Child Speech Synthesis using Deep Learning

MSc Research Project
Data Analytics

Zeba Siddique
Student ID: x20227086

School of Computing
National College of Ireland

Supervisor: Dr. Abubakr Siddig

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Zeba Siddique
Student ID:	x20227086
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr. Abubakr Siddig
Submission Due Date:	15/08/2022
Project Title:	Child Speech Synthesis using Deep Learning
Word Count:	9266
Page Count:	29

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Zeba Siddique
Date:	17th September 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Child Speech Synthesis using Deep Learning

Zeba Siddique
x20227086

Abstract

Speech synthesis or text-to-speech automatically translates the input text into spoken speech. TTS technologies have progressively increased to generate synthesized speech that is intelligible, natural, and human-sounding. Whilst extensive research on TTS has been conducted using the adult speech corpus, little or minimum investigation has been done on the synthesis of child speech. This study proposes a TTS model comprising of a pre-trained and fine-tuned speaker encoder and Tacotron 2 synthesizer, along with a HiFi-GAN neural vocoder model that involves a transfer-learning approach to synthesize child speech. A publicly available multi-speaker child speech corpus was cleaned and a pre-processed subset of 1-hour data was utilized for training and fine-tuning the proposed TTS model. The quality of the synthesized child's speech was evaluated using the MOSNet score. The best quality of synthesized child speech achieved an average MOSNet score of 2.8 for the fine-tuned HiFi-GAN vocoder. The proposed TTS model could generate synthesized child speech on adoption of the transfer-learning approach.

Keywords— Speech synthesis, text-to-speech, multi-speaker, Tacotron 2, HiFi-GAN, MOSNet

1 Introduction

1.1 Background

Speech synthesis or Text-to-speech (TTS) research has gained notoriety in terms of generating high-quality human-sounding synthesized speech. Speech synthesis enables the artificial production of human speech through machines and devices. The first end-to-end TTS system was developed in 1968 which produced monotonous yet intelligible speech (Flanagan et al.; 1970). Recent TTS systems are based on more effective procedures rather than being fundamental. From aiding communication with people with disabilities to assisting in digital instructions during teaching, there is a multitude range of applications for speech synthesis technology. Synthesizing human-sounding intelligible and natural speech is exciting and an essential area of research.

Research advancements in the field of human speech synthesis have largely focused on deep learning mechanisms to develop a top-notch text-to-speech system to produce synthetic voice. However, most of this research was conducted using different adult speech corpora, thereby completely abandoning the system's behavior for child speech. This led to formulating the purpose of this research to synthesize child speech in the original speaker's voice using deep learning technologies.

1.2 Motivation

Speech synthesis or text-to-speech system is a framework that aims to produce speech signals from a given input of language representation. The text-to-speech systems initially produced synthesized voices that were robotic. Over the course of time and with advancements in technology, the text-to-speech system was able to produce natural-sounding synthetic speech. The focus of research in the speech synthesis domain is to obtain synthetic speech that closely resembles the speech pattern of the original speaker.

Text-to-speech systems have applications in a variety of domains. The synthesized voice generated from a text-to-speech system can be used in read-aloud digital assistive technology and computerized teaching. The synthesized voice can also be used in the entertainment industry, for generating voice-overs in the original actor's voice. A considerable amount of research on speech synthesis is majorly conducted on adult speech, and a minimal investigation is done on child speech corpus (Jain et al.; 2022) . This formulates the need to investigate a text-to-speech system capable of synthesizing child speech.

Humans communicate effectively through verbal communication. Verbal communication comprises speech that is produced due to coordinated muscular movements representing different characteristics like amplitude, pitch, loudness, etc. Human speech evolves with age. The speech of an adult is more precise and well-formulated whereas the speech of a child is inarticulate. The vocal track and fold of a child are smaller as compared to that of an adult, thus largely differing in the fundamental frequency response (Jain et al.; 2022). Therefore, the synthesis of child speech through a text-to-speech system requires notable attention.

This research focuses on experimenting with the My Science Tutor (MyST) children's speech corpus that consists of 393 hours of child speech data. Using the MyST corpus and combining the best performing deep learning models, a text-to-speech system is built that may be utilized for generating synthesized child speech in the original speaker's voice. Despite numerous research using adult speech data being done in this field, a text-to-speech model catering to the child speech data is worth exploring. To build and bring a robust TTS model into real-world application, a thorough analysis of existing works and research theories is required.

1.3 Research Question

How well a deep learning based text-to-speech model synthesize the speech of a child in the original speaker's voice?

1.4 Objectives

The main objectives of the research proposal are as follows:

1. To build a text-to-speech model using deep learning technologies.

2. To understand the capabilities of the built TTS model when trained purely on the child's speech corpus.

3. To conduct experiments and assess the applicability of the built model.

1.5 Plan of the paper

The rest of the paper is structured as follows: Section 2 of the paper outlines the previous research carried out in the speech synthesis field. Section 3 gives a detailed description of the research methodology. Section 4 describes the suggested approach in detail. Section 5 describes the implementation details and specifications. Section 6 discusses the findings and observations supporting the research question and evaluation. Section 7 condenses the study and provides proposals for future work useful for further conducting the research.

2 Related Work

Speech synthesis or text-to-speech produces synthesized speech given an input text. Machine interaction with a voice as the communication medium has become increasingly beneficial, the need to develop a text-to-speech system that produces high-quality synthesized speech has become a necessity. Generating natural, human-sounding, and intelligible speech has been the prime focus of research in the speech synthesis domain. Carlson (1995) largely classified developments of a TTS system into different categories. The synthesized speech generated from a TTS model was also classified, which as per Tiomkin et al. (2011) revealed that a TTS model that used pre-trained statistical methods produced barely audible speech of degraded quality. As a speech signal is best represented in the form of a spectrogram, the TTS models are further classified as auto and non-auto regressive.

2.1 Autoregressive TTS Model

The TTS model maintains long-term dependency by generating speech signals of future conditioning based on past speech signals. The ideology of an autoregressive model is largely based on regressed time series. These models are known to amplify and propagate errors but are defined and trained easily.

Oord et al. (2016) introduced the WaveNet model which could grasp speaker characteristics only from the raw audio waveform. The model produced audio forms where each speech signal was conditioned on the previous signal and had a distribution that was predictive. Experiments on the WaveNet model revealed that the produced synthetic speech from a single speaker corpus was significantly natural with an uncanny prosody element. The experiments of Manzelli et al. (2018) to explore the usage of WaveNet for audio modeling applications revealed that the synthesized audio from the WaveNet model was noisy and lacked generalizability post training on electronic music that was multi-track in nature. Sercan et al. (2017) constructed Deep Voice (DV1) which was entirely a deep neural network model that used WaveNet's variant as the synthesizer. The frequency and phonemes extracted from the audio and its transcripts were used by DV1 to efficiently generate synthesized speech. The only drawback of DV1 identified by the author was the lack of end-to-end training capabilities.

Wang et al. (2017) introduced Tacotron (T1), an end-to-end TTS model trainable only on text-audio pairs thereby producing audio post converting it from raw spectrogram frames. Tacotron is a centered model that uses attention and sequence-to-sequence generative frame mechanisms to produce synthesized speech. The model achieved remarkable results high in naturalness. Arik et al. (2017) introduced a multi-speaker neural network model called as Deep Voice 2 (DV2). As DV1 and T1 both were suitable for a single-speaker corpus, DV2 focused on catering to a multi-speaker corpus and was based on the combined ideology of DV1 and T1. The synthesized voice generated from DV2 had distinguishable characteristics for different speakers, but the spectrogram displayed errors due to the erroneous design of the T1 and WaveNet vocoder combination. Ping et al. (2017) developed Deep Voice 3 (DV3) enabling full parallel computation to address errors identified in DV1 and DV2 models. DV3 depicted faster computation and training capabilities due to the use of recurrent cells.

Shen et al. (2017) developed Tacotron 2 (T2) by combining the mel spectrogram generating Tacotron T1 and the WaveNet vocoder instead of the traditional Griffin-Lim vocoder. Apart from human naturalness, the prosody of T1 and audio quality of WaveNet were observed in the synthesized speech generated from T2. Kalchbrenner et al. (2018) introduced a single-layered recurrent neural network along with a dual softmax layer sequential model called WaveRNN. The model achieved high-quality speech generated on low-end resources such as mobile CPUs.

2.2 Non-autoregressive TTS Model

The non-autoregressive models majorly use a parallel spectrogram to speed up the inference process. The training process of these models is guided using a well-trained teacher model.

Ren et al. (2019) introduced a feed-forward transformer structured controllable neural text-to-speech model called FastSpeech that uses teacher model architecture for training and considers the predicted mel spectrograms as the ground truth. The model was capable of stabilizing the training loss but produced a lower quality synthesized speech. Oord et al. (2018) introduced Parallel WaveNet, a probability density distillation neural network model that parallelly generated the best features of the Inverse autoregressive flow (IAF) and WaveNet. The model was based on the ideology of the student-teacher model to match the probability of samples from the distributions. The Parallel WaveNet model demonstrated capabilities of being adopted for a multi-speaker corpus for a variety of other languages. Ping et al. (2018) developed another model called ClariNet, a text-to-wave fully convolutional neural speech synthesis based on the Deep Voice 3 architecture. The model explored the feasibility of WaveNet of Gaussian distribution for raw audio waveform modeling.

Prenger et al. (2019) introduced WaveGlow capable of maximizing the likelihood of training data only from a single network in contrast to ClariNet and Parallel WaveNet which were based on complex networked architecture. Additionally, WaveGlow achieved a stable and simple training procedure with a single cost function. Miao et al. (2020) developed a single feed-forward network capable of learning the alignment amongst text given as input and the spectrogram generated. The performance of WaveGlow surpassed the performance of Tacotron 2 and could preserve the original audio's characteristics. Ren et al. (2020) developed a model called FastSpeech 2 (FS 2) to address the shortcomings of the FastSpeech (FS 1) model. This developed model utilized the speaker characteristics and conditioned it for inputs and trained the model on ground truth directly. FS2 drew inferences faster as compared to FS1 and achieved a training speed three times greater than FS1. Further classification of text-to-speech systems is based on the amount of training data required.

