] https://www.kaggle.com/datasets/bryanpark/spanish-single-speaker-speech-dataset

3.2 Data Preprocessing

3.2.1 Test Experiments

Before conducting formal speech synthesis experiments, I designed two test experiments to initially verify the hypothesis and improve the training method. The following is the detailed procedure of the test experiments.

Experiment 1: Manual Labeling Experiment

Firstly, a small-scale Chinese experiment was conducted. To investigate whether the model can learn the effect of weighting specific audio segments, I randomly selected 100 Chinese audios from the Baker corpus as samples, and manually performed loudness tuning operations on random parts of each audio in Audition, specifically, turning up the volume of the selected audio segments by about 5 dB. When processing these audios, I also modified the corresponding TextGrid file and the lab file to add an asterisk (*) as an identifier before the segment whose volume was turned up. Then, I selected 1000 Chinese audio training without any modification together. A total of 1,100 audios were used to help the model learn more diverse features.

After about 30,000 training steps, the results showed that the waveforms of the synthesized audio were significantly larger in the portion of the synthesized audio where the asterisks were added, proving that the model can learn the effect of audio exacerbation. However, this manual labeling experiment has obvious shortcomings, such as manually labeling the data is very time-consuming and error-prone, and does not apply to large-scale audio processing. In addition, due to the inconsistency of the audio volume in the dataset, the volume in a sentence may fluctuate from high to low, which also affects the effect of the experiment.

Experiment 2: Batch Processing Experiment

To address the problems of insufficient amount of audio and inconsistent volume in the manual labeling experiment, I designed a second experiment.

Firstly, increase the number of training audios. I used the format of one group of every 2,500 audios to increase the number of Chinese audios to 5,000. The reason for choosing a group of 2,500 audio samples is to ensure that there is enough data diversity for each audio effect while avoiding excessive computational burden. This number is based on preliminary small-scale experiments and my own experience in training speech synthesis models, and it can balance the training time and

performance of the model, and achieve a better cost-benefit ratio between computational resources and training time.

Second, by using Audition's batch processing function, the audio is divided into two parts for loudness normalization: the first 2,500 audio is uniformly adjusted to -10 dB to simulate the enhancement of the voice during emphasis, and the other 2,500 audio is uniformly adjusted to -18 dB to simulate the standard loudness in daily conversations, to ensure that the results of the speech synthesis are more closely related to the volume of the actual human language communication. The batch processing approach is designed to improve the efficiency and consistency of data processing. By setting a uniform loudness standard in Audition, the inconsistency and errors of manual processing can be avoided, ensuring that each sample is trained under the same conditions, thus improving the accuracy and replicability of model learning. The loudness settings are based on the ITU 1770 standard developed by the International Telecommunication Union (ITU) and Spotify's practical experience. The ITU 1770 standard is designed to address the problem of inconsistent volume in radio, TV programs, and other multimedia platforms, and to improve the user experience by achieving volume consistency across different sources of audio through loudness regularization techniques. This standard is widely accepted internationally and applied to audio production and distribution, ensuring that listeners receive a balanced and consistent listening experience in a variety of playback environments[5]. In addition, I also refer to the application of this standard by Spotify, which adjusts audio tracks through loudness regularization techniques to make the volume more consistent throughout the playlist, further improving the standardized expression of different sound effects[6].

The experiment was trained over about 150,000 steps and showed positive results. In the synthesized audio tested, it was possible to synthesize audio with an emphasis section. This proves that the batch processing method is effective and solves the time-consuming and inaccurate problems of manual labeling. Notably, the experiment also illustrated that more training steps are not better. Observations can be made of the training log files to determine when to stop early. For example, when training to about 90,000 steps, the model's loss value starts to float up and down and cannot be significantly reduced any further, which indicates that 90,000 steps of training is sufficient to validate hypothesis 1 for this research problem.

[5] https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1770-5-202311-I!!PDF-E.pdf

[6] https://support.spotify.com/us/artists/article/loudness-normalization/#_gl=1*yeym0k*_gcl_au*MTk3NzIxNjM0Mi4xNzE2NDkzNjYw

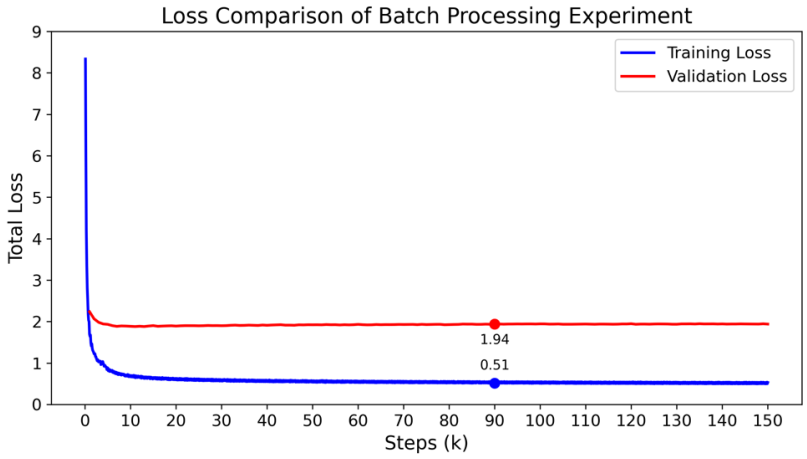

Figure 1: Loss Comparison of Batch Processing Experiment

Altogether, although the first test experiment proved that the method of increasing volume is effective, it is not suitable for large-scale application due to the complexity and instability of manual labeling. The second test experiment successfully verified that the same effect can be applied in large-scale audio processing through the batch processing method. This proves that the batch volume adjustment method is not only effective but also replicable. Through these two experiments, I have initially verified the feasibility of the method and laid the foundation for subsequent formal experiments. Next, based on the results of these two experiments, I will conduct further formal experiments on speech synthesis in Chinese, English, and Spanish to test hypothesis 2.

