university of groningen

campus fryslân

Optimizing Text-to-Speech: Investigating Training Data Volume for Human-Level Synthesis with Fastspeech2

Yi Lei

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Optimizing Text-to-Speech: Investigating Training Data Volume for Human-Level Synthesis with Fastspeech2

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
Supervisor 1's Dr.Do (Voice Technology, University of Groningen)
with the second reader being
()

Yi Lei S5712491

June 11, 2024

Acknowledgements

I am thankful to my supervisor Phat for his helpful and important advice and for the kind words and understanding during the process of writing this thesis. And great thanks to each member of the Voice Technology team, your insight and wisdom helped me a lot to acquire my skill in this field. Thanks to the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high-performance computing cluster, without it, this study can never be finished. Above all, thanks to my family and friends who always support me from the beginning.

Abstract

This study investigates the relationship between training data volume and Text-to-Speech (TTS) system performance, focusing on the FastSpeech 2 model. I aim to determine the amount of data necessary to achieve human-level speech synthesis. Hypothesizing that Mean Opinion Scores (MOS) increase with data augmentation until reaching a human-level threshold, I conduct experiments with varying data volumes. Participants then subjectively rate synthesized speech samples alongside natural speech. The research aims to advance TTS technology by providing insights into the critical role of training data volume, particularly in low-resource language settings.

Keywords: Text-to-Speech, TTS, FastSpeech 2, training data volume, speech synthesis, Mean Opinion Scores, MOS, human-level speech

Contents

1 Introduction

My passion lies within the realm of Text-to-Speech (TTS), particularly in understanding how to effectively utilize data to train models. FastSpeech 2 is a widely used and popular model in this field. I am intrigued by the intricate interplay between training data and the performance of TTS systems. How much training data will yield human-level speech? Through my research, I aim to investigate this relationship through a series of experiments coupled with subjective evaluations. By unraveling the nuances of how training data impacts TTS system performance, I aspire to contribute to the advancement of speech synthesis technology, particularly in the context of low-resource languages (LRL).

TTS technology aims to synthesize intelligible and natural-sounding speech from text (Taylor, 2009). In recent years, TTS has made significant progress due to advances in deep neural networks. The current trend in the TTS community is to adapt end-to-end models, which have demonstrated improved performance compared to traditional approaches.

Two state-of-the-art models, FastSpeech 2 (Ren et al., 2022) and NaturalSpeech (Tan et al., 2022), have achieved remarkable results. Both models were trained using the LJ Speech dataset. In subjective tests, FastSpeech 2 yielded a Mean Opinion Score (MOS) of 3.83 (± 0.08), while NaturalSpeech achieved a MOS of 4.56 (± 0.13), indicating a high level of naturalness in the synthesized speech.

Despite these breakthroughs, the relationship between the size of the training dataset and the performance of TTS models remains an area with limited clarity. Existing studies have yet to provide a definitive answer on the optimal amount of data required to achieve human-level speech synthesis. This gap in knowledge prompts further investigation into how the quantity and quality of training data influence the efficiency and output of TTS systems.

The challenge of data scarcity is particularly relevant in the context of Under-Resourced Languages (URLs) (Besacier et al., 2014). To address this issue, the TTS community has explored new pre-training strategies to improve data efficiency(Prajwal & Jawahar, 2021) and adapted transfer learning techniques from high-resource languages to low-resource languages(LRLs). While significant progress has been made in TTS technology, there is still a need for further research to better understand the relationship between training data and model performance, especially in the context of under-resourced languages.

1.1 Research Question and Hypothesis

TTS technology has seen significant advancements with models such as FastSpeech 2, known for its efficiency and ability to generate human-like speech(Ren et al., 2022). However, the relationship between the volume of training data and the quality of synthesized speech remains under-explored, especially in the context of low-resource languages. Understanding this relationship is crucial for optimizing TTS systems and ensuring their accessibility across different languages and datasets.

1.1.1 Research Questions

RQ1: Is there a relationship between the quality of synthesized speech and the amount of training data used in TTS systems?

RQ2: Is there a point at which increasing the amount of training data does not significantly improve the quality of synthesized speech in TTS systems? If so, what is this point?

RQ3: For low-resource languages, is there a minimum amount of training data required to produce ”acceptable” quality synthesized speech using FastSpeech 2?

1.1.2 Hypotheses

Hypothesis 1: There exists a positive correlation between the amount of training data and the MOS in TTS systems. As the volume of training data increases, the MOS will also increase.

Hypothesis 2: There exists a point of diminishing returns for TTS systems, beyond which additional training data does not result in significant improvements in MOS.

Hypothesis 3: For low-resource languages, TTS systems can achieve ”acceptable” quality synthesized speech with a minimum of 10 hours of annotated speech data. ”Acceptable” quality is defined as speech that is clear and understandable, as measured by a MOS threshold of 3(Kirkland et al., 2023) determined by subjective evaluations.

By addressing these research questions and hypotheses, the study aims to provide insights into the critical role of training data volume in enhancing TTS systems, contributing to the broader field of speech synthesis and its applications in low-resource language settings.

2 Literature Review

This chapter explores the evolution and current advancements in TTS technology, highlighting key developments from early rule-based and concatenative methods to contemporary neural network-based approaches. It delves into the progression of TTS systems through Statistical Parametric Speech Synthesis (SPSS) and the revolutionary impact of deep learning models like WaveNet, Tacotron, and Transformer-based architectures. The chapter also addresses the critical role of training data quantity and the performance of TTS model, particularly in the context of low-resource languages. Additionally, it outlines methodologies for evaluating TTS systems, including subjective tests and the Mean Opinion Score (MOS), and discusses the details of the M-AILABS Speech Dataset that has been used for this research.

2.1 Development of TTS

2.1.1 Early stage of TTS

TTS technology aims to synthesize intelligible and natural-sounding speech from text (Taylor, 2009). In other words, it involves getting computers to read out loud, converting text into speech. TTS has come a long way from focusing on producing basic sounds and phonemes using analog circuits and simple algorithms in the 1970s to adopting neural networks that directly generate speech waveforms from text, bypassing intermediate representations and producing remarkably natural-sounding speech.

The early TTS systems relied on rule-based and concatenative methods. Rule-based systems, such as those developed in the 1960s and 1970s, used predefined linguistic rules to convert text into phonetic representations and then into speech. This method relied on linguistic expertise and was resource-intensive. On the other hand, concatenative systems, popular in the 1980s and 1990s, pieced together prerecorded speech segments to form complete sentences. However, these systems required large amounts of pre-recorded data and could not generate words that had not been recorded.

2.1.2 Statistical parametric speech synthesis

In the early 2000s, Statistical Parametric Speech Synthesis (SPSS) emerged, leveraging statistical models to generate speech. SPSS systems, such as those based on Hidden Markov Models (HMMs), provided more flexibility and naturalness compared to concatenative methods. These systems required large amounts of training data to model the statistical properties of speech effectively. One notable example is the HMM-Based Speech Synthesis System (HTS), developed by the HTS Working Group. HTS is an open-source toolkit for HMM-based speech synthesis that models various aspects of speech, such as spectral, excitation, and duration parameters, using HMMs. It supports multiple speakers and multi-lingual synthesis. However, HTS requires a large, phonetically, and prosodically diverse speech corpus with aligned text transcriptions. Typically, 5 to 10 hours of well-annotated speech data are needed for training to achieve higher quality and more natural-sounding synthesis.

2.1.3 Neural network TTS

Following the SPSS phase, TTS technology adopted deep neural networks as the base of its architecture. The advent of deep learning revolutionized TTS technology. One pioneering model in this domain is WaveNet, introduced by (van den Oord et al., 2016). WaveNet employs dilated convolutions to generate high-quality speech waveforms directly from text inputs. By modeling the raw audio waveform at the sample level, WaveNet captures subtle nuances and inflections of human speech, producing remarkably natural-sounding speech. However, WaveNet's autoregressive nature, where each audio sample is generated sequentially based on previous samples, results in slow inference speeds, making real-time applications challenging.

To address efficiency concerns, researchers proposed various architectures balancing quality and computational efficiency. Notably, Tacotron (Wang et al., 2017) and Tacotron 2 (Shen et al., 2018) employ a sequence-to-sequence architecture with an attention mechanism, enabling direct mapping from text to spectrograms or mel-spectrograms. Tacotron models bypass the autoregressive generation process by producing spectrograms, which are then converted to waveforms using vocoders. Tacotron 2 combines a spectrogram prediction network with a modified WaveNet vocoder, significantly enhancing the naturalness and intelligibility of the synthesized speech while reducing inference time compared to WaveNet.

2.1.4 Recent development of TTS

In recent years, Transformer-based architectures have gained prominence due to their parallelization capabilities and effectiveness in capturing long-range dependencies. Notable examples include Transformer TTS (Li et al., 2018) and FastSpeech (Ren et al., 2019). These models leverage self-attention mechanisms for efficient text-to-spectrogram synthesis. Transformers excel at handling the sequential nature of text and speech data by processing entire sequences in parallel, significantly speeding up the generation process. FastSpeech, in particular, introduces a non-autoregressive approach that generates mel-spectrograms in parallel, further improving inference speed and scalability. FastSpeech 2 (**?**) builds on this foundation by enhancing the quality of speech synthesis and addressing issues related to duration modeling, pitch prediction, and energy prediction.

Further innovations include models like Parallel WaveGAN (Yamamoto et al., 2020) and Multi-band MelGAN (Yang et al., 2021), which focus on enhancing the efficiency of waveform generation. These models employ generative adversarial networks (GANs) to produce high-fidelity audio with reduced computational complexity, enabling faster and more scalable TTS systems.

The evolution of TTS systems from WaveNet to Tacotron and Transformer-based models represents a significant leap in both speech quality and generation efficiency. These advancements underscore the rapid progress in neural network-based TTS, paving the way for more natural and responsive speech synthesis technologies in various applications, from virtual assistants and interactive voice response systems to accessibility tools for individuals with speech impairments.

The landscape of TTS systems has been transformed by the introduction of neural network-based architectures, each addressing different aspects of quality, speed, and scalability. WaveNet set a new standard for natural-sounding speech, while Tacotron and its successors improved efficiency and ease of training. Transformer-based models, with their parallel processing capabilities, further enhanced the speed and accuracy of TTS systems. The continuous innovation in this field promises even more sophisticated and accessible speech synthesis solutions in the future.