2.3 Large amount of data

Koffi (2022) described the steps of building a TTS model for a language other than English and concludes that for a model to generate natural-sounding synthetic speech requires collecting the right amount of data and linguistic information. Chen et al. (2022) proposed a three-layered Bi-LSTM model that first converts the electroglottograph signals to text and used Tacotron 2 to generate synthetic speech. This model was developed to synthesize speech for individuals who no longer have speech capabilities. The research largely contributed to the feasibility of a TTS model in the production of synthesized voice using the EGG signals. Valin et al. (2022) proposed LPCNet that could generate real-time synthesis using linear production. LPCNet reduced the computational requirements by two and a half times. Saeki et al. (2022) developed a TTS model using the regularization method. The system was trained corpus containing environmental distortions and noise. Evaluation of the developed model was assessed on datasets of different qualities; results reveal the developed TTS model outperformed the existing clatter robust systems. Lei et al. (2022) developed a sequence-to-sequence attention-based model called MsEmoTTS that could control emotional expressions and used versions of Tacotron and Global Style Tokens to comprehend the learning style representations. The results revealed dissimilar efficiencies on different datasets. Das et al. (2022) analyzed the quality of synthesized speech using VQ-VAE to evaluate multilingual switch for an utterance. Results depict that synthesized speech of slow speakers tends to be more natural as compared to fast speakers. Xue et al. (2022) focused on developing a speaker encoder model that could efficiently obtain speaker characteristics to synthesize speech in the target speaker's voice. The speaker encoder models squeeze-and-excitation blocks to concentrate more on the speaker's speaking style. The resultant model outperformed the speaker similarity score and naturalness and was the first research to adopt predictors for mean opinion score assessment.

2.4 Few-shot TTS model

Zhang et al. (2022) used only 2 minutes of data to clone a speaker's voice using a multimodal system comprising the VQ-VAE speaker encoder model for obtaining linguistic features, an extended sequence-to-sequence Tacotron 2 model, and Parallel WaveNet as the vocoder. The speaker encoder model was fine-tuned, and results depict the synthesized voice was more natural as compared to the one produced by Tacotron 2. Gabrys et al. (2022) developed VoiceFilter that used only one-minute data from the original speaker. Results depict that model was easily scalable for new speakers.

2.5 Zero-shot TTS model

Gorodetskii (2022) developed a speaker adaptation TTS system that requires data of a few seconds to generate the synthesized voice. The model comprised of speaker encoder model, multi-speaker Tacotron 2, and SC-WaveRNN to generate time-domain waveforms. Results depict the model could generate synthesized speech but could not match the prosody levels of the original speaker. Xiao et al. (2022) researched the use of speaker embeddings for generating synthesized voice. The research concludes a TTS model comprising of speaker encoder model is extremely useful in the zero-shot learning process. Zhao et al. (2022) developed nnSpeech, a model that used latent Z distribution, conditioned phonemes, and variational autoencoder. Results revealed that the model performed well on selected language but performed poorly on the cross-datasets problem.

2.6 GAN-based TTS model

Kumar et al. (2019) evaluated the performance of MelGAN, a non-autoregressive, parallel, and fast generating feed-forward neural architecture that used transposed convolutions for upscaling the input mel spectrograms. The mean opinion scores achieved by MelGAN depict performance comparable with WaveGlow and Tacotron 2. Kong et al. (2020) developed HiFi-GAN, mainly based on MelGAN architecture in conjunction with pre-trained Tacotron 2 achieved better results than WaveGlow. The HiFi-GAN model is trained in an adversarial manner and consists of one generator and two different discriminators. The model also incorporated different losses to improve and stabilize training. Keisuke et al. (2022) performed a comparative study on two HiFi-GAN, LPCNet, and MWDLP neural vocoders. The TTS model for this comparative study used the FastSpeech acoustic model. The performance of these vocoders was evaluated using both single and multi-speaker speech corpus. Experimental analysis for single speaker speech corpus depicts LPCNet and MWDLP resulted in a blurry component in the synthesized speech, whereas the HiFi-GAN v1 (where hidden channels = 512) achieved the highest mean opinion score. The scores achieved in multi-speaker speech corpus by MWDLP, and HiFi-GAN were far better than LPCNet, but altogether low as compared to the scores obtained in the single-speaker speech corpus condition. Under all circumstances, the quality of synthesized speech of HiFi-GAN v1 was better than HiFi-GAN v2 (where hidden channels = 128). Several implementations of HiFi-GAN vocoders demonstrate the production of high-quality synthesized speech.

2.7 Child Speech TTS Model

Jain et al. (2022) developed a TTS pipeline to achieve child speech synthesis. The TTS pipeline comprised of a pre-trained and fine-tuned speaker encoder model, a pre-trained and fine-tuned Tacotron 2 synthesizer, and a pre-trained WaveRNN vocoder which was further fine-tuned for child speech corpus. This research was the first to adopt the computerized mean opinion score predictor called as MOSNet. The synthesized child speech achieved a mean opinion score of 3.95 and MOSNet depicted a high correlation.

To summarize, detailed research was carried out to understand the progress in the domain and chart out a framework for a TTS model. The research was divided into seven categories based on the mechanism or type of a TTS model adopted to generate synthesized speech. TTS models that comprise of speaker encoder model and a GAN based vocoder generated synthesized speech that matched the original speaker's characteristics. Therefore, it can be concluded that for a multi-speaker speech corpus, an ideal TTS pipeline can mainly comprise of a speaker encoder model, an acoustic model such as FastSpeech/ Tacotron to generate acoustic features, and a GAN-based neural vocoder to yield better and high quality of synthesized speech.

3 Child Speech Synthesis Research Methodology

This research follows the stages of Knowledge Discovery in Databases (KDD) process. The detailed explanation of the KDD process application for this research is explained below.

3.1 Data Selection

A compelling dataset is essential to foster machine learning and derive valuable insights. Due to challenges in data collection of a child's speech, only a few multi-speaker child speech datasets are available. This study uses the My Science Tutor children's speech corpus (MyST) [1] which is the largest English-speaking child corpus freely available to the research community. MyST corpus contains child speech collected during the interaction of a total of 1371 students (grade 3-5) with a virtual science tutor. The dataset consists of 456 hours (.flac file format) of speech data resulting in 2,28,874 utterances out of which only 45% are transcribed (.trn file format). The dataset is divided into two phases based on the topic of discussion. The dataset also had the development, training, and test partitions.

3.2 Data Pre-processing

The data in the MyST corpus followed a specific nomenclature to easily differentiate between the speakers and the session. Out of the total data available in the MyST corpus, only 45% was available with the transcripts. The audio data present only in phase 1 of the dataset was transcribed at the word level with utmost precision. To understand the speaking patterns, pitch fluctuations, and voice modulations in a child's speech, this research uses purposive sampling i.e., a non-probability sampling method. This study uses the audio references having a corresponding transcript file from phase 2 as a sample representing the entire MyST corpus. The audio references with transcripts free of dis-fluency markers or stop words were then used for further steps. A detailed pre-processing procedure is explained in 5.1

3.3 Data Transformation

An audio data is usually comprised of complex features, thus, to recognize the audio distinctly, it is necessary to extract features. To extract features and characteristics relevant to a speaker, the pre-processed audio files were initially converted to .wav files. Post conversion, a corresponding mel spectrogram for each audio is generated. The mel spectrogram for each frame of the spectrum contains a short-time Fourier transform (STFT), from the linear frequency scale to the logarithmic mel scale, further going to the filter banks to be expressed on the mel scale frequency as the distribution of signal energy.

3.4 Model Building

The baseline text-to-speech synthesis model developed by Jain et al. (2022) operates as follows: An encoder model first takes the audio files to generate speaker-specific character-

[1] MyST Corpus: `https://boulderlearning.com/request-the-myst-corpus/`

istics, which are then used as input to the acoustic model. The acoustic model uses these speaker-specific characteristics to generate hidden sequences thereby producing acoustic features. These acoustic features are fed as an input to the vocoder model that produces waveforms and then produces the synthesized speech. Extending the baseline approach, a text-to-speech synthesis model is developed using the speaker encoder model, Tacotron 2 as the acoustic model, and HiFi-GAN developed by Kong et al. (2020) as the vocoder model, thereby facilitating a text-to-speech system for child speech synthesis.

3.5 Evaluation

Various subjective and objective measures are used to evaluate the performance of a text-to-speech system. This study uses a mean opinion score predictor called MOSNet developed by Lo et al. (2019), a deep learning based objective assessment that is built using convolution and recurrent neural network models. The MOSNet compares the original and the synthesized voice and generates a score of similarity based on the Mean Opinion Score scale as shown in Table 1. Two audio references are claimed to be similar if MOSNet generates a higher correlation value. The synthesized samples generated and their corresponding MOSNet scores are described in the section on experiments.

Table 1: Mean Opinion Score Rating Scale

Rating	Synthesized Speech Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

4 Design Specification

The research proposal for synthesizing child speech includes a text-to-speech system and is explained in the figure below. The text-to-speech system proposed in this study to synthesize child speech uses a pre-trained speaker encoder model trained on speaker verification task, a pre-trained state-of-the-art Tacotron 2 model and a HiFi-GAN vocoder model. The detailed description of the model architecture is given below.

4.1 Baseline model

Various speech synthesis models are adopted on adult speech corpora, but minimum investigation or research is done on achieving speech synthesis using a child speech corpus. An existing implementation of text-to-speech system done by Jain et al. (2022) to achieve child speech synthesis is considered as the baseline model for this research. The baseline model comprises of the speaker encoder model pretrained on four speech corpora (three adult speech and one child speech), the state-of-the-art Tacotron 2 as an acoustic model pretrained on two speech corpora (one adult speech and one child speech), and WaveRNN pretrained on two speech corpora (one adult speech and one child speech) as the vocoder model that performs sequential modelling of audio from mel spectrograms using a recurrent network that uses Gated Recurrent Unit (GRU) and dense layer transformation.

4.2 Proposed Model

To achieve child speech synthesis, this study integrates the encoder and acoustic model from the baseline model developed by Jain et al. (2022) with the high performing HiFi-GAN vocoder developed by Kong et al. (2020). The speaker encoder model is capable of generating speaker specific characteristics called embeddings which will be used by the Tacotron 2 model to generate acoustic features using the audio reference and speaker embeddings. Further, the HiFi-GAN vocoder converts the acoustic feature to a waveform thus generating the synthesized speech. Figure 1 shows the architecture of the proposed model.