3.2.2 Formal Experiments

In the formal experiments, the processing can be divided into several parts.

Firstly, the pre-processing of the audio files. To verify whether this method can synthesize many different speech features, I chose the sounds with more distinctive features for the verification of the emphasis effect and the speech slowing effect. For each language (Chinese, English, and Spanish), 7,500 audios were selected, and every 2,500 audios were used as a sample of one effect, which was categorized into emphasis, slowing down, and normal audios, and batch processed with Audition. The specific operation is as follows: for each language, I divided the 7,500 audio files into three groups of 2,500 audio each. The first group of 2,500 audio files is adjusted to a uniform loudness of -10dB to simulate the effect of voice overdubbing. The second group of 2,500 audio was adjusted to a uniform loudness of -18dB to maintain the volume of normal speech. The third set of 2,500 audios was processed in two steps: firstly, the loudness was uniformly adjusted to -10dB, and then the speech rate was slowed down by a factor of 1.5 to simulate the effect of slower speech. The batch processing ensured that each language had three audio groups with clearly differentiated speech effects.

The next step is the processing of the lab files. The corpus for each language comes with a document that records the name of each audio file and its corresponding audio text. I generated the corresponding lab file for each audio by splitting this document and utilizing Python. This was handled as follows: for the audio portions that were reread, an asterisk (*) was added in front of each word in the lab file as an identifier; for the audio portions that were slowed down, a percentage symbol (%) was added in front of each word in the lab file as an identifier. The lab files for all languages were processed in this way to clearly distinguish between audio with different processing effects in subsequent experiments.

Different strategies were used to process the TextGrid files for the three languages. For the Chinese Baker corpus, its pre-equipped TextGrid files were utilized. Specific modifications included: adding an asterisk (*) marker after the “text=” tag in the emphasis section; and for the slowed-down section, not only adding a percentage symbol (%) after “text=”, but also adjusting the “xmin” and “xmax” tags of the timeline to extend their duration to 1.5 times and retaining them up to three decimals, to match the extended speech rate after the audio processing. To ensure the accuracy of these modifications, an alignment test was performed via Praat software, which showed good alignment. For the English and Spanish corpus, due to the lack of pre-aligned TextGrid files, the MFA (Montreal Forced Aligner) tool was used to uniformly generate TextGrid files. The file alignment check was performed before processing, and the TextGrid generation and annotation were performed after confirming that there were no errors. For the parts of emphasis and slowing down, the processing method is the same as that for Chinese. In all three languages, symbols indicating

pauses, such as “sp”, “spn”, “sil”, and “<eps>”, were not specially processed or tagged. Similarly, no special symbols were added to blank paragraphs.

The model I chose for speech synthesis was FastSpeech2, which required more files to be prepared before training because FastSpeech2 was designed to be remarkably different from other models such as Tacotron2 and the original FastSpeech. FastSpeech2 introduces an advanced attentional mechanism, which allows the model to learn and predict key audio features, including pitch, duration, mel spectrum, and energy more accurately. In particular, pitch reflects the tonal variation of articulation and is crucial for conveying the semantics and intonation of different languages and dialects; duration relates to the length of articulation of a phoneme or a word, and its precise control is the basis for achieving smooth and natural speech; mel spectrum provides an effective representation of audio frequency by simulating the auditory perception of the human ear, which directly affects the quality and clarity of speech; the energy feature depicts the strength and dynamic range of the speech signal, which is extremely important for reproducing the speaker's emotion and emphasis.

For this reason, I used the processing files provided by the FastSpeech2 model to synthesize the corresponding pitch, duration, mel spectrum, and energy files. In addition, to validate the model, I set 150 audio samples to constitute the validation set. I also added and updated symbols that appeared in the training and validation sets and were missing in the symbols file to ensure the completeness of the symbol table and the accuracy of the model training. This process was consistent across all three different languages.

3.3 Experimental Configurations

Before starting the training, I configured a detailed set of parameters for the experiment, which were largely consistent with the settings in the original FastSpeech2 paper[7]. Below are the specific configuration details of the experiment.

For the **transformer** configuration, the encoder was set to 4 layers with 2 heads per layer and a hidden layer dimension of 256; the decoder was set to 6 layers, also with 2 heads per layer and a hidden layer dimension of 256. The filter size of the convolutional layer was set to 1024 and the kernel size was set to [9, 1]. The dropout rate is set to 0.2 for both the encoder and decoder. In **variance predictor** configuration, the filter size is 256, the kernel size is 3, and the dropout rate is 0.5. This helps in predicting pitch and energy variations. For **variance embedding**, the quantization of pitch and energy is set to "linear" with 256 quantization intervals, which is consistent with the characteristics of the data without normalization. In the **multi-speaker** configuration, the multi-speaker feature is not enabled because the datasets of the three languages are all single speakers to keep the model simple and focused. The **maximum sequence length** is 1000 to accommodate longer speech input. The **vocoder** uses the "Hi-Fi-GAN" model, which is one of the most widely recognized vocoders in the field of high-quality speech synthesis. Finally, for the **optimizer**, the batch size is set to 128, the optimizer's momentum (betas) is set to [0.9, 0.98], the eps is 0.000000001, and no weight decay is used (weight_decay is 0.0). The gradient clipping threshold is 1.0 and the gradient accumulation step (grad_acc_step) is 1. The warm-up step (warm_up_step) is set to 4000, the decay steps (anneal_steps) are [300000, 400000, 500000], and the decay rate (anneal_rate) is 0.3.