TTS technology has made significant progress due to advances in deep neural networks. These advancements have revolutionized the field, allowing for the creation of more natural and expressive synthetic speech. The current trend in the TTS community is to adopt end-to-end models, which have demonstrated improved performance compared to traditional approaches that often require separate components for text analysis, phoneme generation, and waveform synthesis.

Among the state-of-the-art models, FastSpeech 2 (Ren et al., 2022) and NaturalSpeech (Tan et al., 2022) have achieved remarkable results. Both models were trained using the LJ Speech dataset, a widely-used dataset in the TTS community. FastSpeech 2 builds upon the original FastSpeech model by introducing improvements in pitch prediction, energy prediction, and duration modeling, which contribute to the model's ability to generate high-quality and expressive speech. In subjective tests, FastSpeech 2 yielded a Mean Opinion Score (MOS) of 3.83 ($\pm$ 0.08), reflecting a good level of naturalness and intelligibility in the synthesized speech.

On the other hand, NaturalSpeech has pushed the boundaries of TTS quality even further. This model leverages advanced techniques such as adversarial training and variational autoencoders to produce highly natural and human-like speech. NaturalSpeech achieved a MOS of 4.56 ($\pm$ 0.13) in subjective evaluations, indicating a superior level of naturalness and listener satisfaction compared to FastSpeech 2.

2.1.5 Motivation of this study

Despite these breakthroughs, the relationship between the size of the training dataset and the performance of TTS models remains an area with limited clarity. Existing studies have yet to provide a definitive answer on the optimal amount of data required to achieve human-level speech synthesis. This gap in knowledge prompts further investigation into how the quantity and quality of training data influence the efficiency and output of TTS systems.

It is well understood that having a larger and more diverse training dataset can improve the generalization capabilities of a model, enabling it to handle a wider range of inputs and produce more varied speech outputs. However, the diminishing returns on performance gains with increasingly larger datasets have not been thoroughly quantified. Additionally, the quality of the dataset, including the clarity of recordings, the diversity of speaking styles, and the representativeness of different phonetic contexts, plays a crucial role in determining the model's performance.

Further research is needed to explore these dynamics and establish guidelines for dataset construction that balance size and quality. Investigating the trade-offs between dataset size, model complexity, and training time is essential for developing more efficient and effective TTS systems. Moreover, understanding these factors can help optimize the training process, making it more feasible to achieve high-quality speech synthesis with limited computational resources.

In conclusion, while end-to-end models like FastSpeech 2 and NaturalSpeech represent significant advancements in TTS technology, the field continues to grapple with questions about the optimal data requirements for training these models. Addressing these questions through rigorous research will be key to further enhancing the naturalness and efficiency of TTS systems, ultimately bringing us closer to achieving human-level speech synthesis.

2.2 Training Methodologies and Data Size

The performance of TTS models heavily relies on the quantity and quality of training data. Larger datasets often lead to more robust and natural-sounding synthesis. For instance, Google's Tacotron 2 model was trained on a dataset containing thousands of hours of speech paired with text transcripts, contributing to its impressive quality and generalization capabilities (Shen et al., 2018).

However, obtaining vast amounts of high-quality training data can be challenging. To mitigate this issue, data augmentation techniques, such as speed perturbation and speaker adaptation, have been employed to enrich the training dataset and enhance model robustness (Sotelo et al., 2017).

Furthermore, the advent of transfer learning has enabled leveraging pre-trained models on large-scale datasets, such as LibriSpeech or LJSpeech, followed by fine-tuning on smaller, domain-specific datasets. This approach has been demonstrated to improve model performance significantly while reducing the need for massive amounts of domain-specific training data (Ping et al., 2020).

In the context of low-resource languages, speech synthesis poses unique challenges and opportunities. Researchers, such as (Pine et al., 2022), underscore the importance of developing TTS systems for languages with limited digital resources. The primary motivation is to support language revitalization efforts, enabling communities to preserve and promote their linguistic heritage through technology. (Pine et al., 2022) trained the FastSpeech 2 model with varying amounts of data and conducted subjective tests for each model to evaluate performance. The MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) scores indicated that the amount of training data positively affects the model's performance, demonstrating that even incremental increases in data can enhance synthesis quality.

Despite these positive findings, the study highlighted several areas needing further exploration. Firstly, there was a lack of detailed discussion about the threshold of data required to achieve optimal performance. This is particularly important for low-resource languages where data collection can be challenging and resource-intensive. Determining a clear threshold would help allocate resources more efficiently and guide future data collection efforts.

Furthermore, the study's largest dataset comprised only 24 hours of speech data, which, while providing valuable insights, may not be sufficient to generalize findings across different languages and contexts. The subjective tests, involving 30 participants, provided useful feedback but also indicated the need for larger and more diverse participant groups to ensure the robustness and reliability of the results. The limited participant pool may introduce biases and limit the applicability of the findings to broader populations.

The paper also emphasizes the critical role of community involvement in data collection and validation processes. Engaging native speakers in these processes not only improves the cultural and linguistic accuracy of the synthesized speech but also fosters a sense of ownership and empowerment within the community. This participatory approach is essential for ensuring that the TTS systems are well-received and effectively utilized by the communities they are designed to support.

Additionally, (Pine et al., 2022) advocate for the strategic use of pre-trained models and transfer learning in low-resource settings. By leveraging models that have been trained on large, diverse datasets, researchers can adapt these models to new languages with relatively small amounts of data. This approach significantly reduces the data requirements and accelerates the development of high-quality TTS systems for low-resource languages, thereby contributing to the broader goal of language preservation and revitalization.

In conclusion, while the study (Pine et al., 2022) provides valuable insights into the effects of

training data size on TTS performance, it also highlights the need for further research to establish data thresholds, involve larger participant groups, and explore community-driven data collection methods. These efforts are crucial for developing effective and sustainable TTS systems that support the revitalization and preservation of low-resource languages.

2.3 Architecture of FastSpeech 2

FastSpeech 2 is an advanced TTS model that significantly improves upon its predecessor by incorporating a more robust and efficient architecture. It features a fully end-to-end neural network design that includes an encoder, variance adaptor, and decoder. This architecture is designed to generate high-quality speech in a non-autoregressive manner, meaning it can produce speech much faster than autoregressive models. It is the model that has been used for this study.

The encoder in FastSpeech 2 processes the input text to generate a sequence of hidden representations. These representations are then passed to the variance adaptor, which adjusts the prosodic features such as pitch, energy, and duration to modulate the speech synthesis process. The variance adaptor is crucial for generating speech that sounds natural and expressive.

Pitch information is essential in FastSpeech 2 because the variance adaptor uses pitch to control the prosody of the synthesized speech. Accurate pitch extraction allows the model to adjust the intonation patterns of the output speech, making it sound more natural and expressive. Without precise pitch data, the synthesized speech would lack the natural variation in intonation that characterizes human speech, leading to monotonous and unnatural output.

In addition to pitch, the energy feature in FastSpeech 2 helps the model learn how to modulate loudness across different phonetic contexts. This modulation is crucial for emphasizing certain words or phrases, contributing to the overall naturalness and emotional impact of the synthesized speech. Without energy features, the model might produce flat and less engaging speech.

Furthermore, FastSpeech 2 uses Mel-spectrograms as intermediate targets for the decoder, allowing the model to learn the complex patterns of speech sounds more effectively. Mel-spectrograms capture both phonetic and acoustic information, enabling the model to generate high-quality and intelligible speech. These representations serve as a bridge between the text input and the final audio output, ensuring that the synthesized speech accurately reflects the intended phonetic content.

The architecture of FastSpeech 2, with its emphasis on accurate prosody and acoustic features, necessitates the use of pitch, energy, and Mel-spectrograms. These components are integral to the model's ability to produce natural-sounding and expressive speech efficiently, making FastSpeech 2 a significant advancement in TTS technology.

2.4 Subjective test

In this study, a subjective test is used to evaluate the quality and naturalness of synthesized speech produced by FastSpeech 2. It is an effective tool for assessing how well FastSpeech 2 can mimic natural human speech and convey meaning and emotion. The process consists of several steps. First, during the listening evaluation, participants listen to audio samples of synthesized speech produced by different FastSpeech 2 models. These audio samples typically consist of various sentences generated by FastSpeech 2, and in this experiment, the same sentence is used for all samples to eliminate the effect of different sentences.

Next, in the scoring phase, participants provide subjective ratings or feedback on different aspects of the synthesized speech, such as naturalness and quality, using rating scales from 1 to 5 to rate the speech produced by different models. Following this, I will analyze the collected data to identify trends and differences between each speech sample, looking for patterns in participants' ratings or comments to understand which FastSpeech 2 model performs better and why. Finally, the results of subjective tests can be used to compare different FastSpeech 2 models and evaluate the effectiveness of increasing data size on the performance of the FastSpeech 2 model.

2.5 MOS

Follow-up subjective test, MOS is a widely used subjective evaluation metric to assess the quality of synthesized speech. MOS is determined through subjective testing, where human listeners are asked to rate the quality of synthesized speech samples on a numerical scale. In this study, MOS comes from subjective tests.

2.6 URL

"Under-Resourced Languages" (URLs) or "lower-resourced languages" refer to languages that lack substantial digital resources and support compared to more widely used or dominant languages, for example, Frisian in Friesland Netherlands. These languages often have limited or no access to technologies such as machine translation, speech recognition, and text-to-speech systems, primarily due to the scarcity of digital data like texts, audio recordings, and annotated materials needed to train computational models. The challenge of data scarcity is particularly relevant in the context of Under-Resourced Languages (URLs) or lower resourced languages (Besacier et al., 2014). To address this issue, the TTS community has explored new pre-training strategies to improve data efficiency (Prajwal & Jawahar, 2021)and adapted transfer learning techniques from high-resource languages to Low-Resource Languages (LRLs) (Tu et al., 2019). While significant progress has been made in TTS technology, there is still a need for further research to better understand the relationship between training data and model performance, especially in the context of under-resourced languages.

3 Methodology

This chapter outlines the methodology used to address the research question and validate the hypothesis. The study investigates the relationship between training data volume and the performance of TTS systems, specifically focusing on the FastSpeech 2 model. The methodology encompasses the dataset selection and preprocessing, feature extraction, model training, and evaluation techniques. In subsection 3.1, the M-AILABS Speech Dataset is discussed, highlighting its suitability for the research. Subsection 3.2.2 details the feature extraction process using robust algorithms to capture essential speech characteristics. Following this, subsection 3.2.1 explains the application of the Montreal Forced Aligner (MFA) for precise phoneme alignment and duration extraction. Subsection 3.4 elaborates on the evaluation methods, including subjective assessments, to comprehensively analyze the models' performance.