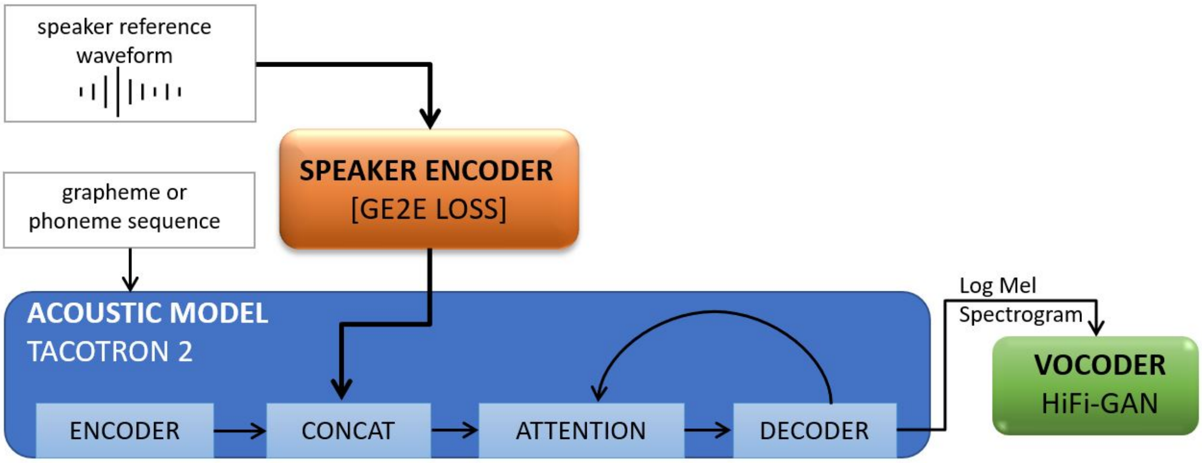


Figure 1: Proposed model for child speech synthesis

The detailed description and functioning of the proposed model are given below.

4.2.1 Encoder Model

The encoder model is responsible for generating embeddings using input audio (mel spectrogram frames). These embeddings represent characteristics of a speaker in the transformed space initiating from a target speaker's utterance. The speaker encoder model developed by Jia et al. (2018) in the Speech Vector to TTS technique works well for multi-speaker datasets regardless of the noise level in the audios. This speaker encoder model aims to generate embeddings for a target speaker using a speech signal that is short in length, irrespective of background noise and phonetic content. This is achieved by training it on a scalable and highly accurate neural network speaker verification task that attempts to optimize the generalized end-to-end (GE2E) loss (Wan et al.; 2018). This ensures that embeddings of utterances from different speakers are far apart in the embedding space while those of the same speaker demonstrate high cosine similarity. The highly accurate neural network computes a sequence of log-mel spectrogram frames and maps it to a fixed dimensional vector, known as the d-vector. The final embedding and the utterance-wise d-vector are generated by L2 normalizing the window-wise d-vectors and averaging them elementwise. A child's speech is inarticulate, consisting of heavy fluctuations in pitch, frequency etc. and noises. Since the SV2TTS speaker encoder model can extract meaningful speaker embedding from the noisy audio references, was therefore selected to be the encoder model used in this research. Initially the SV2TTS speaker encoder model was trained on LibriSpeech, VoxCeleb1 and VCTK. Jain et al. (2022) further trained the SV2TTS speaker encoder model for a million steps on the entire MyST corpora.

4.2.2 Acoustic Model/ Synthesizer

The acoustic/ synthesizer model is responsible for generating mel spectrograms using embeddings produced from the speaker encoder model. Tacotron 2, the state-of-the-art acoustic model developed by Google, is a neural network architecture that maps character embeddings to mel spectrograms. It additionally synthesizes the mel spectrograms to time-domain waveforms by using the WaveNet vocoder model with a few modifications. This research uses the state-of-the-art Tacotron 2 model with several modifications based on the works of Jemine et al. (2019) and Jain et al. (2022). Jemine et al. (2019) performed the initial modification of the Tacotron 2 model by stripping off the WaveNet vocoder from the original implementation and training it on clean LJSpeech, a single speaker dataset. Jain et al. (2022) further modified the Jemine et al. (2019)'s model by adding a speaker encoder to the architecture to make it function for multi-speaker datasets. The acoustic model takes the audio-text pair as input followed by an attention mechanism before decoding a spectrogram. Jain et al. (2022) finetuned this acoustic model on the child speech corpora for an additional 750k steps. The research uses these combined modifications of the state-of-the-art Tacotron 2 model.

4.2.3 Vocoder Model

A vocoder or a voice coder is a major component in speech synthesis. A vocoder model is responsible for the transformation of acoustic features into a waveform thereby producing the synthesized speech. This research mainly focuses on the vocoder model to generate synthesized speech for the multi-speaker child corpora. The generative adversarial networks often called GANs have seen promising applications in the field of speech synthesis. In a comparative study of different vocoders conducted by Matsubara et al. (2022), it was observed that the HiFi-GAN vocoder model proposed by Kong et al. (2020) achieved better-synthesized sample quality and has high computational efficiency even with a noisy input. Furthermore, the HiFi-GAN vocoder model could synthesize natural human-sounding speech at a faster rate and is generalizable for the inversion of mel spectrograms of unseen speakers. Several studies conducted by You et al. (2021), Kim et al. (2021), Beck et al. (2022) etc., adopted the HiFi-GAN model as the vocoder and achieved a synthesized speech like that of a human. Due to the impressive results of the HiFi-GAN model, this research adopts HiFi-GAN developed by Kong et al. (2020) as the vocoder model.
Following is the architectural description of the HiFi-GAN model.

1. Generator: The HiFi-GAN consists of a single CNN generator that upsamples the mel spectrograms received as input through transposed convolutions to match the output sequence length to the raw waveform's temporal resolution. The patterns of various lengths are observed in parallel with the help of a multi-receptive field fusion module in the generator. The summed outputs from multiple residual blocks are returned by the MRF module. The MRF module has some adjustable parameters to achieve a trade-off between the sample quality and the efficiency of the synthesized speech.

2. Discriminator: The HiFi-GAN consists of two discriminators namely multi-period discriminator (MPD) and multi-scale discriminator (MSD). The MPD consists of multiple sub-discriminators that are a stack of strided convolutional layer with leaky rectified linear unit activation (ReLU). These sub-discriminators capture varied tacit structures by considering various parts of the equally spaced input audio. To independently process the periodic signals the kernel size in the width axis is restricted to 1 in each convolutional layer of MPD. The MPD are additionally applied with weight normalizations. To evaluate the audio sequences, the MSD used in HiFi-GAN model is based on the works of MelGAN. The MSD consist of three sub-discriminators which are a stack of leaky ReLU activated grouped convolutional layers that operate on different input scales of audios. Since the first sub-discriminator directly operates on the raw audio, weight normalization is applied to the other two sub-discriminators.

3. Loss: The HiFi-GAN follows the least squares loss function training objectives of LSGAN for non-disappearing gradient flows. The training loss of both generator and discriminator is termed as GAN loss. The discriminator classifies ground truth samples and the synthesized samples from the generator as 1 and 0 respectively. To improve the fidelity of the generated audio and the efficiency of the generator a mel spectrogram loss represents the L1 distance of the mel spectrograms of ground truth and synthesized audio waveforms. The mel spectrogram loss helps in generating a synthesized waveform that is realistic given an input condition. It also helps in the early stabilizing of the adversarial training process. The widely adopted loss in speech synthesis systems is the feature matching loss, which is a metric that measures the feature difference between the ground truth and the synthesized sample. The final loss of the HiFi-GAN comprises the GAN loss, mel spectrogram loss, and feature matching loss.

5 Implementation

5.1 MyST Corpus

The audio references present in the MyST corpus are of varying durations which may impact the training process of the model. For further processing, the audio references between the duration of 10 to 15 seconds were considered, as the ones less than 10 seconds mostly contained unintelligible/ noisy speech and the ones greater than 15 seconds were discarded considering the system's capacity. The transcripts for these selected audio files were then checked for stop words/ disfluency markers such as breath, laugh, noise, silence, etc. and (()) (represents no phonetic meaning). Punctuation marks or special characters in the transcripts were removed. The audio files were discarded whose transcripts incur any disfluency markers. Using Python's Librosa library, the silence or long pauses were removed from this disfluency free 10-15 second lengthed audio files and were then converted to .wav files. A total of 20 hours of data was obtained post-completion of the pre-processing phase. For this study, a random selection of audio files totaling for an hour of data was used to train and evaluate the proposed model's capabilities.

5.2 Baseline model

The baseline text-to-speech consists of a speaker encoder model and an acoustic Tacotron 2 model. The implementation of the two components of the baseline architecture is described below.

The encoder being the first module in a text-to-speech system is crucial in understanding and deciphering the characteristics of a speaker. This research uses the speaker encoder same as that in the works of Jia et al. (2018); Jemine et al. (2019); Jain et al. (2022). The speaker encoder first generates the mel spectrograms from the clean and noisy segmented audio references, trains using the GE2E loss, and computes a fixed dimensional d-vector which is further optimized on the GE2E loss to differentiate between the speakers in the embeddings space. A pre-trained model is available for the speaker encoder that is trained for 1 million iterations on three different adult speech datasets viz. VCTK, LibriSpeech, VoxCeleb1, and a single child corpus viz. My Science Tutor produced a minimal equal error rate (EER) Jain et al. (2022). To ensure better generalization of adult and child speaker embeddings, pre-trained model [2] made available by Jain et al. (2022) was used for the speaker encoder.

The acoustic model is the second text-to-speech system that predicts mel spectrogram from a given input text. This research uses the acoustic model same as that in the work of Jemine et al. (2019); Jain et al. (2022) that takes audio references and transcripts. The state-of-the-art Tacotron 2 uses convolutional layers, LSTM layers, and an attention mechanism to generate the mel spectrograms. A pre-trained model is available for Tacotron 2 that is trained for 250k steps on clean Librispeech data further fine-tuned on a subset of My Science Tutor dataset for an additional 750k steps. This research uses the Tacotron 2 acoustic pre-trained model checkpoints [3] made available by Jain et al. (2022).

5.3 Novel approach

Generative adversarial networks (GAN) have performed remarkably in the field of machine learning. The foundation of GANs is based on a mini-max game which comprises a generator and a discriminator. The generator produces synthetic samples like that of the real data whereas the discriminator classifies data as a real or synthetic sample. The generator enhances to a point where the real and synthetic data samples are identical to the discriminator. The implementation of the HiFi-GAN vocoder in this research uses the python TTS library [4] and is based on the works of Kong et al. (2020) and is described below.