This set of experimental configurations aims to simulate and validate the effect of the FastSpeech2 model in different speech synthesis scenarios, and to ensure that the model's performance is consistent with the original paper through fine parameter adjustments.

In addition to the above experimental configurations, to further monitor and manage the training process of the model, I also set detailed step configurations to ensure data visualization and continuous performance evaluation during the training process. The specific **step** configuration is as follows:

The total step is set to 300,000. The log step is set to record a training log every 100 steps to allow real-time monitoring of loss values and other important metrics during training. The synthesis step and validation step are both set to 1,000 steps, which means that the generation of synthetic samples and performance evaluation on the validation set is performed every 1,000 steps to keep up to date with the model's performance on unseen data. The save step is set to save the model every

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

10,000 steps to ensure that there are enough checkpoints in the training process that can be used for subsequent recovery training or performance comparison.

In addition, I made my own decision on when to stop training early based on the losses observed in the log files during training. This strategy aims to prevent over-training and over-fitting and ensures that the model stops training promptly after reaching the desired performance, thus preserving the optimal model for subsequent testing and deployment. This flexible training management strategy helps optimize the overall performance of the model while saving unnecessary computational resources.

All training was done on Hábrók, using NVIDIA A100 GPUs for training.

4 Results & Discussion

To evaluate the performance of the speech synthesis model more comprehensively, I used both numerical analysis (Section 4.1) and manual listening assessment (Section 4.2) for a comprehensive evaluation.

For numerical analysis, I used Audition to examine the duration and loudness of three speech effects: normal, repetition, and slowed speech. This provided objective data support and allowed me to accurately measure the technical metrics of the different speech expressions.

For artificial hearing assessment, I constructed mixed speech samples containing features of normal speech rate, emphasis, and slowing down, and designed questionnaires to invite listeners to participate and test whether they could accurately recognize the speed change and emphasis parts of the speech samples. This approach not only provides subjective feedback from listeners but also reveals the effects and potential problems of speech synthesis in practical use.

By observing the training logs, I selected the outputs of the speech synthesis models in Chinese, English, and Spanish at 90,000, 60,000, and 70,000 steps of training as evaluation checkpoints. Based on these checkpoints, I used Tatoeba [8] to generate 10 semantically consistent test sentences, thus maintaining the consistency of the test conditions. These sentences are listed below:

	Chinese	Chinese (pinyin)
1	他看起来像个运动员，但是其实是个作家	ta1 kan4 qi3 lai2 xiang4 ge4 yun4 dong4 yuan2 sp dan4 shi4 qi2 shi2 shi4 ge5 zuo4 jia1
2	今天下午我有两个小时的英语课和两个小时的汉语课	jin1 tian1 xia4 wu3 wo3 you3 liang3 ge5 xiao3 shi2 de5 ying1 yu3 ke4 he2 liang3 ge4 xiao3 shi2 de5 han4 yu3 ke4
3	如果可能的话我想去世界各地旅行	ru2 guo3 ke3 neng2 de5 hua4 wo3 xiang3 qu4 shi4 jie4 ge4 di4 lv3 xing2

[8] https://tatoeba.org/zh-cn

4	当我们决定期待从生活中得到什么的时候，生活开始了	dang1 wo3 men2 jue2 ding4 qi1 dai4 cong2 sheng1 huo2 zhong1 de2 dao4 shen2 me5 de5 shi2 hou5 sp sheng1 huo2 kai1 shi3 le5
5	这个童话故事很浅白，七岁的小孩也看得懂	zhe4 ge5 tong2 hua4 gu4 shi4 hen3 qian3 bai2 sp qi1 sui4 de5 xiao3 hai2 ye3 kan4 de5 dong3
6	我宁愿呆在家里也不要在这种天气中出门	wo3 ning4 yuan4 dai1 zai4 jia1 li3 ye3 bu2 yao4 zai4 zhe4 zhong3 tian1 qi4 zhong1 chu1 men2
7	众所周知，空气是多种气体的混合体	zhong4 suo3 zhou1 zhi1 sp kong1 qi4 shi4 duo1 zhong3 qi4 ti3 de5 hun4 he2 ti3
8	我们有一只猫。我们都喜欢这只猫	wo3 men2 you3 yi4 zhi1 mao1 sp wo3 men2 dou1 xi3 huan1 zhe4 zhi1 mao1
9	谢谢你让我度过一个愉快的晚上	xie4 xie4 ni3 rang4 wo3 du4 guo4 yi2 ge5 yu2 kuai4 de5 wan3 shang3
10	寒冷干燥，灿烂的阳光，多么美丽的冬日天气	han2 leng3 gan1 zao4 sp can4 lan4 de5 yang2 guang1 , duo1 me5 mei3 li4 de5 dong1 ri4 tian1 qi4

Table 1: Test Sentences in Chinese

	English	**Spanish**
1	He looks like a sportsman, but he is a writer	Parece un deportista, pero es escritor
2	This afternoon I have English class for two hours and then two hours of Chinese	Esta tarde tengo dos horas de clase de inglés y dos horas de clase de chino
3	I want to go on a journey around the world if possible	Quiero ir a un viaje alrededor del mundo, si es posible
4	Life starts when you decide what you are expecting from it	La vida empieza cuando decides lo que esperas de ella
5	This fairy tale is easy enough for a seven year old child to read	Este cuento de hadas es bastante simple, un niño de siete años puede leerlo

6	I'd rather stay home than go out in this weather	Prefiero quedarme en casa que salir con este tiempo
7	As everyone knows, air is a mixture of gases	Como todo el mundo sabe, el aire es una mezcla de gases
8	We have a cat. We are all fond of it	Tenemos un gato. A todos nos gusta el gato
9	Thank you for the pleasant evening	Gracias por una noche agradable
10	Cold and dry, splendid sunshine, what beautiful winter weather	Frío y seco, una esplendida luz del sol, qué hermoso clima invernal

Table 2: Test Sentences in English and Spanish

4.1 Numerical Analysis

Firstly, I selected the outputs of the Chinese, English, and Spanish speech synthesis models at 90,000, 60,000 and 70,000 steps of training as the evaluation checkpoints, and generated the effects of normal, emphasis and slowing down the speech rate by 1.5 times for ten test sentences, respectively. Then, the audio was scanned for duration and loudness using Audition. Finally, the average of the ten test sentences was calculated.