3.1 Dataset

3.1.1 Overview

The dataset chosen for this study is the M-AILABS Speech Dataset, a comprehensive and freely available dataset suitable for training both speech recognition and speech synthesis models. The dataset comprises audio files in WAV format, recorded in mono at a sampling rate of 16000 Hz. This format ensures high-quality audio suitable for the training requirements of TTS models.

3.1.2 Dataset Details

For this study, I focus on the English (female voice) subset of the M-AILABS Speech Dataset. This subset contains a total of 45 hours and 34 minutes of recorded speech, amounting to 4.9 GB of data. It includes 23,561 sentences, providing a diverse range of phonetic contexts and speaking styles, which are essential for training robust and natural-sounding TTS systems.

3.1.3 Motivation for Choosing the M-AILABS Dataset

The M-AILABS Speech Dataset was selected for several key reasons. Firstly, its free and open access aligns with the ethical considerations of using open-source data for research purposes. This accessibility ensures that the research can be replicated and extended by other researchers in the future. Secondly, the audio files are consistently formatted and maintained at a high quality (mono, 16000 Hz), which is crucial for training effective TTS models. High-quality, consistent data helps in achieving reliable and reproducible results. Lastly, the total duration of 45 hours and 34 minutes provides ample data to create multiple subsets, enabling the study of the impact of training data volume on TTS performance. The substantial size of the dataset ensures that models can be trained effectively, even when data is segmented into smaller subsets.

3.1.4 Split of dataset and motivation

To systematically investigate the effect of training data volume on TTS performance, the English (female voice) subset is divided into five distinct subsets of varying durations: 1 hour, 15 hours, 25

hours, 35 hours, and 45 hours. This division facilitates a detailed analysis of how different amounts of training data influence the quality of the generated speech.

One-Hour Dataset: This subset is designed to mimic the conditions of under-resourced languages, which often face challenges in collecting sufficient training data. By analyzing the performance of TTS models trained on this minimal dataset, I aim to understand the limitations and potential of TTS systems in low-resource settings, it is to evaluate if one hour is enough for a TTS model to generate intelligible speech.

Fifteen-Hour Dataset: This subset provides a moderate increase in data volume, allowing for the observation of improvements in model performance as more data becomes available. This helps in identifying the point at which additional data begins to have a noticeable impact on TTS quality.

Twenty-Five-Hour Dataset: Serving as a mid-point in the data volume spectrum, this subset allows for the assessment of continued improvements in TTS performance with further increases in training data.

Thirty-Five-Hour Dataset: This subset represents a higher data volume, enabling the study of diminishing returns as data volume increases. It helps in identifying whether a substantial amount of data is necessary for achieving near-human-level synthesis.

Forty-Five-Hour Dataset: Utilizing the entire available dataset, this subset serves as the benchmark for maximum data volume. It provides insights into the upper limits of TTS performance with extensive training data.

3.2 Data pre-processing

There are two steps for pre-processing the data before training.

3.2.1 Montreal Forced Aligner (MFA)

The Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) is a tool used to generate precise alignments between phonemes and their corresponding audio segments. MFA is particularly effective for creating accurate phone duration targets, which are essential for the training of TTS models like FastSpeech 2. The MFA is trained on the same data that will be used for training the TTS models, ensuring that the alignments are tailored to the specific characteristics of the dataset, leading to more accurate and reliable phoneme boundaries. Using the trained MFA, phone durations are extracted for each subset of the dataset by aligning the text transcripts with the corresponding audio files to generate time-aligned phonetic transcriptions.

One significant advantage of MFA is its ability to provide suitable alignments even when trained on limited data. This is particularly beneficial for under-resourced languages, where the availability of extensive training data is often a challenge. By leveraging MFA, high-quality alignments and duration targets can be ensured, facilitating effective TTS model training even in low-resource scenarios. In this experiment, MFA was applied to the entire dataset. MFA serves two main functions: text normalization, which converts text sequences into a standardized format to ensure consistency across the dataset, and phoneme conversion, which transforms text into phonetic sequences to facilitate accurate speech synthesis.

By utilizing the Montreal Forced Aligner (MFA) for preprocessing, the quality and accuracy of the training data are enhanced, which is crucial for the performance of FastSpeech 2 models. This approach not only ensures precise phone duration targets but also addresses the challenges of

working with limited data, making it an ideal choice for this study. It is worth noting the issue of out-of-vocabulary (OOV) words, which are words that occur in the dataset but are not included in the dictionary. For such words, the alignment assigns a special label, making the TTS model unable to learn anything useful from them. To avoid this and increase the performance of the model, these words have been manually added to the dictionary.

3.2.2 Feature Extraction

Feature extraction is a crucial step in preparing the audio data for training FastSpeech 2 models. The extracted features, including pitch, energy, and Mel-spectrograms, are essential for capturing the nuanced characteristics of speech, enabling the model to generate natural and high-quality synthesized speech.

Pitch is an important prosodic feature that influences the perceived intonation and emotional tone of speech. I use robust algorithms to extract pitch contours from the audio files. Accurate pitch extraction ensures that the synthesized speech maintains the natural prosody and intonation patterns of the original recordings.

Energy features capture the loudness variations in speech, which are critical for conveying emphasis and emotional content. By extracting energy levels at each frame, the model can learn to modulate loudness appropriately, enhancing the expressiveness and naturalness of the synthesized speech. Energy extraction is crucial for controlling the dynamics and expressiveness of speech.

Mel-spectrograms represent the spectral properties of speech in a way that aligns with human auditory perception. They are generated by applying a Short-Time Fourier Transform (STFT) to the audio signal, followed by mapping the frequencies to the Mel scale. Mel-spectrograms provide a time-frequency representation of the audio, capturing detailed phonetic and acoustic information that is essential for high-quality speech synthesis. Mel-spectrograms provide a detailed and perceptually relevant representation of the speech signal.

These features are essential to train the FastSpeech 2 models, ensuring that the models have access to comprehensive and high-quality acoustic information. This detailed feature extraction process contributes to the overall performance and naturalness of the generated speech, making it a critical component of the TTS system.

3.3 Model Training

The preprocessed data, including the extracted features, is used to train multiple FastSpeech 2 models. Each model is trained on a different subset of the dataset, allowing for a comprehensive analysis of the impact of training data volume on TTS performance. The training process involves several key steps to ensure optimal model performance and reliable results. In this study, I used the implementation of FastSpeech 2 from (**?**), which is available on Git Hub.

3.3.1 Training Configuration

The FastSpeech 2 model architecture is designed to convert text sequences into corresponding speech waveforms using a non-autoregressive approach, which enables faster and more efficient speech synthesis compared to traditional autoregressive models. The key components include the Text Encoder, which converts text sequences into hidden representations; and the Variance Adaptor, which adjusts

the pitch, energy, and duration of the hidden representations; the Mel-Spectrogram Decoder, which generates Mel-spectrograms from the adapted hidden representations; and the Post-Net, which enhances the Mel-spectrograms to improve audio quality.

Hyperparameters are carefully tuned to optimize the performance of the FastSpeech 2 models. These hyperparameters are crucial for ensuring that the models converge effectively and generate high-quality synthesized speech. The chosen values are informed by the original FastSpeech 2 study and are tailored to suit the specifics of this experiment. The batch size determines the number of training samples processed simultaneously during each iteration. A larger batch size can stabilize training by averaging gradients, but it also requires more memory. For this study, the batch size is set to 16, balancing computational efficiency with the need to fit the model within the memory constraints of the available hardware.

Dropout is a regularization technique used to prevent overfitting by randomly dropping a fraction of neurons during training, encouraging the model to learn robust features that generalize well to unseen data. The dropout rate for the encoder and decoder is set to 0.2, meaning 20% of the neurons in these layers are dropped during each training step, reducing the risk of overfitting while maintaining sufficient model capacity for learning. For the variance predictor, the dropout rate is set to 0.5. Given that the variance predictor plays a crucial role in adapting the prosody features (pitch, energy, duration), a higher dropout rate helps ensure it generalizes well across different speech variations.

The number of training steps determines how long the model is trained. More steps allow the model to learn more from the data, but excessive training can lead to overfitting. Based on the original FastSpeech 2 paper, the total training steps are set to 160,000 (160K). This number has been found to be sufficient for the model to converge, achieving a balance between learning capacity and training duration. By setting a fixed number of steps, I ensure consistency across all models trained on different data subsets, facilitating a fair comparison of their performance. By tuning these hyperparameters, I aim to optimize the training process for the FastSpeech 2 models, ensuring they perform well across different data volumes. Each hyperparameter setting is applied consistently across all models trained in this experiment to maintain comparability and ensure that observed performance differences are due to the varying amounts of training data rather than changes in the training configuration.

3.3.2 Training Procedure

The data preparation process begins by dividing each subset of the dataset (1 hour, 15 hours, 25 hours, 35 hours, and 45 hours) into training and validation sets, with a ratio of 0.5:9.5. The extracted features, such as pitch, energy, and Mel-spectrograms, are then fed into the FastSpeech 2 model. The text encoder processes the input text sequences, while the variance adaptor adjusts the hidden representations based on these extracted features.

The model is trained to minimize a composite loss function that includes several components. The Mean Squared Error (MSE) measures the difference between the predicted and target Mel-spectrograms. Duration loss evaluates the accuracy of the predicted phone durations compared to the ground truth. Additionally, pitch and energy loss ensure that the predicted pitch and energy values align with the extracted features.

The training loop involves iterating over the training set, updating the model's parameters based on the loss function, and validating the model's performance on the validation set. Early stopping is employed to prevent overfitting, halting training if the validation loss does not improve for a specified

number of epochs.

3.3.3 Training Time

Training the FastSpeech 2 models is a computationally intensive task that requires substantial resources to ensure timely and effective model convergence. For this study, the training process is carried out on a single NVIDIA V100 GPU, known for its high performance in deep learning applications.

The batch size for each training session is set to 16 sentences. This batch size is chosen to balance the computational load and memory constraints, ensuring efficient utilization of the GPU while maintaining stability during training.

The training duration for each model varied depending on the amount of training data used. The model trained on 45 hours of data took 27.54 hours to reach coverage. In comparison, the model with 35 hours of data required 23.7 hours to complete training. For the model that used 25 hours of training data, it took 22.15 hours to achieve full coverage. The training time decreased further for the model with 15 hours of data, which took 19.79 hours. Notably, the model with only 1 hour of training data required 14.85 hours to train fully.