The HiFi-GAN-based neural vocoder consists of a single generator and two discriminators. The light-weighted generator comprises layers of convolution and transposed convolution. The generator uses acoustic features as input and generate upsamples with the help of transposed convolutions until the output sequence length matches the temporal resolution of the audio waveforms. To correctly capture the various components of the frequency in the speech waveforms, the outputs of the convolution layers are summed up by performing the multi-receptive field fusion (MRF). The first discriminator is called the multi-period discriminator (MPD) which comprises several sub discriminators each

[2] Speaker encoder checkpoints https://drive.google.com/drive/folders/1FuAY2XXcU0vLVo1f9QYQjhs_g9eURbio

[3] Acoustic model checkpoints https://drive.google.com/drive/folders/1wcxVnJ5mQZNdl1r_aLzY86iIAgRm4hQH

[4] Python TTS library https://github.com/coqui-ai/TTS

of which is responsible for capturing different structures from different parts of the audio. Each sub-discriminator in the MPD has the leaky rectified linear unit (ReLU) activation and weight normalization is applied to the entire MPD. As the MPD sub-discriminator accepts disjoint audio samples, to consecutively evaluate the audio sequence, an MSD comprising of three leaky ReLU activated sub-discriminator is added. The MSD uses smoothed waveforms to operate. Weight (except the first sub-discriminator) and spectral normalization is applied to stabilize training. The HiFi-GAN vocoder was implemented using the open-source python TTS library [5] with configuration parameters described below:

5.3.1 Generator

The generator block of the vocoder performs the multi-receptive field fusion (MRF) using residual block that consists of 3 convolutional layers in each 1D convolution block. Each convolutional block is activated with a leaky rectified linear unit (ReLU) [6]. Following are the hyper-parameters adjusted for the multi-receptive field fusion (MRF) module of the generator:

- Channels/ Hidden dimensions: The 1D representation of the convolution layers is represented as a stack of hidden channels. To ensure that both input and output channel is of the same size, the audio features are scaled and padded. Comparing the performance of HiFi-GAN v1 (512 hidden channels) with HiFi-GAN v2 (128 hidden channels), it was observed that the synthesized audio from HiFi-GAN v1 was far better Keisuke et al. (2022), and hence the channel for this research was set to 512.

- Kernel size (residual block): The kernel or more commonly called the filter size of a convolutional filter. The size of the kernel can be experimented with to understand the performance of the convolutional neural network (CNN). According to Pons et al. (2021), the receptiveness of the CNN can be increased by increasing the kernel size. The kernel size of the generator is kept in the increasing order as 3, 7, and 11 like that of HiFi-GAN v1 and v2.

- Dilation size For convolutional networks to obtain a substantial receptive field, a stack of dilated convolutions can be added. This stack of dilated convolutions considers the computational efficacy as well as the input resolution. The residual connection and 2 convolutional layers are stacked 3 in the residual block of the implemented vocoder. The dilation size is set to [[1, 1], [3, 1], [5, 1]] $\times 3$ like that of HiFi-GAN v2 Kong et al. (2020).

- Kernel size (transposed convolutions): The transposed convolutions are used in the residual block to upsample its output. The kernel size of the transposed convolutions is set to 16,16,4,4 due to its wide acceptance in the implementation of various vocoders (You et al.; 2021; Kim et al.; 2021; Beck et al.; 2022).

- Stride/ upsample kernel factor: The kernels of the transposed convolutions are required to shift when moving across different input segments. The stride for the

[5] Python TTS library `https://pypi.org/project/TTS/`

[6] HiFi-GAN Generator `https://github.com/coqui-ai/TTS/blob/dev/TTS/vocoder/models/hifigan_generator.py`

kernels in the transposed convolutions is set to 8,8,2,2 due to its wide acceptance in the implementation of various vocoders (You et al.; 2021; Kim et al.; 2021; Beck et al.; 2022).

- Optimizer: Most of the neural vocoder implementations You et al. (2021); Kim et al. (2021); Beck et al. (2022) used AdamW as the optimizer, which is why the HiFi-GAN vocoder implemented in this research also uses the AdamW optimizer with beta values as 0.8 and 0.99 with a weight decay of 0.01
- Learning rate/ step size: The magnitude of changing/ updating the weights of the model during the back-propagation training process is set to 0.0001.
- Batch size: To fasten the completion of an epoch during the training, the batch size was set to 32.

5.3.2 Discriminator

The discriminator block of the vocoder consists of the multi-period and multi-scale discriminators which is based on the works of Kumar et al. (2019) and uses the same configuration parameters of the HiFi-GAN v1 discriminator [7] (Kong et al.; 2020). Some of the parameters are mentioned below:

- Learning rate: The discriminator is implemented with an exponential learning rate with a multiplicative decaying factor set to 0.999.

Other parameters of the implemented HiFi-GAN vocoder are described below:

- Fast Fourier Transformation (FFT) size: FFT is an optimized algorithm that converts an audio signal into spectral components and provides frequency information of the signal. The value of FFT like that of HiFi-GAN is set to 1024.
- Window size: The window size is an important parameter for analysis as it influences the frequency or temporal resolution of the audio signal analysis and is set to 1024.
- Hop length: The hop size/ length refers to the distance between the centres of consecutive windows. An overlap of windows occurs if the window size is greater than the hop size. A large overlap of windows leads to smoother spectral transitions but is computationally expensive. The hop size for this research is therefore set to 256 which also follows the multiplication of all the upsample rates mentioned for the generator.
- The number of mel spectrograms: A range of different mel spectrograms filter banks was analysed by Cheuk et al. (2020) and based on their analysis, the value was set to 80.
- Sample rate: The sample rate mentions the number of times an audio reference is sampled in a second. The best sample rate is 44.1kHz for musical audio. Since the audio references for this research are mostly conversational, the sample rate was set to 22kHz like that of most text-to-speech systems.

[7] HiFi-GAN Discriminator `https://github.com/coqui-ai/TTS/blob/dev/TTS/vocoder/models/hifigan_discriminator.py`

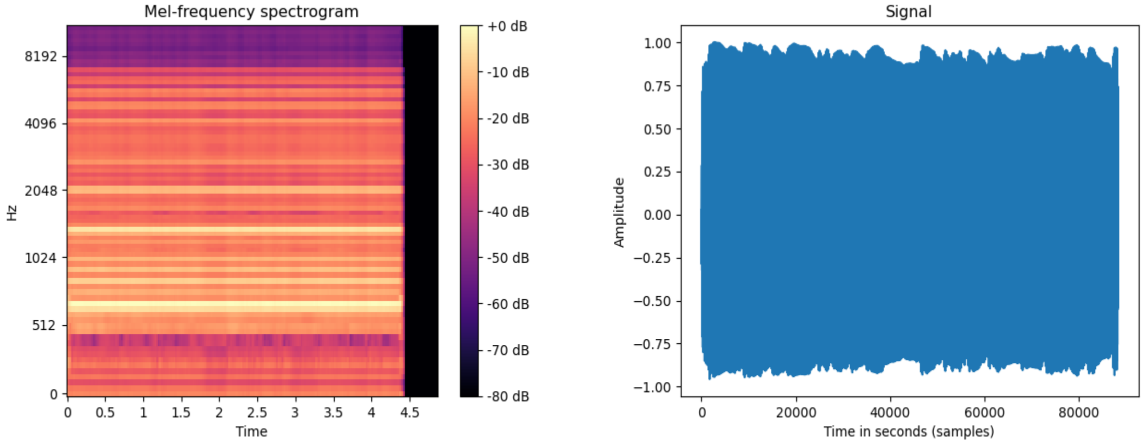
- Activation function: Like most of the neural GAN vocoders, Leaky ReLU is the activation function used for the implementation of the vocoder in this research with the hyper-parameter alpha value set to 0.1
- Loss terms: The implemented vocoder is trained by adding the GAN, mel spectrogram, and feature matching loss.

6 Evaluation

This segment summarizes the findings supporting the research question. Several experiments were conducted to synthesize speech using the novel approach of using a HiFi-GAN vocoder in the baseline architecture of child speech synthesis. The proposed model's performance is evaluated by conducting experiments post training it on the MyST corpus.

6.1 Experiment 1: HiFi-GAN vocoder trained only on MyST corpus – 100 epochs

To achieve child speech synthesis purely on a child speech corpus, the HiFi-GAN vocoder model was initially trained from scratch on a random subset (10%) of the pre-processed MyST dataset. The first round of experiments was conducted once the vocoder model was trained for 100 epochs. As the vocoder model was being trained from the scratch, this experiment focused on generating an experimental voice from the model irrespective of the audio reference of child/ adult speech. The experimental voice obtained post 100 epochs were purely metallic, rough, and unpleasant to the human ears. The mel spectrogram and the waveform of the experimental voice as shown in Figure 2 were plotted using the Librosa, a python library. The mel spectrogram depicts the generated sound as devoid of any resonance or fluctuations, with a constant type of sound propagating from start to end. Similarly, the waveform of the experimental voice reveals it to have a constantly high pitch sound throughout the duration.



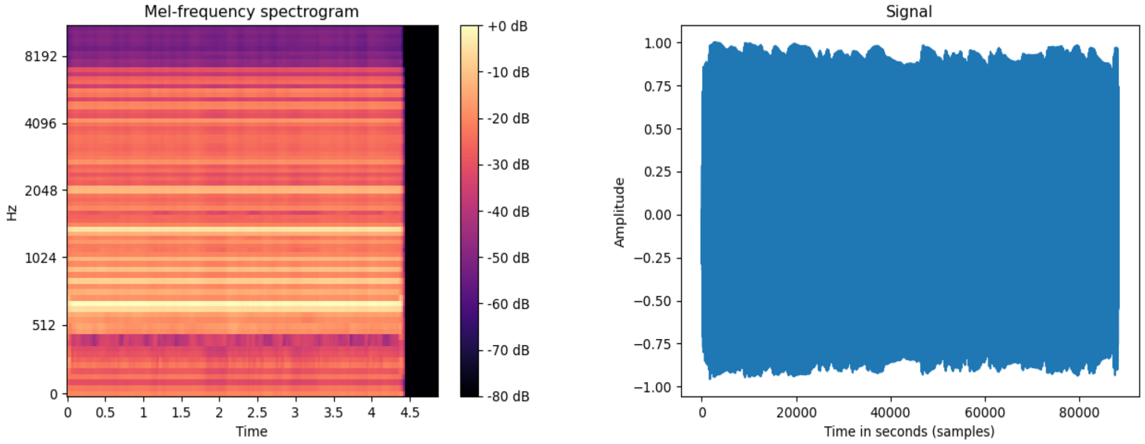
(a) Mel spectrogram of experimental voice

(b) Waveform of experimental voice

Figure 2: Experimental voice characteristics for HiFi-GAN - 100 epochs

6.2 Experiment 2: HiFi-GAN vocoder trained only on MyST corpus – 200 epochs

The HiFi-GAN vocoder was then trained for an additional 100 epochs on the same subset (10%) of the pre-processed MyST dataset. This experiment was conducted to obtain an experimental voice resembling human speech unlike the output of the previous experiment. A experimental voice was generated from the model irrespective of the audio reference of child/ adult speech. The experimental voice obtained from this model was similar to the experimental voice obtained from the experiment. The mel spectrogram and the waveform of the experimental voice were plotted using the python library called Librosa. The mel spectrogram and the waveform of the experimental voice obtained from the model post training for 200 epochs as shown in Figure 3 did not have any improvements and still sounded rough and metallic.