Average Duration (s)			
Feature	Chinese	English	Spanish
Normal	3.382	3.405	2.896
Emphasis	3.634	3.283	2.865
Slow 1.5x	5.457	5.272	4.562

Table 3: Average Duration for the Test Sentences

It is clear from Table 3 that the audio duration of slowing down the speech rate by 1.5 times is about 1.5 times longer than the normal effect and the emphasis effect. This indicates that the model learned the speech features of the slowed-down speech rate very well. In contrast, there is no clear pattern in the change in audio duration for the emphasis part.

Average Loudness (dB)			
Feature	Chinese	English	Spanish
Normal	-18.011	-18.741	-19.088
Emphasis	-12.773	-13.763	-13.699
Slow 1.5x	-12.597	-12.846	-12.735

Table 4: Average Loudness for the Test Sentences

Similarly, it is evident from Table 4 that the audio decibel values (dB) of the emphasis and the slowed-down 1.5 times speech speed were improved by about 6 dB compared to the normal speech speed. This result indicates that the model effectively modeled the loudness characteristics of these speech effects. However, it is important to note that despite the loudness improvements, they are still approximately 2 dB lower than the corresponding average decibel value for emphasis in the training set (-10 dB). This may be indicative of room for improvement of the model in terms of loudness amplification.

To further verify the statistical significance of these results, I performed independent sample t-tests on the duration and loudness of different speech effects (normal, emphasis, slowed down by 1.5x) in various languages. This test is used to compare the difference between the means of two independent groups of samples to determine whether the observed effect is statistically significant, and the statistical significance of the result is usually judged by a p-value of less than 0.05 (5% level of significance).

Statistical Comparison of Loudness			
	Chinese	English	Spanish
Normal & Emphasis	<0.001	<0.001	<0.001
Normal & Slow 1.5x	<0.001	<0.001	<0.001
Emphasis & Slow 1.5x	0.535	0.022	0.23

Table 5: Statistical Comparison of Loudness for Different Speech Effects

As can be seen in Table 5, for all three languages, there are significant differences between normal speech speed and emphasis and between normal speech speed and slowing down, suggesting that there is a significant distinction between the loudness of the models in modeling these different speech effects. The difference in loudness between emphasis and slowing down is not significant in Chinese and Spanish, suggesting that in these two languages the two speech effects behave similarly in terms of loudness, while in English, although statistically significant, the actual effect is likely to be smaller. Therefore, from the perspective of numerical analysis, despite the differences with the original training data, the model has learned and reproduced

the loudness of different speech features quite accurately, showing its ability to capture subtle differences in speech features.

In summary, I used numerical evaluation to evaluate the duration and loudness of three languages when synthesizing different speech effects. The results show that the model effectively learns and simulates the loudness and duration characteristics of different speech effects, especially when slowing down the speech rate. Using independent samples t-tests, I confirmed that the model's changes in loudness between different speech effects were statistically significant, although the difference in loudness between repetition and slowing down was not significant in some languages. This suggests that the model has had some success in modeling specific speech features, but there is still room for improvement in some details.

4.2 Manual Listening Evaluation

To evaluate the effectiveness of speech synthesis techniques more fully, I also conducted an artificial hearing assessment. Two main types of evaluation were used: an evaluation of the effects of different speech features and a Mean Opinion Score (MOS) test.

The choice of these methods was based on the following considerations:

Firstly, the evaluation of the effects of speech features focused on analyzing the performance of the speech synthesis model when simulating different speech variations (e.g., emphasis and speech rate adjustment). Through these effect-specific tests, the study aims to gain insight into the model's ability to mimic real human speech variations and its limitations. This evaluation was carried out by setting up experimental and control groups and collecting participant feedback in the form of a questionnaire to determine whether they could accurately recognize specific effects in synthetic speech.

Secondly, the MOS test focuses on evaluating the overall quality of the synthesized speech, especially its naturalness and clarity. As a widely used criterion in the field of speech synthesis, MOS evaluates the auditory quality of speech by inviting listeners to compare synthesized speech with real human voices. This evaluation method is effective in revealing the performance of synthesized speech in real-world application scenarios, especially its suitability and acceptability in multilingual environments.

To collect participants' feedback, I designed and released a questionnaire[9] through the Qualtrics platform, which received a total of 40 valid responses. The subjects who participated in the experiment all had a bachelor's degree or higher, were of Chinese nationality, and possessed good language skills in Chinese and English, which provided a solid foundation for evaluating the effects of speech. Considering the participants' unfamiliarity with Spanish, the Spanish parts of the test of speech effects were removed and only Chinese and English were included. The experiment was designed to contain an experimental and a control group, with the audio samples from the experimental group combining the Chinese and English emphasis effects, the speech slowing effect, and their combinations, while the audio samples from the control group did not contain these specific effects. Participants were asked to identify and select audio clips that they thought had a specific speech effect. In addition, to assess the naturalness and clarity of the synthesized speech, this study also conducted a MOS test, which covered Chinese, English, and Spanish, more fully in which participants were invited to rate the naturalness and clarity of the synthesized speech in comparison to the real human voice.