Several factors contribute to the training time, including the complexity of the model, the size of the dataset, and the efficiency of the GPU. The NVIDIA V100 GPU, with its 32 GB of memory and optimized architecture for deep learning, significantly accelerates the training process by efficiently handling the large volumes of data and complex computations involved in training FastSpeech 2 models. To summarize, the training time for each FastSpeech 2 model is influenced by a batch size of 16 sentences, the use of an NVIDIA V100 GPU, and a total of 160,000 training steps, resulting in a total training time of 108 hours for all the models.

3.3.4 Test Sentences

After all the models finished training, two sentences were selected from the English (female voice) subset of the M-AILABS Speech Dataset. This subset was chosen to eliminate the effect of different voices on the results.

The first sentence is: ”He ought now to have been at school; but his mama had taken him home for a month or two, on account of his delicate health.” The second is: ”Mister Preston replied, ’Certainly. I am that and many other things besides, at your service.’”

Each sentence was generated by all the models, resulting in 12 speech audio samples (including the ground truth) used in the subjective test. They can be found via this link.

3.4 Evaluation

The evaluation phase is crucial for assessing the performance and effectiveness of the trained FastSpeech 2 models. This section outlines the methods used to evaluate the models, including subjective assessments. The goal is to determine how varying volumes of training data impact the quality of synthesized speech.

3.4.1 Subjective Evaluation

Subjective evaluation is conducted to assess the naturalness and intelligibility of the synthesized speech, involving human listeners rating the audio samples on several criteria. Participants provide Mean Opinion Scores (MOS), rating the naturalness and quality of the synthesized speech on a scale from 1 to 5, where 1 signifies ”poor” and 5 signifies ”excellent.” It is explained at the beginning of the test that naturalness refers to how closely the synthesized speech resembles human speech in terms of tone, rhythm, intonation, and fluidity, assessing whether the audio sounds like it was spoken by a real person. Quality pertains to the technical aspects of the audio, including clarity, lack of distortions or artifacts, consistency of volume, and overall fidelity, measuring how pleasant and clear the audio sounds to the listener. Multiple samples are rated to ensure a comprehensive assessment, and the average MOS is calculated for each model.

The subjective evaluation involves a diverse group of participants to ensure a wide range of feedback and to minimize biases. The ratings from these subjective tests provide a holistic view of the model's performance.

3.4.2 Statistical Analysis

The data collected from subjective evaluations are subjected to rigorous statistical analysis to draw meaningful conclusions. Firstly, Analysis of Variance (ANOVA) is used to determine whether there are statistically significant differences in the performance metrics across models trained with different data volumes. Following this, post-hoc tests, specifically Tukey's HSD, are conducted to perform pairwise comparisons between different models, helping to identify specific differences in performance metrics between models trained with varying data volumes. Additionally, correlation analysis is conducted to explore the relationship between training data volume and performance metrics, aiding in understanding how changes in data volume impact the quality of synthesized speech. This thorough statistical analysis ensures that the conclusions drawn from the evaluation are robust and reliable, providing a solid foundation for answering the research question and validating the hypothesis.

3.4.3 Results Interpretation

The results from the evaluation are interpreted in the context of the research question and hypothesis, leading to several key outcomes. Firstly, the analysis identifies the optimal training data volume that achieves the best balance between model performance and computational efficiency. Secondly, by comparing models trained on minimal data with those trained on larger datasets, the study provides insights into the challenges and potential of TTS systems in low-resource language settings. Lastly, based on these findings, recommendations are made for optimizing training data volume in TTS research and for further exploration of data-efficient training strategies. Through this comprehensive evaluation, the study aims to provide valuable insights into the relationship between training data volume and the performance of FastSpeech 2-based TTS systems, thereby contributing to the advancement of TTS technology, particularly for low-resource languages.

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

4 Results

This section presents the experimental results, comparing the performance of different TTS models trained with varying amounts of data from the M-AILABS Speech Dataset. I have recruited 32 participants via the internet. The evaluation focuses on two main metrics: quality rating and naturalness rating, both measured through subjective assessments by human listeners. The results are summarized in Table 1, showcasing the mean and standard deviation (SD) of these ratings for each model, including a ground truth comparison. These findings provide insights into how the volume of training data influences the perceived quality and naturalness of the synthesized speech.1.

Table 1: Subjective test results of TTS systems.

Sample	Model	Quality Rating (Mean)	Quality Rating (SD)	Naturalness Rating (Mean)	Naturalness Rating (SD)
Sample 1	1 Hour Model	2.38	1.11	2.32	1.21
Sample 1	15 Hour Model	2.71	1.04	2.82	1.25
Sample 1	25 Hour Model	2.65	1.00	2.65	1.11
Sample 1	35 Hour Model	2.50	0.85	2.68	0.93
Sample 1	45 Hour Model	3.03	0.95	3.00	1.03
Sample 1	Ground Truth	4.21	0.63	4.24	0.88
Sample 2	1 Hour Model	3.18	1.12	3.26	0.88
Sample 2	15 Hour Model	2.97	0.98	3.06	1.11
Sample 2	25 Hour Model	2.91	0.85	3.06	0.94
Sample 2	35 Hour Model	2.94	1.00	2.97	0.79
Sample 2	45 Hour Model	2.94	1.00	3.00	1.06
Sample 2	Ground Truth	3.97	0.87	3.94	0.78

Examining Table 1, it is evident that the amount of training data has a significant impact on the quality and naturalness of the synthesized speech. The analysis reveals several key trends and observations.

First, there is a clear positive correlation between the amount of training data and the quality and naturalness ratings. For Sample 1, the 1 Hour Model scores a mean quality rating of 2.38 and a mean naturalness rating of 2.32, both with relatively high standard deviations (SD) of 1.11 and 1.21, respectively. As the training data volume increases, these ratings generally improve. The 45 Hour Model achieves the highest ratings among the TTS models with a mean quality rating of 3.03 (SD 0.95) and a mean naturalness rating of 3.00 (SD 1.03). This demonstrates a substantial improvement over the models trained with less data. The ground truth, unsurprisingly, receives the highest ratings of 4.21 (SD 0.63) for quality and 4.24 (SD 0.88) for naturalness.

Similarly, for Sample 2, the 1 Hour Model starts with a relatively high mean quality rating of 3.18 and a naturalness rating of 3.26, with standard deviations of 1.12 and 0.88, respectively. The ratings for the other models do not show a consistent increase with more data; however, the variations in standard deviations indicate differing listener perceptions. The 45 Hour Model does not show a clear superiority in this case, scoring a mean quality rating of 2.94 (SD 1.00) and a mean naturalness rating of 3.00 (SD 1.06). The ground truth again achieves the highest ratings of 3.97 (SD 0.87) for quality and 3.94 (SD 0.78) for naturalness.

Overall, the improvement in ratings for the models trained with more data is more pronounced in Sample 1 than in Sample 2. This indicates that while increased data volume generally enhances TTS performance, the extent of improvement can vary depending on the specific sample.

The standard deviations also provide insights into the consistency of the ratings. The models trained with larger datasets tend to have lower standard deviations, indicating more consistent listener perceptions of quality and naturalness. For example, the 45 Hour Model in Sample 1 has standard deviations of 0.95 and 1.03 for quality and naturalness, respectively, compared to 1.11 and 1.21 for the 1 Hour Model.

In conclusion, the analysis shows that increasing the training data volume generally leads to better and more consistent quality and naturalness ratings for synthesized speech. However, the extent of improvement can vary between different samples, and even the best-performing TTS models do not yet match the quality of the ground truth recordings. These findings highlight the importance of training data volume in developing high-quality TTS systems and suggest areas for further research, such as optimizing data utilization and exploring model enhancements.

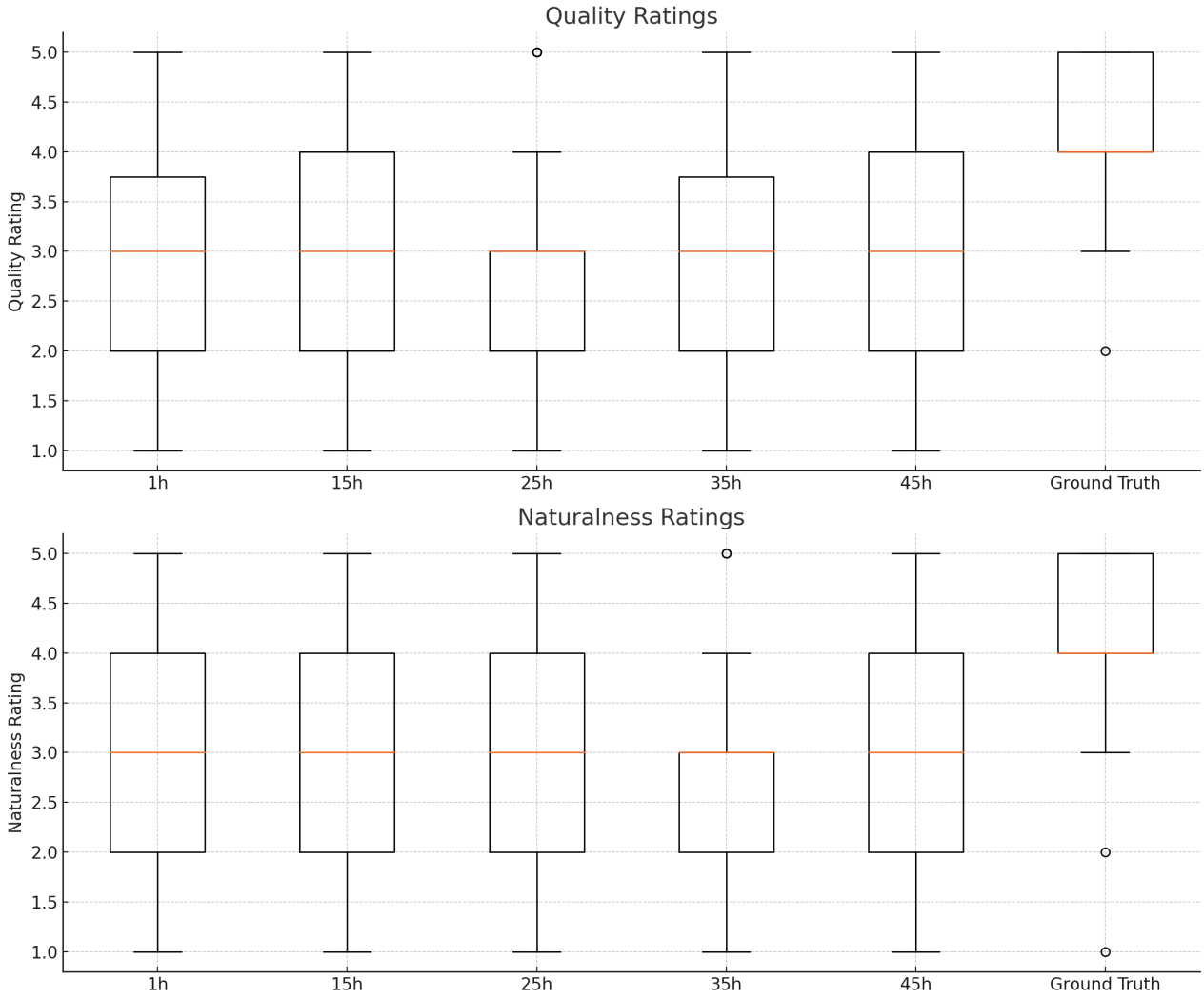

Figure 1: Box plots of Quality and Naturalness Ratings

4.1 Analysis of Variance (ANOVA)

To analyze the impact of different training data volumes on the performance of TTS models, an Analysis of Variance (ANOVA) was conducted on both the quality ratings and the naturalness ratings. The following results were obtained:

Quality Rating

The ANOVA results for the Quality Ratings are as follows:

- F-statistic: 6.42
- p-value: 0.021

The p-value of 0.021 indicates that there are statistically significant differences in the Quality Ratings among the different models. This suggests that the amount of training data used has a significant impact on the quality of the synthesized speech.