(a) Mel spectrogram of experimental voice

(b) Waveform of experimental voice

Figure 3: Experimental voice characteristics for HiFi-GAN - 200 epochs

6.3 Experiment 3: Pre-trained HiFi-GAN

This experiment was conducted to explore the feasibility of the existing pre-trained HiFi-GAN vocoder to generate synthesized speech using child audio reference. An original audio recording from the pre-processed MyST corpus was given as input to the speaker encoder model which generated speaker embeddings. These speaker embeddings were then fed to the acoustic model to generate acoustic features. The acoustic features were then used by the HiFi-GAN vocoder model to generate synthesized voice. The mel spectrograms and waveform of the original and the synthesized voices were plotted. The waveforms of the two voices were checked for similarity using the MOSNet CNN-BLSTM evaluation [8]. The mel spectrograms as shown in Figure 4 reveals that the original and synthesized voice are poorly similar. The mel spectrogram of the synthesized voice shows a steady signal at the start and the end of the audio reference. This depicts that the synthesized voice produced consists of a similar sound that is unintelligible resembling a noise. Though the synthesized voice produces some valid audio output, the noise at the ends is not acceptable.

[8] MOSNet Implementation `https://github.com/lochenchou/MOSNet.git`

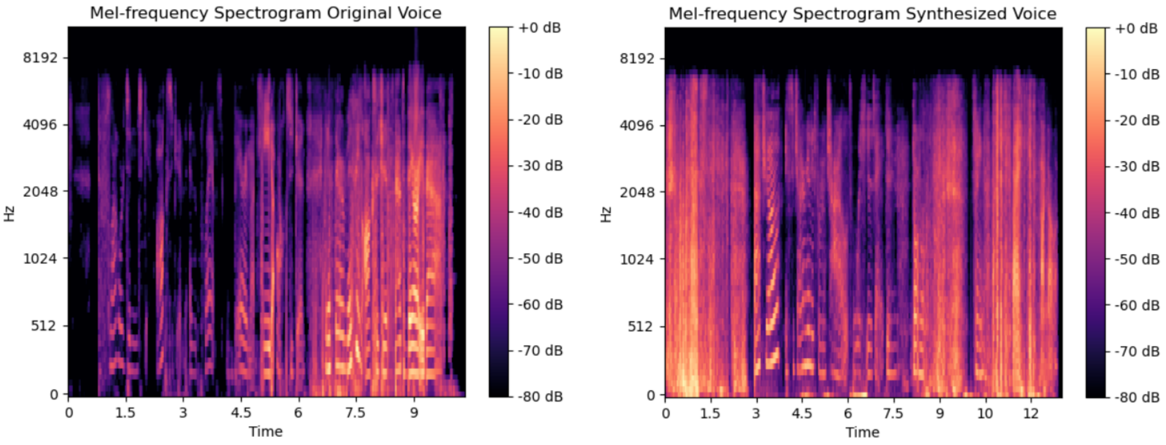


Figure 4: Mel spectrograms of original and synthesized voice from pre-trained HiFi-GAN

The waveforms of the two voices are shown in Figure 5 reveals that periodicity trends in the waveform match to a very small extent and confirm the presence of high pitch only at the start of the audio.

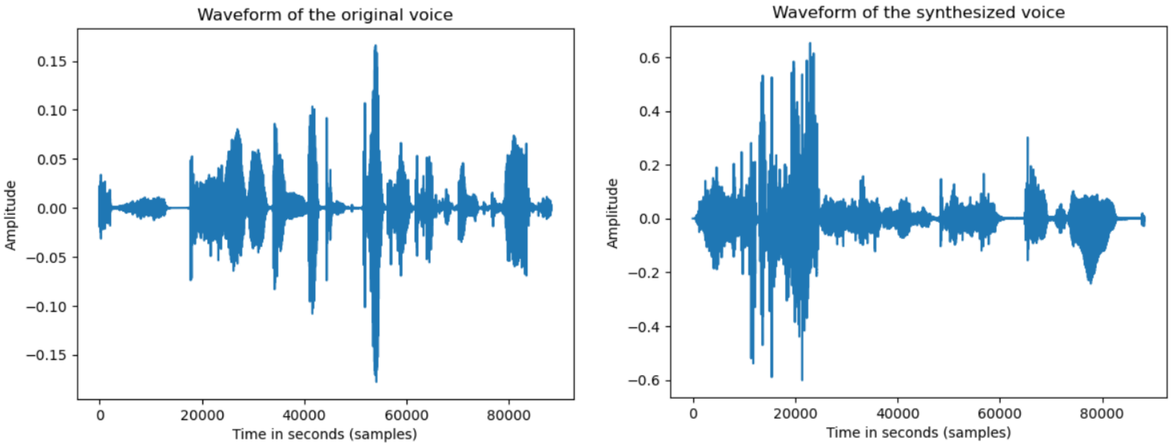
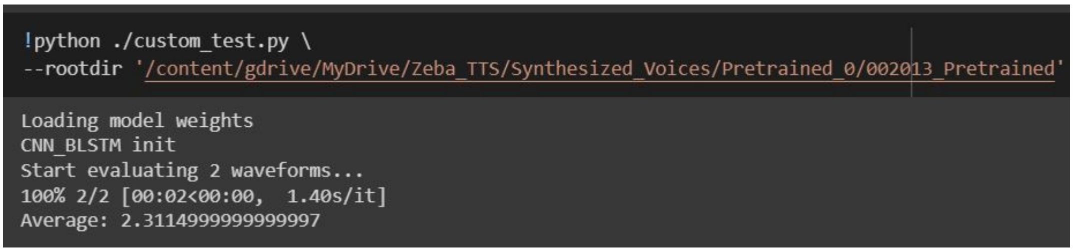


Figure 5: Waveforms of original and synthesized voice from pre-trained HiFi-GAN

Using the CNN-BLSTM-based MOSNet evaluation, the two audio references achieved a score of 2.31 as shown in Figure 6 representing a significantly smaller correlation between them, thus confirming a poor quality of the synthesized speech.



```
!python ./custom_test.py \
--rootdir '/content/gdrive/MyDrive/Zeba_TTS/Synthesized_Voices/Pretrained_0/002013_Pretrained'
```

```
Loading model weights
CNN_BLSTM init
Start evaluating 2 waveforms...
100% 2/2 [00:02<00:00,  1.40s/it]
Average: 2.3114999999999997
```

Figure 6: MOSNet scores of the original and synthesized voice from pre-trained HiFi-GAN

6.4 Experiment 4: Transfer learning approach for HiFi-GAN fine-tuned on MyST corpus – 100 epochs

To generate a human-sounding synthesized voice, the vocoder model had to be trained on the MyST corpus for at least 600k steps.[9] Considering the resources and its limitation, the transfer learning approach was adopted. A pre-trained HiFi-GAN vocoder model was fine-tuned on the MyST corpus based on the hyper-parameter configuration described in the section implementation. This experiment evaluates the synthesized voice generated from the proposed model post fine-tuning the HiFi-GAN vocoder model for 100 epochs on the MyST corpus. An original audio recording from the pre-processed MyST corpus was given as input to the speaker encoder model which generated speaker embeddings. These speaker embeddings were then fed to the acoustic model to generate acoustic features. The acoustic features were then used by the HiFi-GAN vocoder model to generate synthesized voice. The mel spectrogram of the original and the synthesized voice as shown in Figure 7 still depicts poor similarity. The mel spectrogram of the synthesized voice is blurry throughout depicting a sense of noise throughout the audio reference.

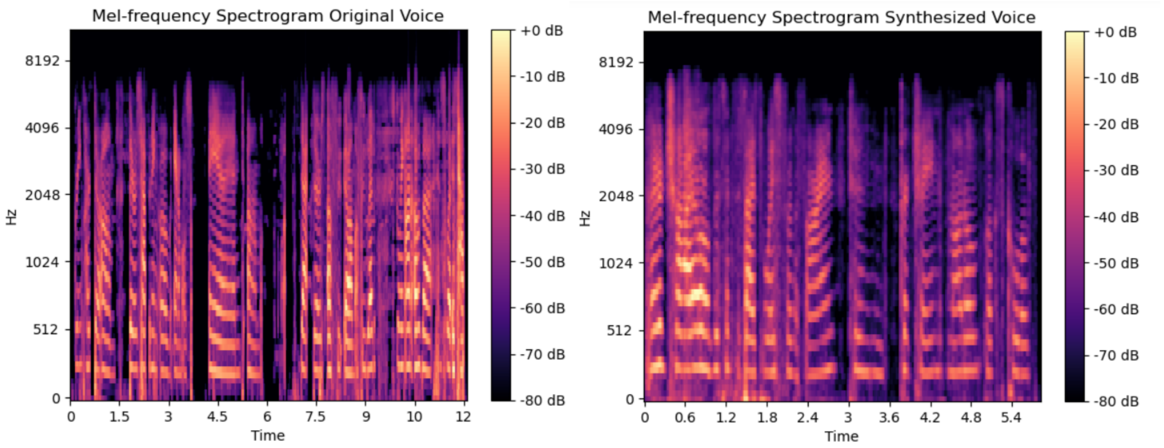


Figure 7: Mel spectrograms of original and synthesized voice post 100 epochs

The waveform of the original and synthesized voice is shown in Figure 8. The waveform of both the audio references shows a similarity of an insignificant level. The synthesized waveform represents noise and a high amplitude at the start of the audio.