[9] https://github.com/weiyilan9/master_thesis/blob/main/questionnaire_demo/Questionnaire.pdf

Although the questionnaires in this study did not collect personal information from the participants, there are still some ethical issues that need to be considered in the process of data collection, management, and storage. Firstly, the questionnaire did not capture any personally identifiable information, avoiding as much as possible the risk of participants' privacy being compromised. All data were collected anonymously and used only for this study and will not be disclosed to third parties. In addition, participants were given full informed consent before their participation and understood the purpose and content of the study. Finally, during data analysis, all results were presented in such a way as to ensure that they did not reflect individual-specific information, thus avoiding any potential adverse effects on participants. These measures ensured the ethical compliance of the study and the protection of participants' rights and interests.

The statistical results and analyses are presented below:

Correct Rate					
	Language	Emphasis	Emphasis (mix)	Slow	Slow (mix)
Experimental Group	Chinese	0.4	0.675	0.8	0.85
	English	0.925	0.475	0.9	0.95
Control Group	Chinese	0.475	0.325	0.075	0.175
	English	0.3	0.275	0	0.625

Table 6: Correctness of Different Speech Effects

Statistical Significance Comparison				
	Emphasis	Emphasis (mix)	Slow	Slow (mix)
Chinese	0.499	0.002	<0.001	<0.001
English	<0.001	0.065	<0.001	<0.001

Table 7: Comparison of Statistical Significance Between the Experimental Group and the Control Group on Different Speech Effects

We can see in Table 6 a comparison of the correct rate between the experimental and control groups for different language and phonological effects. In the experimental group, for the Chinese emphasis effect, the correct rate is 40%, while for English, this correct rate increases significantly to 92.5%. This indicates that the English speech synthesis model performed more accurately in simulating the emphasis effect. For the mixed emphasis effect, the correct rates were 67.5% and 47.5% for Chinese and English, respectively, implying that Chinese is relatively more accurate in handling mixed speech features. For the slowing down effect, the correct rate was 80% for Chinese and 90% for English, showing that the speech synthesis models of both languages are better able to learn and simulate speech rate changes. Mixed with the

slowing down effect test, Chinese and English had 85% and 95% correct rates respectively, showing that the mixed effect was more effectively modeled in English.

In statistics, p-values are used to determine the significance of results. Typically, results are considered statistically significant if the p-value is less than 0.05 (i.e., a 5% level of significance). This means that the observed effect or difference is highly unlikely to be caused by random factors and reflects a real, systematic effect or difference. To verify the significance of the experimental results, I statistically analyzed the results in Table 5.

As can be seen from the comparison of statistical significance in Table 7, for the emphasis effect, English showed a very high statistical significance for the single emphasis effect ($p < 0.001$), whereas Chinese's performance was only significant for the emphasis mixture effect ($p = 0.002$), and not significant for the single emphasis effect ($p = 0.499$). For the slowing down effect, both for the single and mixed effects, the p-values for both Chinese and English were less than 0.001, showing strong statistical significance, which suggests that the slowing down simulation was well learned and applied in both languages.

In summary, these results reveal differences in the performance of different languages and speech effects in speech synthesis modeling. The speech-slowing effect is particularly strong in Chinese and English with remarkable success, which is a very promising result. The difference in the emphasis effect was not significant in Chinese but showed some potential in English. This difference may be related to the fact that the synthetic loudness variations of the reread speech mentioned in Section 4.1.1 failed to reach the loudness frequencies of the training set audio. However, I believe that more discriminating results can be obtained by increasing the audio loudness adjustment of the training set for the emphasis effect in the future.

Overall, the experimental results demonstrate that this method of synthesizing multiple speech features can be applied to different languages and validate hypotheses one and two (Section 2.4). These findings provide valuable insights for further optimization of speech synthesis methods.

MOS Scores			
	Language	Naturalness	Clarity
Synthetic Speech	Chinese	3.5	4.575
	English	3.875	4.35
	Spanish	4.125	4
Ground Truth	Chinese	4.35	4.825
	English	4.55	4.775
	Spanish	4.4	4.375