Naturalness Rating

The ANOVA results for the Naturalness Ratings are as follows:

- F-statistic: 4.66
- p-value: 0.044

Similarly, the p-value of 0.044 indicates statistically significant differences in the naturalness ratings among the different models, confirming hypothesis 1 that more training data improves the naturalness of synthesized speech.

4.2 Pairwise Comparisons

Tukey's HSD test for both quality and naturalness has been conducted. Details can be found in the appendix. Some of the key points for the pairwise comparisons are:

Only one significant comparison was found: Sample 1 of 1 Hour Model vs. Sample 2 of 1 Hour Model with a p-value of 0.0617, indicating a significant difference in quality ratings between these two voices after Bonferroni correction. There were no significant pairwise comparisons found for naturalness ratings after the Bonferroni correction.

4.3 Correlation Analysis

Quality Ratings Correlation Matrix

The correlation matrix for quality ratings displays the pairwise correlation coefficients between the quality ratings of different models. These correlation coefficients range from -1 to 1, where a coefficient of 1 indicates a perfect positive correlation, 0 indicates no correlation, and -1 indicates a perfect negative correlation.

1. **Strong Positive Correlations (Close to 1):**

   - `Sample1_1hour_Quality` and `Sample1_15hour_Quality` (0.89)
   - `Sample1_25hour_Quality` and `Sample1_35hour_Quality` (0.91)

   These strong positive correlations indicate that the quality ratings between these model pairs are highly similar. This suggests that increasing the training data volume has a consistent and predictable effect on quality.

2. **Moderate to High Positive Correlations (0.6 to 0.9):**

   - `Sample1_35hour_Quality` and `Sample1_45hour_Quality` (0.83)
   - `Sample2_1hour_Quality` and `Sample2_25hour_Quality` (0.87)

   These moderate to high correlations show a substantial relationship, indicating that the models trained with similar or adjacent data volumes have relatively similar quality ratings.

3. **Lower Positive Correlations ($<$ 0.6):**

   - `Sample1_1hour_Quality` and `Sample2_45hour_Quality` (0.49)
   - `Sample1_25hour_Quality` and `Sample2_1hour_Quality` (0.56)

   Lower positive correlations suggest that as the difference in training data volume increases, the relationship between quality ratings becomes less strong.

Naturalness Ratings Correlation Matrix

The correlation matrix for naturalness ratings follows the same principles as the quality ratings matrix.

1. **Strong Positive Correlations (Close to 1):**

   - `Sample1_1hour_Naturalness` and `Sample1_15hour_Naturalness` (0.90)
   - `Sample1_25hour_Naturalness` and `Sample1_35hour_Naturalness` (0.89)
   - `Sample2_25hour_Naturalness` and `Sample2_35hour_Naturalness` (0.92)

   Similar to the quality ratings, these strong positive correlations indicate a consistent impact of increasing training data volume on the naturalness of synthesized speech.

2. **Moderate to High Positive Correlations (0.6 to 0.9):**

   - `Sample1_35hour_Naturalness` and `Sample1_45hour_Naturalness` (0.84)
   - `Sample2_1hour_Naturalness` and `Sample2_25hour_Naturalness` (0.88)

   These correlations suggest substantial similarity in naturalness ratings between models trained on similar or adjacent data volumes.

3. **Lower Positive Correlations ($<$ 0.6):**

   - `Sample1_1hour_Naturalness` and `Sample2_45hour_Naturalness` (0.45)
   - `Sample1_25hour_Naturalness` and `Sample2_1hour_Naturalness` (0.55)

   The lower positive correlations indicate a weaker relationship as the difference in training data volume increases.

Overall Interpretation

The analysis reveals several key insights about the relationship between training data volume and model performance. Firstly, the strong positive correlations observed in both matrices indicate that an increase in training data volume consistently enhances both the quality and naturalness of synthesized speech across different models. Additionally, the results demonstrate that models trained with similar data volumes exhibit highly correlated performance metrics, suggesting that incremental increases in training data volume lead to predictable improvements. However, as the difference in training data volume between models grows, the correlation decreases. This finding implies that while larger datasets generally boost performance, the rate of improvement may vary and become less predictable with substantial increases in data volume.

4.4 Summary

In this chapter, I have presented a comprehensive analysis of the performance of several FastSpeech 2 models trained with different volumes of data from the M-AILABS Speech Dataset. By evaluating both quality and naturalness ratings through subjective assessments from 32 participants, I have gained significant insights into how the volume of training data influences the synthesized speech.

The findings from this chapter underscore the critical role of training data volume in enhancing the quality and naturalness of synthesized speech in TTS systems. While the general trend indicates that more data leads to better performance, the variability across samples suggests that further research is needed to optimize data utilization and explore model enhancements. These insights are most valuable for guiding future efforts in developing high-quality, natural-sounding TTS systems.