[9] Ideal Vocoder Steps: `https://discourse.mozilla.org/t/training-russian-tts/70381`

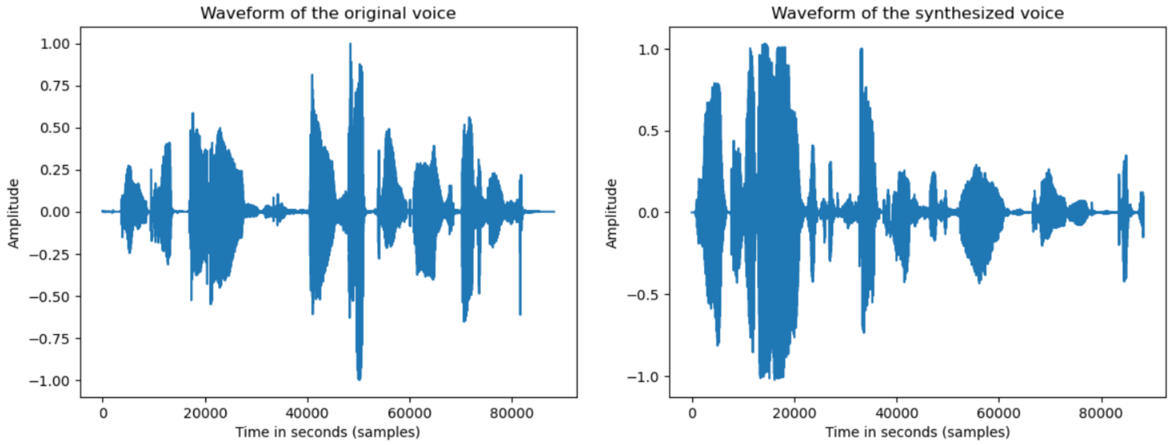
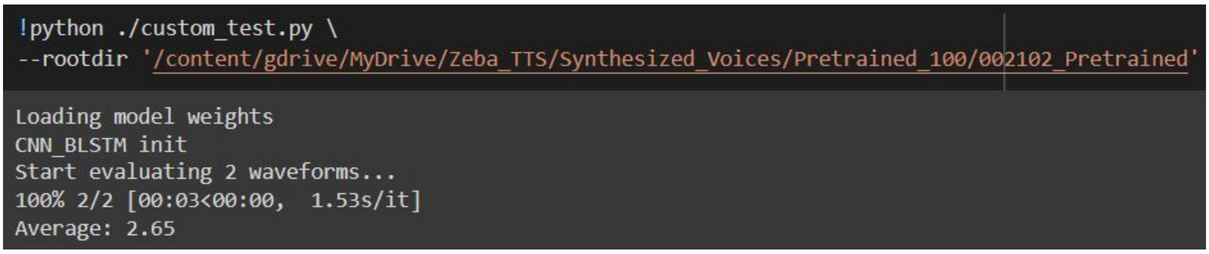


Figure 8: Waveforms of original and synthesized voice post 100 epochs

Using the CNN-BLSTM-based MOSNet evaluation, the two audio references achieved a score of 2.65 as shown in Figure 9 representing a significantly smaller correlation between them, thus confirming a poor quality of the synthesized speech. However, the MOSNet score achieved is slightly better than the MOSNet score obtained in the previous experiment.



```
!python ./custom_test.py \
--rootdir '/content/gdrive/MyDrive/Zeba_TTS/Synthesized_Voices/Pretrained_100/002102_Pretrained'

Loading model weights
CNN_BLSTM init
Start evaluating 2 waveforms...
100% 2/2 [00:03<00:00,  1.53s/it]
Average: 2.65
```

Figure 9: MOSNet scores of the original and synthesized voice from fine-tuned HiFi-GAN

6.5 Experiment 5: Transfer learning approach for HiFi-GAN fine-tuned on MyST corpus – 200 epochs

This experiment was conducted to analyse if a better MOSNet score can be achieved by further training the fine-tuned model on the MyST corpus. Based on the hyperparameters explained in the section implementation, the pre-trained HiFi-GAN vocoder model was further fine-tuned on MyST corpus for an additional 100 epochs. The original and the synthesized voice generated from the text-to-speech system using this fine-tuned vocoder model was analysed and mel spectrograms and waveforms were plotted using the python library called Librosa. The mel spectrogram of the two audio references is shown in Figure 10. The mel spectrograms depict resembling patterns but still comprises some distortion across the signal.

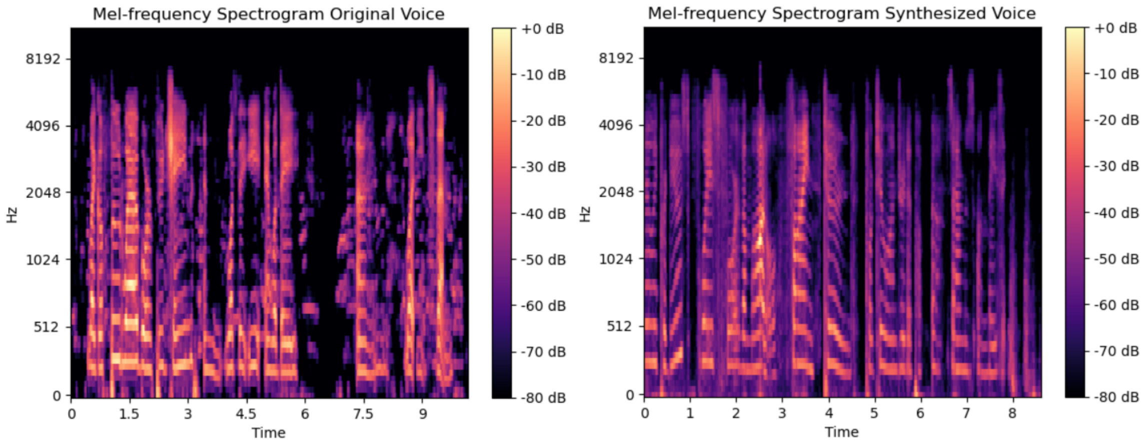


Figure 10: Mel spectrograms of original and synthesized voice post 200 epochs

The waveforms of the two audio references are shown in Figure 11. The absence of the high amplitude fluctuations at the ends of the signal for the synthesized waveform resembles that the noise element is disappearing as the model is trained further.

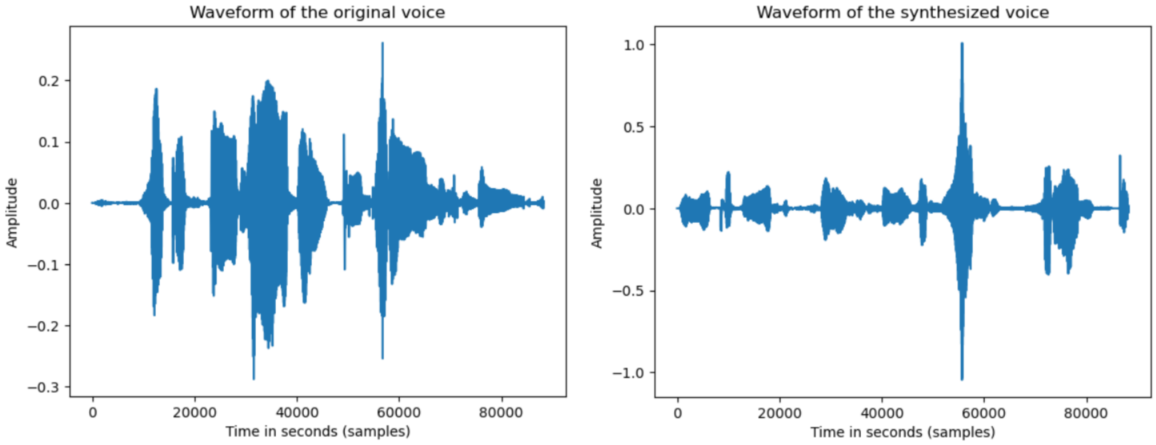
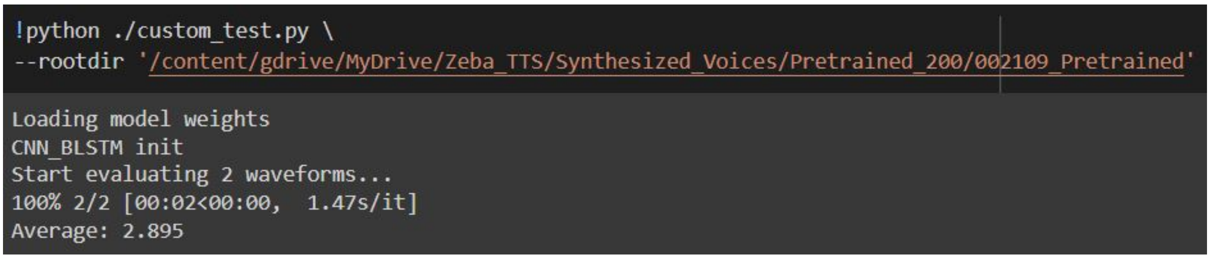


Figure 11: Waveforms of original and synthesized voice post 200 epochs

Using the CNN-BLSTM-based MOSNet evaluation, the two audio references achieved a score of 2.89 as shown in Figure 12 representing a slightly better yet smaller correlation amongst them, depicting the quality of the synthesized speech to be relatively better than the previous experiments but is still poor.



```
!python ./custom_test.py \
--rootdir '/content/gdrive/MyDrive/Zeba_TTS/Synthesized_Voices/Pretrained_200/002109_Pretrained'
```

```
Loading model weights
CNN_BLSTM init
Start evaluating 2 waveforms...
100% 2/2 [00:02<00:00,  1.47s/it]
Average: 2.895
```

Figure 12: MOSNet scores of the original and synthesized voice from fine-tuned HiFi-GAN

6.6 Discussion

The experiments conducted so far yielded intelligible insights that can aid in answering the research question of this study. Vocoders of varied configurations exist in the current text-to-speech systems. The applicability of these vocoders can experiment on different datasets. Existing vocoders have delivered promising results when trialed on various adult speech corpus. One such best-performing vocoder of adult speech corpus was chosen to be investigated on a child speech corpus.

The vocoder in the proposed model when trained from scratch for 100 or 200 epochs produced unintelligible, rough, high-pitched, and purely metallic sounds. Research carried out to understand the ideal number of epochs for which the vocoder should be trained revealed a minimum of 600k epochs can produce human-sounding speech. Due to the limitation of the scope and the resources, the transfer learning approach was adopted.