Table 8: Comparison of MOS Scores for Naturalness and Clarity in Three Languages

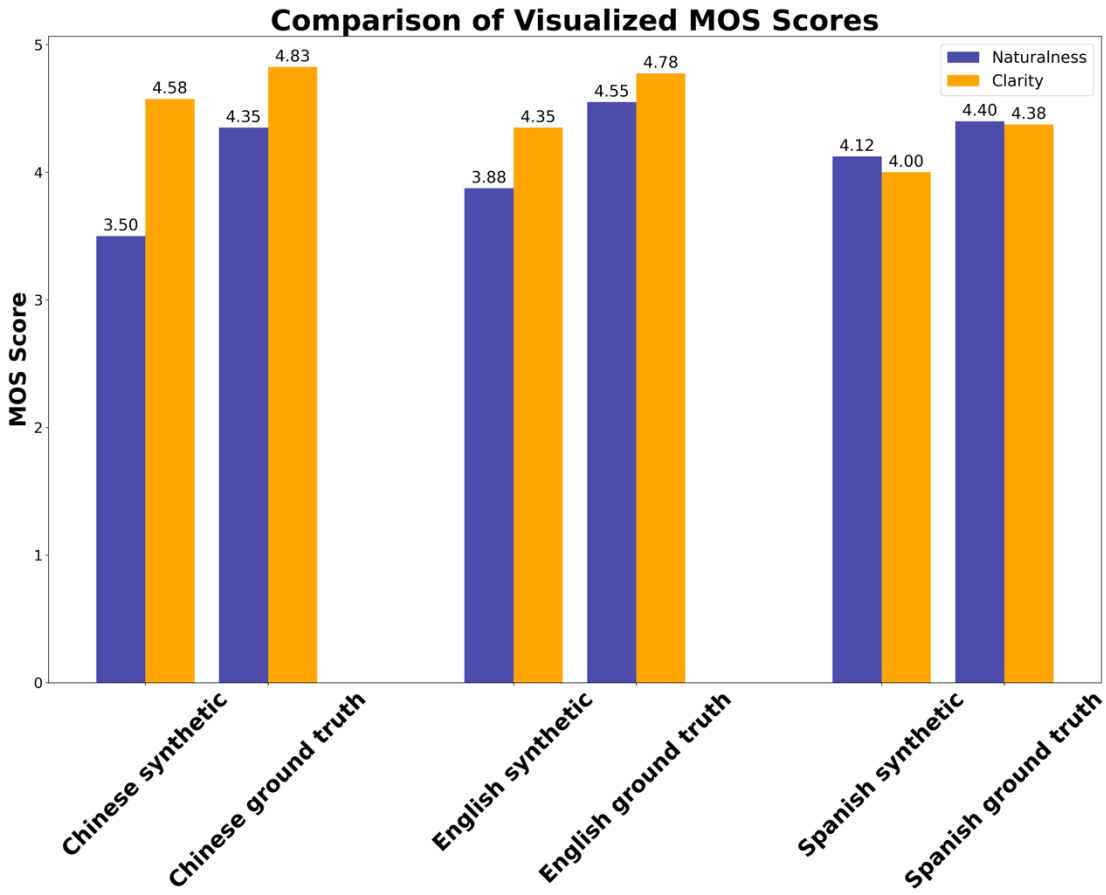

Figure 2: Comparison of Visualized MOS Scores

The Mean Opinion Score (MOS) is a commonly used method for evaluating speech quality and is particularly widely used in assessing speech communication systems and speech synthesis techniques. It provides listeners with a scale ranging from 1 to 5 to measure the quality of the speech samples heard. In the field of speech synthesis, the MOS score is commonly used to assess the naturalness and intelligibility of synthesized speech. Naturalness measures how smooth and natural the speech sounds, while clarity focuses on the noise impact and intelligibility of the speech.

The reason for choosing naturalness and clarity as scoring criteria Is that these two attributes do not require the listener to have an in-depth understanding of the linguistic content. This is particularly applicable to assessing non-native listeners' ratings of Spanish because although participants may not understand the exact meaning of the Spanish language, they are still able to rate the fluency of speech and the clarity of sound.

I visualized the data from the MOS scores. As can be visualized from the visualization charts in Figure 2, the clarity ratings of the synthetic speech in all languages were close to the ratings of the real speech, although there were small gaps. For example, the clarity score of synthetic speech for Chinese is 4.575 compared to 4.825 for real

speech, and the clarity score of synthetic speech for Spanish is 4 compared to 4.375 for real speech. However, the naturalness scores perform differently across languages. While there is a significant gap between the synthesized and real speech in Chinese and English, the naturalness performance in Spanish is closer, pointing out that the experimental speech synthesis models and methods have some room for improvement in the future.

In summary, the results of the MOS scores are relatively encouraging. Although there is still room for further improvement, this speech synthesis technique has been able to produce results that are very close to real speech in multiple languages. Especially in terms of clarity, the performance of the synthesized speech is very close to that of real speech, showing the potential of the technology.

Overall, in the research presented in this paper, I provide an in-depth performance evaluation of speech synthesis models in Chinese, English, and Spanish through careful numerical analyses (Section 4.1) and manual listening evaluations (Section 4.2). Test sentences generated through selected training checkpoints and subsequent evaluations show that these models exhibit significant differences and potential for simulating different speech rates and emphasis effects. Especially in the simulation of slowed-down speech rate and mixed effects, each language model demonstrated good learning ability and usefulness.

In addition, the Mean Opinion Score (MOS) test results further validate the naturalness and clarity of the synthesized speech, especially in different languages. Although synthetic speech still needs to be optimized in some aspects, overall, it is close to the real human voice, proving the effectiveness and applicability of the method.

5 Conclusion

This study explores the synthesis of a variety of speech effects (e.g., emphasis, slowing down, etc.) by fine-tuning the training files and corresponding audio data based on the FastSpeech2 model. Experiments conducted in three languages, Chinese, English, and Spanish, show that the methodology is feasible. Precise data annotation and modification can significantly improve the ability of the model to reproduce the target speech effects. The publicly available code also allows for good replicability of the research, providing rich experimental material for future researchers.

The results validate hypothesis 1, which states that by accurately modifying files such as TextGrid and the corresponding audio data, the learning process of the speech synthesis model in the training phase can be significantly affected so that the model can accurately reproduce the target speech effect in the synthesis phase. The synthesized speech has clarity and naturalness close to the real speech. Numerical analyses and manual listening evaluation results show that the model can successfully simulate the effect of speech slowing down in different languages, displaying significant feature variations in terms of duration and loudness, which proves the model's sensitivity and accuracy in adjusting the speech rate. In addition, although the performance of the emphasis effect varies across languages, it is generally possible for the model to learn and reproduce the emphasis effect, especially in English. The relatively insignificant emphasis effect in Chinese may be related to the labeling and processing methods of the training data.

The results also validate hypothesis 2, that is, the proposed method has cross-language generalizability and effectiveness. Through experiments in Chinese, English, and Spanish, I demonstrated the applicability of the method in multilingual environments, which implies that this technique can be generalized for speech synthesis in more languages with a wide range of potential applications.