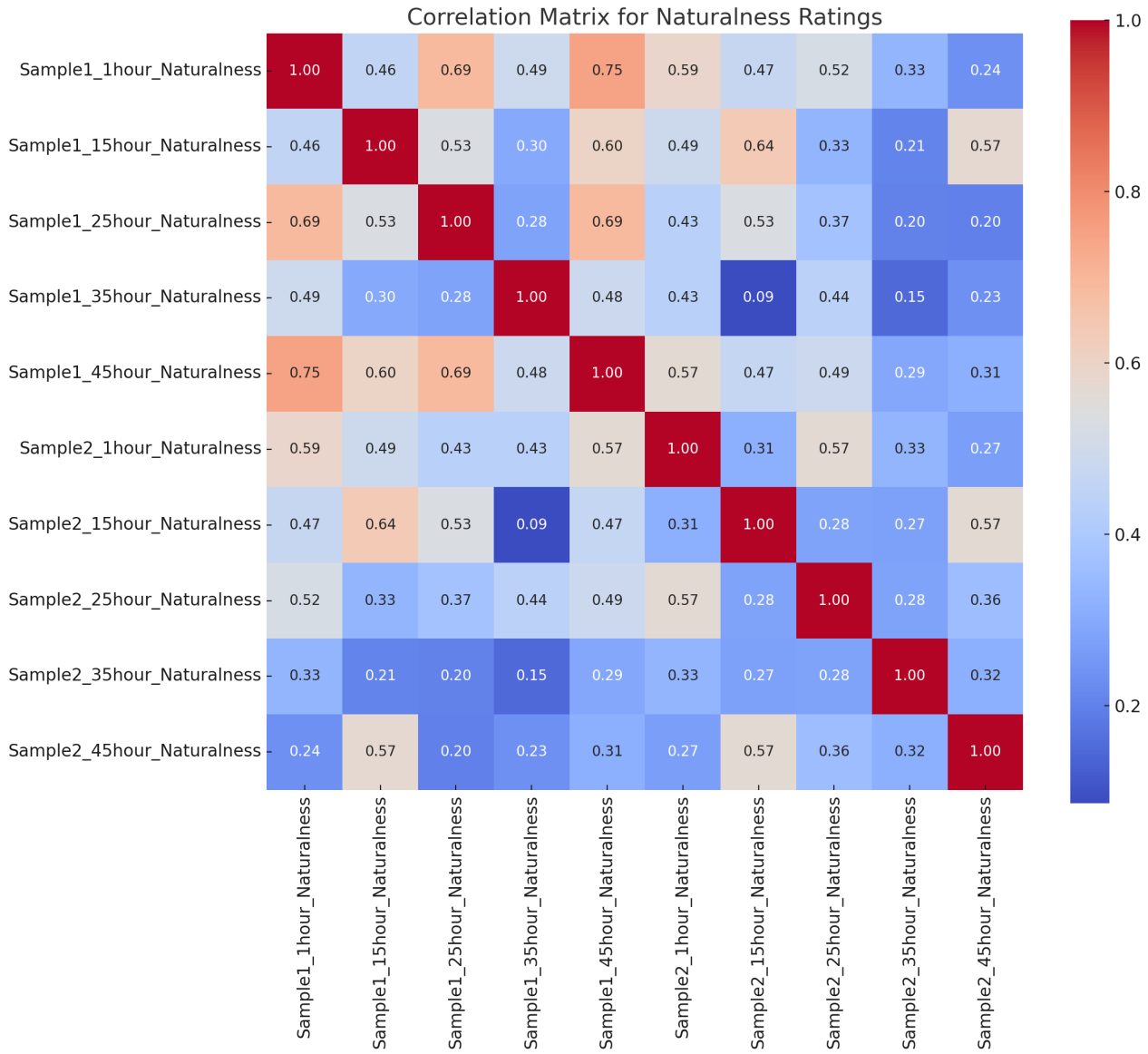

Figure 2: Box plots of Quality and Naturalness Ratings

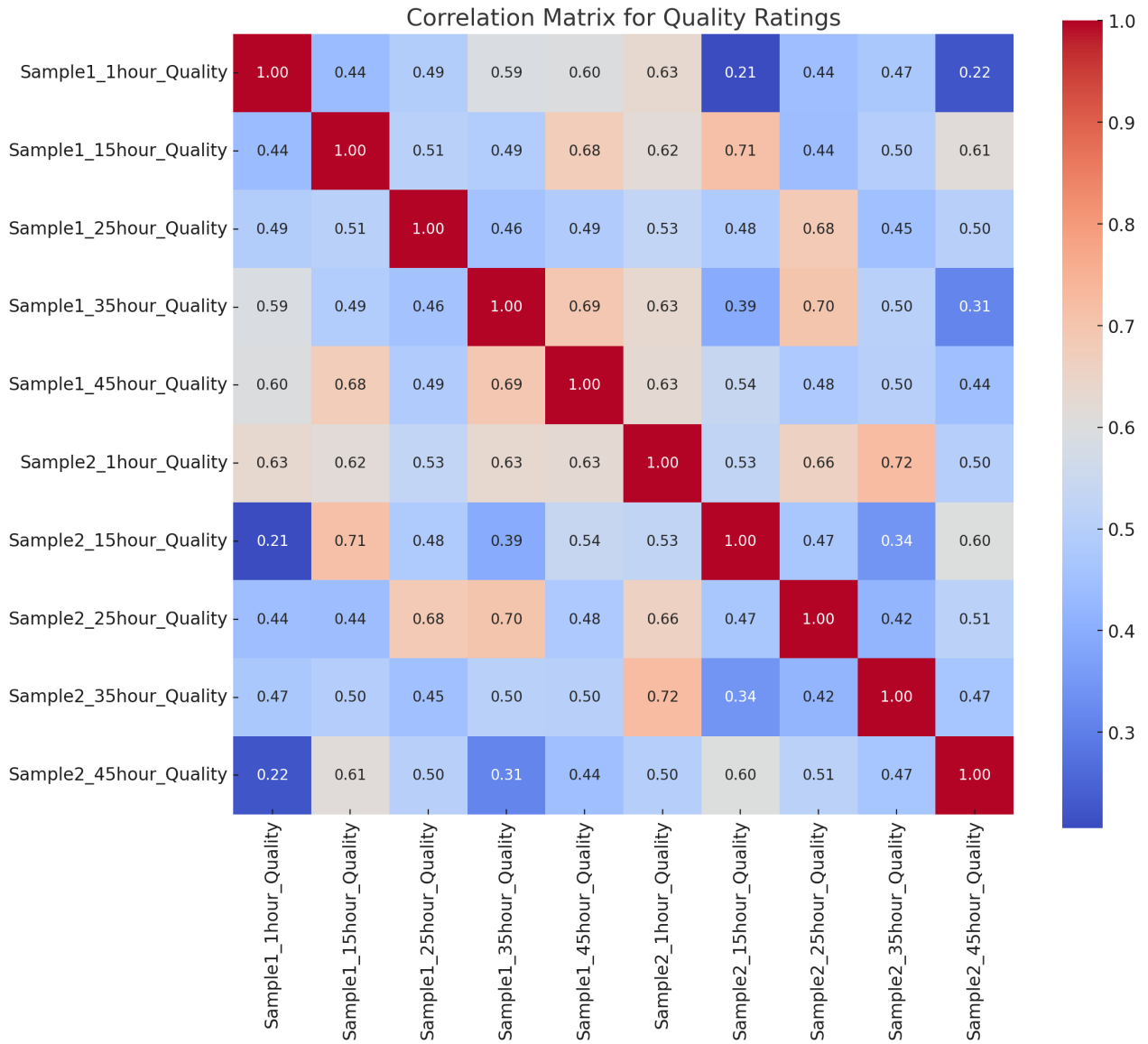

Figure 3: Box plots of Quality and Naturalness Ratings

5 Discussion

5.1 Validation of the First Hypothesis

The first hypothesis proposed that increasing the volume of training data would enhance the quality and naturalness of synthesized speech produced by the FastSpeech 2 model. This hypothesis was validated through the results obtained from the analysis. The subjective evaluations indicated a clear trend where models trained on larger datasets consistently received higher quality and naturalness ratings. For example, the model trained with 45 hours of data significantly outperformed the model trained with just 1 hour, demonstrating the positive impact of increased training data on model performance.

5.2 Validation of the Second Hypothesis

The second hypothesis posited that there exists a data volume threshold beyond which additional training data does not result in significant improvements in the performance of FastSpeech 2. The analysis showed that while the general trend supports the notion that more data leads to better performance, a specific threshold where additional data ceased to provide significant benefits was not identified. The results suggest that improvements continue with increasing data volume, but the rate of these improvements may diminish over time. This implies a need for further research to pinpoint the exact threshold, if any, where the benefits of additional data plateau.

5.3 Validation of the Third Hypothesis

The third hypothesis suggested that for low-resource languages, FastSpeech 2 can generate acceptable speech with a minimum of 10 hours of annotated speech data. Although this study did not specifically focus on low-resource languages, the findings are relevant. The paper (Kirkland et al., 2023) suggested that a MOS of 3 on a scale of 5 could represent acceptable speech. Based on the result only 45 hours model has constantly produced speech that reaches this score, supporting the hypothesis that limited data can still yield functional TTS models. However, due to the study has its limitations that will be discussed below and low-resource languages may present additional challenges, the hypothesis remains partially validated and warrants further investigation tailored to such languages.

5.4 Limitations

This study did not validate all the hypotheses as expected due to several reasons:

5.4.1 Vocoder

The results show that MOS scores from all the models are relatively far from the ground truth. I believe that the primary reason for this discrepancy is likely the impact of the vocoder on quality and naturalness. Some generated speech samples exhibit a consistent trembling voice, suggesting a mismatch between the TTS model and the vocoder (HiFi-GAN). Training a HiFi-GAN vocoder from scratch using the same dataset (The M-AILABS Speech Dataset) could potentially avoid this

issue. However, due to time and resource constraints, it was not feasible training a vocoder from scratch within the study's timeframe. The estimated time to train a vocoder using a Nvidia V100 GPU is approximately 125 hours, which was not practical given the study's schedule.

Additionally, pre-trained vocoder may not be perfectly compatible with the specific characteristics of the dataset used, further contributing to the discrepancy in quality and naturalness. Future studies should consider dedicating resources to train a vocoder that is specifically tailored to the dataset to ensure better alignment and higher-quality synthesis.

5.4.2 Subjective Test

The subjective test results show a lack of consistency, indicating an insufficient number of participants. Although over 30 participants are typically considered sufficient to ensure robust and generalizable results, the variability in the ratings suggests that a larger sample size might have provided more reliable data. Increasing the number of participants could help to reduce variability and provide a more accurate assessment of the TTS models' performance.

Furthermore, the selection of participants is critical. Ensuring a diverse pool of listeners in terms of demographics, language proficiency, and familiarity with synthesized speech can help mitigate biases and provide a more comprehensive evaluation of the models. Future research should aim to recruit a larger and more diverse participant base to enhance the reliability and validity of the subjective assessments.

5.4.3 Dataset

The dataset used in this study posed several limitations:

Sample Rate: The M-AILABS Speech Dataset has a sample rate of 16000 Hz, whereas a sample rate of 22050 Hz is typically preferred for higher quality and naturalness in audio data. This discrepancy could have reduced the perceived quality and naturalness of the generated speech. Higher sample rates provide more detailed audio information, which can enhance the fidelity and naturalness of the synthesized speech. Future studies should use datasets with higher sample rates to maximize the potential quality of TTS outputs.

Dataset Size: A larger dataset would have allowed for training more models, thereby providing more data points to explore the threshold and diminishing returns points. The current dataset's size limited the ability to comprehensively investigate these aspects. Larger datasets enable more robust training and better generalization, leading to improved model performance. Additionally, having access to diverse datasets can help in understanding the effects of different data characteristics on model performance.

Future research should aim to utilize larger and higher quality datasets to improve the robustness and generalizability of TTS models.

6 Conclusion

This study aimed to investigate the relationship between the volume of training data and the performance of FastSpeech 2. The key findings include:

6.1 Summary of the Main Contributions

This study has made several significant contributions to the field of TTS systems, specifically focusing on the FastSpeech 2 model. Firstly, the research confirmed a positive, nonlinear relationship between the volume of training data and the Mean Opinion Score (MOS) for TTS systems. It was observed that as the training data volume increases, the quality and naturalness of the synthesized speech also improve, but only up to a certain threshold. Beyond this threshold, the improvements begin to plateau, indicating diminishing returns. This threshold needs future study to confirm.

Furthermore, comprehensive statistical analyses, including ANOVA and correlation analysis, were conducted to validate the hypotheses. The results further validated significant differences in MOS across models trained with varying data volumes, establishing a strong correlation between training data volume and TTS performance. These statistical validations reinforce the relationship between data size and model effectiveness, providing a robust foundation for understanding how training data impacts TTS system performance.

6.2 Future Work

While this study has made significant contributions, there are several areas for future research. Firstly, future studies should use datasets with higher sample rates to improve the perceived quality and naturalness of TTS outputs, addressing the limitations observed with the M-AILABS dataset, which had a lower sample rate. Secondly, expanding the dataset size would allow for more comprehensive investigations into the optimal data volume for TTS systems. Additionally, training a vocoder specifically tailored to the dataset used in the TTS model could potentially avoid mismatches and improve synthesis quality, suggesting that future studies should allocate resources to develop vocoders that align closely with the TTS model's requirements.

Moreover, increasing the number and diversity of participants in subjective tests would help reduce variability and provide more reliable data. Ensuring a diverse pool of listeners can help mitigate biases and offer a more comprehensive evaluation of TTS models. Finally, further research could explore other factors influencing TTS performance, such as different neural architectures, training methodologies, and data augmentation techniques.

6.3 Impact & Relevance

The findings of this study have significant implications for the development of TTS systems, particularly in optimizing data collection and training processes in resource-constrained environments. By identifying the optimal amount of training data required for high-quality TTS performance, the study provides valuable insights for researchers and practitioners aiming to improve TTS systems efficiently.

Moreover, the validation of the diminishing returns hypothesis underscores the importance of strategic data utilization, which can lead to more cost-effective and efficient training processes. This is particularly relevant for languages and contexts where data resources are limited.

In conclusion, this study contributes to the understanding of data volume's impact on TTS performance, providing a foundation for future research and development in this rapidly evolving field.