A HiFi-GAN pre-trained (on adult speech corpus) vocoder model was chosen for further experiments. The pre-trained HiFi-GAN vocoder could produce intelligible speech of very poor quality having a MOSNet score of 2.31 when given a child speech audio reference. On fine-tuning this pre-trained vocoder model on the MyST corpus, the proposed model could produce human-sounding speech having MOSNet scores greater than 2.6 and 2.8 post training it for 100 and 200 epochs respectively.

As the vocoder model is trained for a greater number of epochs, the MOSNet score tends to increase. The training loss was additionally observed during the fine-tuning of the pre-trained HiFi-GAN vocoder model. The training loss shows a decreasing trend, thereby resulting in the production of better, intelligible, human-sounding, natural synthesized speech similar to the original speaker. Also, the MyST corpus purely contained audio references from a male speaker, the behavior of the proposed model may differ completely for a female speaker. Additionally, fine-tuning and training the pre-trained HiFi-GAN vocoder on both male and female speaker child corpus further for a greater number of epochs will yield better-synthesized child speech resembling the original speaker's characteristics.

To summarize, child speech synthesis can be efficiently achieved using the proposed model.

7 Conclusion and Future Work

The purpose of this study was to synthesize child speech in the original speaker's voice using deep learning models. This study presents a text-to-speech system to achieve child speech synthesis using the HiFi-GAN vocoder in conjunction with the pre-trained speaker encoder and pre-trained Tacotron 2 as the synthesizer. The raw audio reference and their transcripts are pre-processed and transformed and are used as an input to the proposed text-to-speech system. Using the proposed text-to-speech system, experiments were conducted and revealed the necessity of using the transfer learning approach for the vocoder model. These experiments also disclosed that synthesizing speech from the beginning for a given dataset will at least require training a vocoder model for a minimum of 600k epochs.

Additional experiments revealed that the transfer learning and fine-tuning approach deems fit to extend the existing works on the vocoder model for child speech synthesis. Objective evaluation of the synthesized speech from a given audio reference for the implemented HiFi-GAN vocoder achieved a MOSNet score on an average of 2.31, 2.6, and 2.8 for the existing pre-trained, fine-tuned on MyST corpus for 100 and 200 epochs respectively.

The transfer learning approach for the HiFi-GAN vocoder model can be further applied to fine-tune and train on the MyST corpus for a greater number of epochs to achieve a highly intelligible, natural, human-like synthesized voice in the original speaker's voice.

Acknowledgement: The author would like to express deepest gratitude to supervisor Dr. Abubakr Siddig for providing guidance and valuable feedback during the project.

8 Supervisor Recommendation

8.1 Where do your artefact solution and overall outcome of your project fit in the body of knowledge?

The current coursework exhibited various ways to deal with text and image data. This study on achieving child speech synthesis using deep learning educates me on ways of dealing with and handling audio data.

1. Sampling the data: The examples encountered during the coursework always had predetermined sample data of the population. The dataset used in this study was too huge and due to constraints of the resources, a need of using sample data arose. Different types and techniques of sampling were understood and for the problem statement of this study, purposive sampling was adopted as it was best suited.

2. Cleaning and pre-processing audio data: The coursework acquainted me with different ways of dealing with text and image data. This study specifically focuses on audio data. The characteristics associated with an audio signal were understood. Different python libraries to clean the audio data were tried, from which Librosa was chosen. The difference between the spectrograms and mel spectrograms was comprehended due to which the study focused on using mel spectrograms. While using audio data, the sampling rate is the key factor in deciding the audio quality. Most of the TTS models use a sampling rate of 22KHz, but a sampling rate range of 16KHz – 22KHz is advisable. This study initially started by adopting a sampling rate of 16KHz, which made the audio sound more comic. Therefore, a rate of 22KHz was adopted.

3. Generative Adversarial Network: Usage of GAN for image and caption generation was studied and implemented during the coursework. The usage of GAN model for audio and speech signal processing was explored in this study. Based on the existing performance of GAN-based models, an optimum configuration of the parameters was chosen to conduct this research.

4. TTS python library: The functioning of the TTS python library was explored and understood. The TTS library consists of various text-to-speech models developed so far, which makes it easier to synthesize speech for new and different datasets. Before proceeding with TTS library, NVIDIA's Nemo toolkit was also explored, but due to ease of implementation and user friendliness, the python TTS library was adopted.

To summarize, this study equips me with an in-depth knowledge of developing a deep learning-based model to achieve audio and speech generation. Handling huge and complex signal through python libraries, channelling it through a deep learning architecture and creating meaningful value from the existing data which has a wide variety of real-world application is the key takeaway from this project. TTS models at present are quite accurate for generating synthesized speech for native English speakers with a neutral level of accent. With the knowledge acquired from this study, a TTS model tailored for low level language considering accents can be developed.

8.2 Why did you make the design decision to not include a comparative objective?

The proposed model (SV2TTS encoder – Tacotron 2 synthesizer – HiFi-GAN vocoder) was built on the combined works of HiFi-GAN (Kong et al.; 2020) and child speech synthesis model (SV2TTS – Tacotron 2 - WaveRNN) (Jain et al.; 2022). The original implementation of various versions of the HiFi-GAN model was trained on the single speaker LJSpeech dataset and its synthesis capabilities using the multi-speaker VTCK dataset were compared to the publicly available and best performing TTS models viz. WaveNet (trained on multiple Google TTS datasets), WaveGlow (trained on LJSpeech dataset), and MelGAN (trained on LJSpeech dataset). HiFi-GAN outperformed these models under every condition, and from the different versions of HiFi-GAN, the v1 achieved the best synthesis capabilities.

Various other implementations of TTS models using HiFi-GAN architecture depicted outperforming synthesizing capabilities when evaluated using various publicly available adult speech datasets. As the performance evaluation of HiFi-GAN was largely done on adult speech, evaluating it solely on a child's speech corpus became a necessity.

Also, a combination of adult and child speech corpora (My Science Tutor, VCTK, VoxCeleb1, and LibriSpeech dataset) was used to train the existing child TTS, the model's evaluation was done using the MyST corpus. On similar lines, this study fine-tunes the pre-trained HiFi-GAN vocoder on the MyST corpus.

This study focuses on generating synthesized voice in the original speaker's (child) voice. An adult speech is way different than a child's speech therefore, comparing speech synthesizing capabilities of existing models (trained on adult speech) with the proposed model (trained on adult and child speech) may result in unfair judgment, i.e., comparing the TTS model trained on adult speech dataset with the proposed model trained on child speech dataset is similar to comparing apples with oranges.

As a result, instead of comparing performances of the existing and proposed model, the performance assessment of the proposed model by comparing the synthesized outputs achieved by it at different epoch levels is deemed to be a better approach.

8.3 According to MOS presented in Table 1, a result of 2 is poor and 3 is fair, yet you claim that 2.89 is a good result. Can you please discuss and comment on that?

Instead of comparing different TTS models, this study compares the performance of the proposed model at different epoch levels. The experiments conducted in this study depict the MOS scores increase for synthesized speech at different epoch levels. The straightforward usage of pre-trained model generated an MOS score of 2.31, whereas the pre-trained fine-tuned proposed model achieved an MOS score of 2.65 and 2.89 for 100 and 200 epochs respectively.

Based on the Mean Opinion Score scale, the MOS score and its quality equivalence is as follows: 1- Bad, 2 – Poor, 3 – Fair, 4 – Good, and 5 – Excellent.

This study with the help of experiments conducted at different epoch levels, reveals that the MOS score tends to get better than the one achieved by it in the previous experiment. The MOS score obtained during the experiments shows an upward increasing trend.

Based on the experiments conducted, this study claims the following:

1. Though the MOS score of 2.89 achieved by fine-tuning and training the HiFi-GAN vocoder for 200 epochs was better in comparison with the one achieved (2.65) for fine-tuned and trained HiFi-GAN vocoder for 100 epochs, the (2.89) MOS score is still poor (Figure 12).

2. The study further claims that a better MOS score is expected to be achieved by further training the HiFi-GAN vocoder for a larger number of epochs. A glimpse of forecast of MOS score for further epoch is shown in Figure 13.

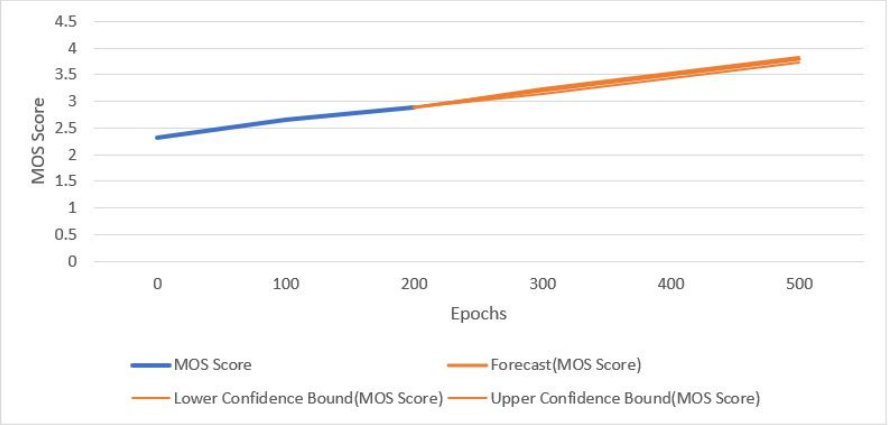


Figure 13: Forecast of MOSNet scores of fine-tuned HiFi-GAN vocoder

8.4 Can you summarize what you feel is the overall novelty of this research undertaken?

Current child speech synthesis models are either built using Hidden Markov Model (HMM) or Recurrent Neural Network (RNN), whilst the adult speech synthesis models are progressively increasing by adapting various GAN-based models. In comparison with adult speech, a child's speech is often inarticulate and disfluent which demonstrates variability in frequency and acoustic features. The situational analysis of current child speech synthesis models conducted by (Terblanche et al.; 2022) revealed that the adaptation of adult speech synthesis models to develop child speech synthesis models is a feasible method that can generate natural-sounding and intelligible synthetic speech.

This study sheds a light on developing a child TTS model by combining the outperforming adult TTS model (HiFi-GAN) with the existing blocks from child TTS models (SV2TTS – Tacotron 2). The novelty of this research is as follows:

1. Training the original single-speaker HiFi-GAN for a multi-speaker speech corpus:

    - Since the original HiFi-GAN TTS model has been trained on a single-speaker LJSpeech dataset, this study initially focused on training the HiFi-GAN vocoder purely on the multi-speaker MyST child speech corpus from scratch. Post training and evaluating it for several epochs, it was understood that the vocoder can generate natural speech only after 600k epochs.