Despite the encouraging results of the study, there is still room for improvement and potential for further research. Future research could further optimize the synthetic performance of the emphasis effect by varying the loudness of the training audio for the emphasis effect, especially in Chinese. In addition, the comparison of this method between different models can be explored in the future. For example, comparing FastSpeech2 with other state-of-the-art models to evaluate the differences in their performance on different speech effects. Further research can also be extended to the field of affective speech synthesis to explore methods of incorporating emotional features in synthesizing different speech effects so that the synthesized audio can not only have a variety of speech effects but also reach a higher level of emotional expression.

The value of the findings suggests that this method can be used to adjust speech effects with relative freedom, simulating natural and flexible variations in speech rate, intonation, and voice intensity, thus more closely resembling the natural speech habits of human beings, with a wide range of applications. Through these improvements and extensions, future research can further enhance the application value of multi-effect speech synthesis technology, promote the development of intelligent interaction technology, and benefit more application areas such as education, healthcare, and business.

In conclusion, this study not only demonstrates the feasibility of fine-tuning training files for multiple effect speech synthesis but also provides valuable insights for future improvements and extensions. It is expected to contribute to the realization of more expressive speech audio synthesis with more diverse effects.

References

[1] Watts, O., Henter, G. E., Merritt, T., Wu, Z., & King, S. (2016, March). From HMMs to DNNs: where do the improvements come from?. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5505-5509). IEEE.

[2] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Wu, Y. (2018, April). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4779-4783). IEEE.

[3] Van Den Oord, A., 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, 12.

[4] Prenger, R., Valle, R., & Catanzaro, B. (2019, May). Waveglow: A flow-based generative network for speech synthesis. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3617-3621). IEEE.

[5] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558.

[6] Kayte, S., & Gawali, B. (2015). A text-to-speech synthesis for Marathi language using festival and Festvox. International Journal of Computer Applications, 975, 8887.

[7] Gibbon, D., & Bachan, J. (2008, May). An Automatic Close Copy Speech Synthesis Tool for Large-Scale Speech Corpus Evaluation. In LREC.

[8] Steiner, I., & Maguer, S. L. (2017). Creating new language and voice components for the updated MaryTTS text-to-speech synthesis platform. arXiv preprint arXiv:1712.04787.

[9] Šimko, J., Törö, T., Vainio, M., & Suni, A. (2023). Prosody under control: Controlling prosody in text-to-speech synthesis by adjustments in latent reference space. In Proceedings of the 20th International Congress of Phonetic Sciences (pp. 3086-3090).

[10] Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013, October). Recent developments in opensmile, the munich open-source multimedia feature extractor. In Proceedings of the 21st ACM international conference on Multimedia (pp. 835-838).

[11] Vogt, T., André, E., & Bee, N. (2008). EmoVoice—A framework for online recognition of emotions from voice. In Perception in Multimodal Dialogue Systems: 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, PIT 2008, Kloster Irsee, Germany, June 16-18, 2008. Proceedings 4 (pp. 188-199). Springer Berlin Heidelberg.

[12] Eskimez, S. E., Sturge-Apple, M., Duan, Z., & Heinzelman, W. B. (2016, July). WISE: Web-based Interactive Speech Emotion Classification. In SAAIP@ IJCAI (pp. 2-7).

[13] Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: an introduction to voice assistants. Medical reference services quarterly, 37(1), 81-88.

[14] Plummer, A. R., & Beckman, M. E. (2016). Sharing Speech Synthesis Software for Research and Education Within Low-Tech and Low-Resource Communities. In INTERSPEECH(pp. 1618-1622).

[15] Buhalis, D., & Moldavska, I. (2022). Voice assistants in hospitality: using artificial intelligence for customer service. Journal of Hospitality and Tourism Technology, 13(3), 386-403.

[16] Kuligowska, K., Kisielewicz, P., & Włodarz, A. (2018). Speech synthesis systems: disadvantages and limitations. Int J Res Eng Technol (UAE), 7, 234-239.

Questionnaire

(The correct answer for the experimental group is marked in red)

Hello, thank you for participating in my survey on evaluating speech synthesis features. The purpose of this survey is to explore the effects of synthesizing a variety of speech features, including emphasis and speech rate adjustment, and to evaluate their performance in Chinese, English, and Spanish. Please note that you do not need to be fluent in Spanish, just answer the questions based on your intuition. Your feedback is very important for me to better evaluate this speech synthesis technology. I look forward to hearing from you!

您好，感谢您参与我的语音合成特征评估调查。本调查旨在探索包括强调、语速调整等多种语音特征的合成效果，并评估它们在中文、英文和西班牙语中的表现。请注意，您无需精通西班牙语，只需根据您的直觉回答相关问题即可。您的每一项反馈对我来说都非常重要，有助于我更好地评估这项语音合成技术。期待您的真实感受和宝贵意见！

Q1 **Which part of the sentence do you think is emphasized?**
您认为句子中的哪部分是强调的？

○ 今天下午我有 (This afternoon I have)

○ 两个小时的英语课 (English class for two hours)

○ 和两个小时的汉语课 (and then two hours of Chinese)

Q2 **Which part of the sentence do you think is emphasized?**
您认为句子中的哪部分是强调的？

○ 今天下午我有 (This afternoon I have)

○ 两个小时的英语课 (English class for two hours)

○ 和两个小时的汉语课 (and then two hours of Chinese)

Q3 **Which part of the sentence do you think is spoken more slowly?**
您认为句子中的哪部分是语速放慢的？

○ 他看起来 (He looks like)

○ 像个运动员 (a sportsman)

○ 但是其实 (but)

○ 是个作家 (he is a writer)

Q4 **Which part of the sentence do you think is spoken more slowly?**
您认为句子中的哪部分是语速放慢的？

○ 他看起来 (He looks like)