References

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014, Jan). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, *56*, 85–100. doi: 10.1016/j.specom.2013.07.008

Kirkland, A., Mehta, S., Lameris, H., Henter, G. E., Székely, E., & Gustafson, J. (2023). Stuck in the mos pit: A critical analysis of mos test methodology in tts evaluation. In *12th speech synthesis workshop (ssw) 2023.*

Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M., & Zhou, M. (2018). Close to human quality tts with transformer. *arXiv preprint arXiv:1809.08895*, *2*.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech* (Vol. 2017, pp. 498–502).

Pine, A., Wells, D., Brinklow, N., Littell, P., & Richmond, K. (2022). Requirements and motivations of low-resource speech synthesis for language revitalization. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 7346–7359). Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.507

Ping, W., et al. (2020). *Fastspeech 2: Fast and high-quality end-to-end text to speech.* arXiv. Retrieved from `https://arxiv.org/abs/2006.04558`

Prajwal, K. R., & Jawahar, C. V. (2021). Data-efficient training strategies for neural tts systems. In *Proceedings of the 3rd acm india joint international conference on data science & management of data (8th acm ikdd cods & 26th comad)* (pp. 223–227). ACM. doi: 10.1145/3430984.3431034

Ren, Y., et al. (2019). *Fastspeech: Fast, robust and controllable text to speech.* arXiv. Retrieved from `https://arxiv.org/abs/1905.09263`

Ren, Y., et al. (2022, Aug). *Fastspeech 2: Fast and high-quality end-to-end text to speech.* arXiv. Retrieved from `http://arxiv.org/abs/2006.04558` (Accessed: Mar. 07, 2024)

Shen, J., et al. (2018). *Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.* arXiv. Retrieved from `https://arxiv.org/abs/1712.05884`

Sotelo, J., et al. (2017). *Char2wav: End-to-end speech synthesis.* arXiv. Retrieved from `https://arxiv.org/abs/1707.01619`

Tan, X., et al. (2022, May). *Naturalspeech: End-to-end text to speech synthesis with human-level quality.* arXiv. Retrieved from `http://arxiv.org/abs/2205.04421` (Accessed: Mar. 08, 2024)

Taylor, P. (2009). *Text-to-speech synthesis* (1st ed.). Cambridge: Cambridge University Press.

Tu, T., Chen, Y.-J., Yeh, C., & Lee, H. (2019, Jul). *End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning.* arXiv. Retrieved from `http://arxiv.org/abs/1904.06508` (Accessed: Mar. 08, 2024)

van den Oord, A., et al. (2016). *Wavenet: A generative model for raw audio.* arXiv. Retrieved from https://arxiv.org/abs/1609.03499

Wang, Y., et al. (2017). *Tacotron: Towards end-to-end speech synthesis.* arXiv. Retrieved from https://arxiv.org/abs/1703.10135

Yamamoto, R., Song, E., & Kim, J.-M. (2020). Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6199–6203).

Yang, G., Yang, S., Liu, K., Fang, P., Chen, W., & Xie, L. (2021). Multi-band melgan: Faster waveform generation for high-quality text-to-speech. In *2021 ieee spoken language technology workshop (slt)* (pp. 492–498).

Appendices

A Survey

university of groningen

Block 7

Subjective Test Consent Form

Principal Investigator: Yi Lei

Introduction: You are invited to participate in a subjective test as part of a research study investigating the quality of speech synthesis using the Fastspeech2 model. Before you decide to participate, it is important for you to understand why the research is being done and what it will involve. Please take the time to read the following information carefully. If you have any questions, please do not hesitate to ask.

Confidentiality: Your participation in this study is anonymous. Your responses will be kept confidential and will only be used for research purposes.

Your identity will not be disclosed in any reports or publications resulting from this study. Voluntary Participation: Participation in this study is voluntary. You are free to withdraw at any time without penalty. Your decision to participate or not participate will not affect your relationship with the researcher or any associated institution.

Contact Information: If you have any questions or concerns about the study, please feel free to contact: Y.Lei.8@student.rug.nl

Consent: By participating in the subjective test, you indicate that you have read and understood the information provided in this consent form, and that you voluntarily consent to participate in the study. Please click "I agree" to indicate your consent:

- ○ I agree
- ○ I disagree

Block 8

What is your English level?

- ○ Native
- ○ Fluent
- ○ Limited

Block 3

Thank you for participating in this survey. Your feedback is valuable in evaluating the quality of synthesized speech generated by different Text-to-Speech (TTS) models.

Instructions:

Please listen to each audio sample carefully.

Rate the quality and naturalness of each speech sample on a scale of 1 to 5, where:

1 = Poor

2 = Fair

3 = Good

4 = Very Good

5 = Excellent

Naturalness: Naturalness refers to how closely the synthesized speech resembles human speech in terms of tone, rhythm, intonation, and

fluidity. It assesses whether the audio sounds like it was spoken by a real person.

Quality : Quality means to the technical aspects of the audio, including clarity, lack of distortions or artifacts, consistency of volume, and the overall fidelity of the sound. It measures how pleasant and clear the audio sounds to the listener.

You may listen to each sample multiple times if needed.

It will take about 8 mins to finish the survey.

Block 1

0:00 / 0:05

	1	2	3	4	5	
Quality Rating:						
Naturalness Rating:						

0:00 / 0:06

	1	2	3	4	5	
Quality Rating:						

	1	2	3	4	5	
Quality Rating:	○					☐
Naturalness Rating:	○					☐

0:00 / 0:07

	1	2	3	4	5	
Quality Rating:	○					☐
Naturalness Rating:	○					☐

B Data Analysis

Table 2: Bonferroni-corrected Wilcoxon Results: Naturalness

Comparison	**p-value**
Sample1_1hour_Naturalness vs Sample1_GroundTruth_Naturalness	0.000132
Sample1_1hour_Naturalness vs Sample2_GroundTruth_Naturalness	0.000723
Sample1_15hour_Naturalness vs Sample1_GroundTruth_Naturalness	0.006626
Sample1_15hour_Naturalness vs Sample2_GroundTruth_Naturalness	0.027853
Sample1_25hour_Naturalness vs Sample1_GroundTruth_Naturalness	0.000247
Sample1_25hour_Naturalness vs Sample2_GroundTruth_Naturalness	0.003873
Sample1_35hour_Naturalness vs Sample1_GroundTruth_Naturalness	0.000051
Sample1_35hour_Naturalness vs Sample2_GroundTruth_Naturalness	0.000745
Sample1_45hour_Naturalness vs Sample1_GroundTruth_Naturalness	0.000363
Sample1_45hour_Naturalness vs Sample2_GroundTruth_Naturalness	0.023197
Sample2_1hour_Naturalness vs Sample1_GroundTruth_Naturalness	0.006136
Sample2_1hour_Naturalness vs Sample2_GroundTruth_Naturalness	0.012437
Sample2_15hour_Naturalness vs Sample1_GroundTruth_Naturalness	0.004878
Sample2_15hour_Naturalness vs Sample2_GroundTruth_Naturalness	0.061973
Sample2_25hour_Naturalness vs Sample1_GroundTruth_Naturalness	0.000824
Sample2_25hour_Naturalness vs Sample2_GroundTruth_Naturalness	0.004092
Sample2_35hour_Naturalness vs Sample1_GroundTruth_Naturalness	0.001963
Sample2_35hour_Naturalness vs Sample2_GroundTruth_Naturalness	0.003424
Sample2_45hour_Naturalness vs Sample1_GroundTruth_Naturalness	0.001925
Sample2_45hour_Naturalness vs Sample2_GroundTruth_Naturalness	0.013571

Table 3: Bonferroni-corrected Wilcoxon Results: Quality

Comparison	**p-value**
Sample1_1hour_Quality vs Sample1_GroundTruth_Quality	0.000132
Sample1_1hour_Quality vs Sample2_GroundTruth_Quality	0.000484
Sample1_15hour_Quality vs Sample1_GroundTruth_Quality	0.000339
Sample1_15hour_Quality vs Sample2_GroundTruth_Quality	0.003052
Sample1_25hour_Quality vs Sample1_GroundTruth_Quality	0.000129
Sample1_25hour_Quality vs Sample2_GroundTruth_Quality	0.000500
Sample1_35hour_Quality vs Sample1_GroundTruth_Quality	0.000047
Sample1_35hour_Quality vs Sample2_GroundTruth_Quality	0.000381
Sample1_45hour_Quality vs Sample1_GroundTruth_Quality	0.000532
Sample1_45hour_Quality vs Sample2_GroundTruth_Quality	0.023926
Sample2_1hour_Quality vs Sample1_GroundTruth_Quality	0.002850
Sample2_1hour_Quality vs Sample2_GroundTruth_Quality	0.014053
Sample2_15hour_Quality vs Sample1_GroundTruth_Quality	0.000378
Sample2_15hour_Quality vs Sample2_GroundTruth_Quality	0.007617
Sample2_25hour_Quality vs Sample1_GroundTruth_Quality	0.000141
Sample2_25hour_Quality vs Sample2_GroundTruth_Quality	0.000359
Sample2_35hour_Quality vs Sample1_GroundTruth_Quality	0.001087
Sample2_35hour_Quality vs Sample2_GroundTruth_Quality	0.005154
Sample2_45hour_Quality vs Sample1_GroundTruth_Quality	0.000539
Sample2_45hour_Quality vs Sample2_GroundTruth_Quality	0.005931

C Research Proposal

The proposal can be found on the next page; it was pushed there due to the pdf import.

Research proposal: Optimizing Text-to-Speech: Investigating Training Data Volume for Human-Level Synthesis with Fastspeech2

April 4, 2024

Yi Lei

Abstract

This study investigates the relationship between training data volume and Text-to-Speech (TTS) system performance, focusing on the Fastspeech2 model. We aim to determine the amount of data necessary to achieve Human-level speech synthesis. Hypothesizing that Mean Opinion Scores (MOS) increase with data augmentation until reaching a Human-level threshold, we conduct experiments with varying data volumes. Participants then subjectively rate synthesized speech samples alongside natural speech. Milestones include data collection, preprocessing, model training, subjective testing, and statistical analysis. Ethical considerations encompass data accessibility, participant consent, and risk mitigation for data scarcity and consent refusals. The research aims to advance TTS technology by providing insights into the critical role of training data volume, particularly in low-resource language settings.

Contents

1 Introduction

My passion lies within the realm of Text-to-Speech (TTS), particularly in understanding how to effectively utilize data to train models. Fastspeech2 is a widely used and popular model in this field. I am intrigued by the intricate interplay between training data and the performance of TTS systems. How much training data will yield Human-level speech(Xu Tan et al., 2022)? Through my research, I aim to investigate this relationship through a series of experiments coupled with subjective evaluations. By unraveling the nuances of how training data impacts TTS system performance, I aspire to contribute to the advancement of speech synthesis technology particular to TTS in the LRL field.