      Due to the limitation of the resource, the focus of the study based on the analysis of Terblanche et al. (2022), was shifted to use the pre-trained HiFi-GAN model to fine-tune it for the child's speech corpus.

2. Using single-speaker pre-trained HiFi-GAN to produce synthetic speech for an unseen speaker:

    - A straightforward plug-and-play of the pre-trained (adult + child) SV2TTS speaker encoder and Tacotron 2 synthesizer along with the pre-trained (only adult) HiFi-GAN vocoder could generate a synthetic child speech in the original speaker's voice which heavily consisted of noise and long pauses.
    - It was understood that the synthetic speech produced with this simple plug-and-play was only possible due to the use of the pre-trained speaker encoder model that created speaker-specific embedding which helped the HiFi-GAN vocoder to produce the synthetic speech in the original speaker's voice.

3. Fine-tuning the pre-trained (single-speaker) HiFi-GAN for a multi-speaker child speech corpus:

    - The pre-trained single-speaker HiFi-GAN vocoder was fine-tuned for the multi-speaker child speech corpus for various epochs. The synthetic speech produced by the proposed models was evaluated for different epochs. The objective evaluation (MOSNet scores) revealed that the quality of the synthesized speech tends to increase as the vocoder is trained further.

This research confirms the claims of Terblanche et al. (2022), that adaptation of adult TTS models to develop a child TTS model is a viable approach. This research additionally reveals that a pre-trained single-speaker adult TTS model could generate synthetic speech for unseen multi-speakers by fine-tuning it on the multi-speaker child speech corpus along with the usage of a speaker encoder model.

8.5 When comparing the waveforms (fig 5,8 and 11) are you using the same sentence? Is it taken from the corpus?

For conducting experiments, the original voices were selected randomly from the dataset (voices which were not a part of the training set). The sentences for the experiments were given at random and were not a part of any transcripts from the dataset.

The following sentences were used to generate synthesized voices:

1. For pre-trained and fine-tuned HiFi-GAN vocoder (Experiment 6.3):

   "This sound is generated for the child using the original pre-trained HiFi-GAN model."

2. For pre-trained and fine-tuned HiFi-GAN vocoder trained for 100 epochs 6.4):

   "Trying to produce synthesized voice on the fine-tuned model for 100 epochs. Sounds cool?"

3. For pre-trained and fine-tuned HiFi-GAN vocoder trained for 200 epochs 6.5):

   "Trial from the model fine-tuned for 200 epochs, sounds amazing."

8.6 Can you please provide an audio sample of an original sound and a synthesized one?

The folder consisting of synthesized voices obtained from different experiments is uploaded at [10].

References

Arik, Diamos, Gregory, Gibiansky, Andrew, Miller, John, Peng, Kainan, Ping, Wei, Raiman, Jonathan, Zhou and Yanqi (2017). Deep voice 2: Multi-speaker neural text-to-speech.

Beck, G. T. D., Wennberg, U., Malisz, Z. and Henter, G. E. (2022). Wavebender gan: An architecture for phonetically meaningful speech manipulation, pp. 6187–6191.

Carlson, R. (1995). Models of speech synthesis., *Proceedings of the National Academy of Sciences* **92**(22): 9932–9937.

Chen, Ren, Pengfei, Mao and Zhao (2022). Limited text speech synthesis with electroglottograph based on bi-lstm and modified tacotron-2.

Cheuk, K. W., Agres, K. and Herremans, D. (2020). The impact of audio input representations on neural network based music transcription, pp. 1–6.

[10] Synthesized Voice Output: `https://studentncirl-my.sharepoint.com/:u:/r/personal/x20227086_student_ncirl_ie/Documents/x20227086_ChildSpeechSynthesis/Synthesized_Voices-ZebaThesis.zip?csf=1&web=1&e=IdgOij`

Das, Williams and Lai (2022). Analysis of voice conversion and code-switching synthesis using vq-vae.

Flanagan, J. L., Coker, C. H., Rabiner, L. R., Schafer, R. W. and Umeda, N. (1970). Synthetic voices for computers, *IEEE spectrum* **7**(10): 22–45.

Gabrys, Huybrechts, Ribeiro, Sam, Chien, Roth, Comini, Barra-Chicote, Perz and Lorenzo-Trueba (2022). Voice filter: Few-shot text-to-speech speaker adaptation using voice conversion as a post-processing module.

Gorodetskii (2022). Zero-shot long-form voice cloning with dynamic convolution attention.

Jain, R., Yiwere, M. Y., Bigioi, D., Corcoran, P. and Cucu, H. (2022). A text-to-speech pipeline, evaluation methodology, and initial fine-tuning results for child speech synthesis, *IEEE Access* **10**: 47628–47642.

Jemine, C. et al. (2019). Master thesis: Real-time voice cloning.

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Lopez Moreno, I., Wu, Y. et al. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis, *Advances in neural information processing systems* **31**.

Kalchbrenner, Nal, Elsen, Erich, Simonyan, Karen, Noury, Seb, Casagrande, Norman, Lockhart, Edward, Stimberg, Florian, Oord, Dieleman and Kavukcuoglu (2018). Efficient neural audio synthesis, **80**: 2410–2419.

Keisuke, Takuma, Ryoichi, Tetsuya, Tomoki and Hisashi (2022). Comparison of real-time multi-speaker neural vocoders on cpus, *Acoustical Science and Technology* **43**(2): 121–124.

Kim, J.-H., Lee, S.-H., Lee, J.-H. and Lee, S.-W. (2021). Fre-gan: Adversarial frequency-consistent audio synthesis, *arXiv preprint arXiv:2106.02297* .

Koffi (2022). A tutorial on formant-based speech synthesis for the documentation of critically endangered languages.

Kong, Kim and Bae (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis.

Kumar, Rithesh, Thibault, Gestin, Teoh, Sotelo, Alexandre, Bengio and Courville (2019). Melgan: Generative adversarial networks for conditional waveform synthesis, **32**.

Lei, Yang, Wang and Xie (2022). Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **30**: 853–864.

Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y. and Wang, H.-M. (2019). Mosnet: Deep learning based objective assessment for voice conversion, *arXiv preprint arXiv:1904.08352* .

Manzelli, Rachel, Thakkar, Vijay, Siahkamari, Ali, Kulis and Brian (2018). An end to end model for automatic music generation: Combining deep raw and symbolic audio networks.

Matsubara, K., Okamoto, T., Takashima, R., Takiguchi, T., Toda, T. and Kawai, H. (2022). Comparison of real-time multi-speaker neural vocoders on cpus, *Acoustical Science and Technology* **43**(2): 121–124.

Miao, Liang, Chen, Ma, Wang and Xiao (2020). Flow-tts: A non-autoregressive network for text to speech based on flow, pp. 7209–7213.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio.

Oord, Li, Babuschkin, Simonyan, Vinyals, Kavukcuoglu, Driessche, Lockhart, Cobo, Stimberg, Casagrande, Grewe, Seb, Dieleman, Elsen, Kalchbrenner, Zen, Graves, King, Walters, Belov and Hassabis (2018). Parallel wavenet: Fast high-fidelity speech synthesis, **80**: 3918–3926.

Ping, W., Peng, K. and Chen, J. (2018). Clarinet: Parallel wave generation in end-to-end text-to-speech.

Ping, Wei, Peng, Kainan, Gibiansky, Andrew, Arik, Sercan, Ajay, Sharan, Raiman, Jonathan, Miller and John (2017). Deep voice 3: Scaling text-to-speech with convolutional sequence learning.

Pons, J., Pascual, S., Cengarle, G. and Serrà, J. (2021). Upsampling artifacts in neural audio synthesis, pp. 3005–3009.

Prenger, Valle and Catanzaro (2019). Waveglow: A flow-based generative network for speech synthesis.

Ren, Hu, Tan, Qin, Zhao, Zhou and Liu (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech.

Ren, Ruan, Tan, Qin, Zhao, Zhou and Liu (2019). Fastspeech: Fast, robust and controllable text to speech, **32**.

Saeki, Tachibana and Yamamoto (2022). Drspeech: Degradation-robust text-to-speech synthesis with frame-level and utterance-level acoustic representation learning.

Sercan, Mike, Thakkar, Adam, Siahkamari, Gregory, Diamos, Andrew, Yongguo, Xian, John, Ng, Jonathan, Shubho and Shoeybi (2017). Deep voice: Real-time neural text-to-speech.

Shen, Jonathan, Pang, Ruoming, Weiss, Schuster, Jaitly, Yang, Chen, Zhang, Wang, Skerry-Ryan, Saurous, Agiomyrgiannakis and Wu (2017). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.

Terblanche, C., Harty, M., Pascoe, M. and Tucker, B. V. (2022). A situational analysis of current speech-synthesis systems for child voices: A scoping review of qualitative and quantitative evidence, *Applied Sciences* **12**(11): 5623.

Tiomkin, S., Malah, D., Shechtman, S. and Kons, Z. (2011). A hybrid text-to-speech system that combines concatenative and statistical synthesis units, *IEEE Transactions on Audio, Speech, and Language Processing* **19**(5): 1278–1288.

Valin, Isik, Smaragdis and Krishnaswamy (2022). Neural speech synthesis on a shoestring: Improving the efficiency of lpcnet.

Wan, L., Wang, Q., Papir, A. and Moreno, I. L. (2018). Generalized end-to-end loss for speaker verification, pp. 4879–4883.

Wang, Y., Skerry-Ryan, RJ, Stanton, Daisy, Wu, Yonghui, Weiss, J., R., Jaitly, Navdeep, Yang, Zongheng, Xiao, Ying, Chen, Zhifeng, Bengio, Samy, Le, Quoc, Agiomyrgiannakis, Yannis, Clark, Rob, Saurous and A., R. (2017). Tacotron: Towards end-to-end speech synthesis.

Xiao, Zhang and Lin (2022). Dgc-vector: A new speaker embedding for zero-shot voice conversion.

Xue, Deng, Han, Li, Sun and Liang (2022). Ecapa-tdnn for multi-speaker text-to-speech synthesis.

You, J., Kim, D., Nam, G., Hwang, G. and Chae, G. (2021). Gan vocoder: Multi-resolution discriminator is all you need, *arXiv preprint arXiv:2103.05236* .

Zhang, Haitong and Yue (2022). Improve few-shot voice cloning using multi-modal learning.

Zhao, Zhang, Wang, Cheng and Xiao (2022). nnspeech: Speaker-guided conditional variational autoencoder for zero-shot multi-speaker text-to-speech.