○ 像个运动员 (a sportsman)

○ 但是其实 (but)

○ 是个作家 (he is a writer)

Q5-1 **Which part of the sentence do you think is emphasized?**
您认为句子中的哪部分是强调的？

○ 我宁愿呆在家里 (I'd rather stay home)

○ 也不要 (than)

○ 在这种天气 (in this weather)

○ 中出门 (go out)

Q5-2 **Which part of the sentence do you think is spoken more slowly?**
您认为句子中的哪部分是语速放慢的？

○ 我宁愿呆在家里 (I'd rather stay home)

○ 也不要 (than)

○ 在这种天气 (in this weather)

○ 中出门 (go out)

Q6-1 **Which part of the sentence do you think is emphasized?**
您认为句子中的哪部分是强调的？

○ 我宁愿呆在家里 (I'd rather stay home)

○ 也不要 (than)

○ 在这种天气 (in this weather)

○ 中出门 (go out)

Q6-2 **Which part of the sentence do you think is spoken more slowly?**
您认为句子中的哪部分是语速放慢的？

○ 我宁愿呆在家里 (I'd rather stay home)

○ 也不要 (than)

○ 在这种天气 (in this weather)

○ 中出门 (go out)

Q7-1 **Do you think this speech sounds natural?**
您认为这个语音听起来自然吗？

○ 5 Excellent 优秀

○ 4 Good 良好

○ 3 Fair 一般

○ 2 Poor 较差

○ 1 Bad 差

Q7-2 **Do you think this speech sounds clear?**
您认为这个语音听起来清晰吗？

○ 5 Excellent 优秀

○ 4 Good 良好

○ 3 Fair 一般

○ 2 Poor 较差

○ 1 Bad 差

Q8-1 **Do you think this speech sounds natural?**
您认为这个语音听起来自然吗？

○ 5 Excellent 优秀

○ 4 Good 良好

○ 3 Fair 一般

○ 2 Poor 较差

○ 1 Bad 差

Q8-2 **Do you think this speech sounds clear?**
您认为这个语音听起来清晰吗？

○ 5 Excellent 优秀

○ 4 Good 良好

○ 3 Fair 一般

○ 2 Poor 较差

○ 1 Bad 差

Q9 **Which part of the sentence do you think is emphasized?**
您认为句子中的哪部分是强调的？

○ This fairy tale

○ is easy enough for

○ a seven year old child

○ to read

Q10 **Which part of the sentence do you think is emphasized?**
您认为句子中的哪部分是强调的？

○ This fairy tale

○ is easy enough for

○ a seven year old child

○ to read

Q11 **Which part of the sentence do you think is spoken more slowly?**
您认为句子中的哪部分是语速放慢的？

○ This fairy tale

○ is easy enough for

○ a seven year old child

○ to read

Q12 **Which part of the sentence do you think is spoken more slowly?**
您认为句子中的哪部分是语速放慢的？

○ This fairy tale

○ is easy enough for

○ a seven year old child

○ to read

Q13-1 **Which part of the sentence do you think is emphasized?**
您认为句子中的哪部分是强调的？

○ Cold and dry

○ splendid sunshine

○ what beautiful winter weather

Q13-2 **Which part of the sentence do you think is spoken more slowly?**
您认为句子中的哪部分是语速放慢的？

○ Cold and dry

○ splendid sunshine

○ what beautiful winter weather

Q14-1 **Which part of the sentence do you think is emphasized?**
您认为句子中的哪部分是强调的？

○ Cold and dry

○ splendid sunshine

○ what beautiful winter weather

Q14-2 **Which part of the sentence do you think is spoken more slowly?**
您认为句子中的哪部分是语速放慢的？

○ Cold and dry

○ splendid sunshine

○ what beautiful winter weather

Q15-1 **Do you think this speech sounds natural?**
您认为这个语音听起来自然吗？

○ 5 Excellent 优秀

○ 4 Good 良好

○ 3 Fair 一般

○ 2 Poor 较差

○ 1 Bad 差

Q15-2 **Do you think this speech sounds clear?**
您认为这个语音听起来清晰吗？

○ 5 Excellent 优秀

○ 4 Good 良好

○ 3 Fair 一般

○ 2 Poor 较差

○ 1 Bad 差

Q16-1 **Do you think this speech sounds natural?**
您认为这个语音听起来自然吗？

○ 5 Excellent 优秀

○ 4 Good 良好

○ 3 Fair 一般

○ 2 Poor 较差

○ 1 Bad 差

Q16-2 **Do you think this speech sounds clear?**
您认为这个语音听起来清晰吗？

○ 5 Excellent 优秀

○ 4 Good 良好

○ 3 Fair 一般

○ 2 Poor 较差

○ 1 Bad 差

Q17-1 **Do you think this speech sounds natural?**
您认为这个语音听起来自然吗？

○ 5 Excellent 优秀

○ 4 Good 良好

○ 3 Fair 一般

○ 2 Poor 较差

○ 1 Bad 差

Q17-2 **Do you think this speech sounds clear?**
您认为这个语音听起来清晰吗？

○ 5 Excellent 优秀

○ 4 Good 良好

○ 3 Fair 一般

○ 2 Poor 较差

○ 1 Bad 差

Q18-1 **Do you think this speech sounds natural?**
您认为这个语音听起来自然吗？

○ 5 Excellent 优秀

○ 4 Good 良好

○ 3 Fair 一般

○ 2 Poor 较差

○ 1 Bad 差

Q18-2 **Do you think this speech sounds clear?**
您认为这个语音听起来清晰吗？

○ 5 Excellent 优秀

○ 4 Good 良好

○ 3 Fair 一般

○ 2 Poor 较差

○ 1 Bad 差