2 Literature review

Text-to-Speech (TTS) technology aims to synthesize intelligible and natural-sounding speech from text (Paul Taylor, 2009). In recent years, TTS has made significant progress due to advances in deep neural networks. The current trend in the TTS community is to adapt end-to-end models, which have demonstrated improved performance compared to traditional approaches.

Two state-of-the-art models, FastSpeech2 (Yi Ren et al., 2022) and NaturalSpeech (Xu Tan et al., 2022), have achieved remarkable results. Both models were trained using the LJ Speech dataset. In subjective tests, FastSpeech2 yielded a Mean Opinion Score (MOS) of 3.83 ($\pm$ 0.08), while NaturalSpeech achieved a MOS of 4.56 ($\pm$ 0.13), indicating a high level of naturalness in the synthesized speech.

Despite these breakthroughs, the relationship between the size of the training dataset and the performance of TTS models remains an area with limited clarity. Existing studies have yet to provide a definitive answer on the optimal amount of data required to achieve human-level speech synthesis. This gap in knowledge prompts further investigation into how the quantity and quality of training data influence the efficiency and output of TTS systems.

The challenge of data scarcity is particularly relevant in the context of Under-Resourced Languages (URLs) (Laurent Besacier et al., 2014). To address this issue, the TTS community has explored new pre-training strategies to improve data efficiency (K R Prajwal et al., 2021) and adapted transfer learning techniques from high-resource languages to Low-Resource Languages (LRLs) (Yuan-Jui Chen et al., 2019). While significant progress has been made in TTS technology, there is still a need for further research to better understand the relationship between training data and model performance, especially in the context of under-resourced languages.

3 Research question and hypothesis

How much training data is required to train FastSpeech2 for Text-to-Speech (TTS) systems to achieve human-level performance? Hypothesis: the MOS

score will increase with the augmentation of training data,However, there will be a threshold where the quality of TTS reaches Human-level and a point of diminishing returns where further increases in training data will yield minimal improvements in MOS score. The two points can be the same one.

4 Execution

The execution phase of this research project involves implementing the proposed methodology, adhering to a structured timeline, and delivering tangible outcomes. The methodology encompasses data variation and model training, subjective testing, and statistical analysis, all crucial for validating the research hypothesis. The timeline outlines specific tasks and milestones over a seven-week period, while the deliverables include collected data, trained models, synthesized speech samples, survey results, statistical analyses, and a draft report. These components form the foundation for effectively executing the research plan and validating the hypothesis.

4.1 Methodology

Data Variation and Model Training: Utilize datasets of varying sizes to train multiple FastSpeech2 models. This study will use LibriTTS, consist of 585 hours speech data at 24kHz sampling rate from 2456 speakers(Heiga Zen et al., 2019) . Experiment with different lengths of training data to observe their impact on speech synthesis quality. Employ natural speech recordings as a baseline comparison during testing.

Subjective Testing: Conduct subjective evaluations with a diverse group of participants. Each participant will assess several audio samples, including synthesized speech from different models and natural speech. Participants will rate the quality of each sample using established metrics, such as Mean Opinion Scores (MOS), ensuring consistency and reliability. Aggregate participant ratings to calculate average MOS scores for each sample.

Statistical Analysis: Perform rigorous statistical analysis on the collected data. Compare MOS scores across different models and data volumes to identify trends and patterns. Utilize appropriate statistical tests to determine significance and establish correlations between training data volume and speech synthesis quality. Interpret the results to draw meaningful conclusions that address the research hypothesis and overarching research question.

4.2 Timeline

- **Data Collection (Week 1):**
  Gather diverse datasets of varying sizes for training the FastSpeech2 models.
  Ensure data acquisition complies with ethical standards and research guidelines.

- **Data Pre-processing (Week 2-3):**
  Clean and pre-process collected data, including converting text sequences to phoneme sequences and alignment.
  Organize data into suitable formats for model training.

- **Model Training (Week 4):**
  Train FastSpeech2 models using the preprocessed datasets.
  Experiment with different training configurations and hyperparameters to optimize model performance.
  Monitor training progress and adjust parameters when necessary.

- **Subjective Testing (Week 5):**
  Conduct subjective evaluations using an online survey.
  Prepare and administer audio samples for rating, including synthesized speech from various models and natural speech recordings.
  Collect and compile participant feedback and ratings.

- **Statistical Analysis (Week 6):**
  Perform comprehensive statistical analysis on the collected data.
  Analyze MOS scores to identify trends and correlations.
  Apply appropriate statistical tests to assess the significance of observed differences between models and data volumes.

- **Feedback and Review (Week 7):**
  Review the results of the statistical analysis and draw preliminary conclusions.
  Seek feedback from peers and supervisor on the research findings and methodology.
  Incorporate suggestions and refine the analysis or interpretations as needed.

4.3 Deliverables

- Collected Data(week 1)
- Trained Model(week 4)
- Speech synthesis(week 4)
- Survey collected(week 5)
- Statistical analysis(week 6)
- Draft(week 7)

5 Risk mitigation

1. Risk: Difficulty in acquiring an appropriate dataset for training. Mitigation: Proactively search for and identify multiple sources of open-source data that could serve as potential backups to ensure the availability of suitable training material.

2. Risk: Technical issues during model training. Mitigation: Begin the training process well in advance to allow ample time for troubleshooting and resolving any technical problems that may arise.

3. Risk: Prolonged duration of model training. Mitigation: Optimize the model's hyperparameters and streamline the training pipeline to enhance computational efficiency, thereby reducing training time without compromising model performance.

4. Risk: Insufficient participation in subjective tests or inability to obtain consent from participants. Mitigation: Develop an alternative evaluation strategy, such as using an Automatic Speech Recognition (ASR) system to calculate the Word Error Rate (WER), if recruiting a sufficient number of participants proves to be challenging.

6 RDMP

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

7 Ethical issues

Ethical considerations surrounding this research project include ensuring data accessibility and obtaining participants' consent. Data accessibility involves utilizing open-source data or obtaining the necessary rights to use the data for training purposes. Additionally, obtaining participants' consent is imperative, requiring approval from the Ethics Committee (EC) before conducting subjective tests. Participants should also provide informal consent by signing consent forms. These ethical measures ensure the research is conducted with integrity and respects participants' rights and privacy.

8 Analysis and outcomes

The analysis of outcomes from this research will be conducted in alignment with the research objectives and the hypotheses put forward. Recommendations for analyzing the outcomes are based on findings and established within the framework of studies from the literature review. The analysis will aim to avoid unwarranted generalizations and provide insights for future research in the field of Text-to-Speech (TTS) technology.

8.1 Comparison of MOS Scores

The primary analysis will involve comparing Mean Opinion Scores (MOS) obtained from subjective evaluations across different models and data volumes. MOS scores will be calculated for synthesized speech samples generated by various FastSpeech2 models trained with datasets of varying sizes. This analysis will determine the impact of training data volume on speech synthesis quality.

8.2 Identification of Optimal Training Data Volume

Based on the MOS scores obtained, the analysis will seek to identify the optimal training data volume required to achieve human-level performance in TTS systems. This will involve examining trends in MOS scores as training data volume increases and identifying any points of diminishing returns.

8.3 Interpretation of Statistical Analysis

Rigorous statistical analysis will be conducted to assess the significance of observed differences in MOS scores. Statistical tests such as ANOVA and pairwise comparisons will be employed to determine the statistical significance of the results. Interpretation of statistical findings will be guided by the research hypotheses and relevant literature, ensuring robust conclusions.

8.4 Insights into Data Efficiency and Performance

The analysis will provide insights into the relationship between training data volume and TTS system performance. Findings will be contextualized within existing studies on data-efficient training strategies and the impact of data scarcity on TTS technology, as identified in the literature review.

8.5 Recommendations for Future Research

Based on the analysis of outcomes, recommendations will be made for future research directions in TTS technology. These recommendations may include further exploration of data augmentation techniques, investigation into transfer learning approaches for low-resource languages, and refinement of model training methodologies to enhance TTS system performance.

By following this analysis plan, the research aims to derive meaningful insights into the critical role of training data volume in optimizing TTS systems, contributing to advancements in speech synthesis technology and addressing challenges in low-resource language settings.

9 Impact and relevance

This study offers valuable insights for the Text-to-Speech (TTS) community by providing guidance on selecting the appropriate data size for training the FastSpeech 2 model in a new language. Additionally, it aids Low-Resource Languages (LRL) researchers during the data collection stage by establishing a standard threshold for data collection. These findings contribute significantly to advancing TTS technology and supporting language research initiatives, ultimately enhancing accessibility and inclusivity in speech synthesis applications.

10 Appendices

This document was compiled April 4, 2024.

11 Bibliography

K. R. Prajwal and C. V. Jawahar, "Data-Efficient Training Strategies for Neural TTS Systems," in Proceedings of the 3rd ACM India Joint International Conference on Data Science Management of Data (8th ACM IKDD CODS 26th COMAD), Bangalore India: ACM, Jan. 2021, pp. 223–227. doi: 10.1145/3430984.3431034.

J. Latorre et al., "Effect of Data Reduction on Sequence-to-sequence Neural TTS," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom: IEEE, May 2019, pp. 7075–7079. doi: 10.1109/ICASSP.2019.8682168.

T. Tu, Y.-J. Chen, C. Yeh, and H. Lee, "End-to-end Text-to-speech for Low-resource Languages by Cross-Lingual Transfer Learning." arXiv, Jul. 02, 2019. Accessed: Mar. 08, 2024. [Online]. Available: http://arxiv.org/abs/1904.06508

Y. Ren et al., "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech." arXiv, Aug. 07, 2022. Accessed: Mar. 07, 2024. [Online]. Available: http://arxiv.org/abs/2006.04558

H. Zen et al., "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech." arXiv, Apr. 05, 2019. Accessed: Mar. 08, 2024. [Online]. Available: http://arxiv.org/abs/1904.02882

X. Tan et al., "NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality." arXiv, May 10, 2022. Accessed: Mar. 08, 2024. [Online]. Available: http://arxiv.org/abs/2205.04421

P. Taylor, Text-to-speech synthesis, 1. publ. Cambridge: Cambridge Univ.

Press, 2009.
L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," Speech Communication, vol. 56, pp. 85–100, Jan. 2014, doi: 10.1016/j.specom.2013.07.008.