SpeechPerfect: Collaboration of Multiple Modules for Perfect Speech Synthesis

Jeihee Cho
Carnegie Mellon University
Pittsburgh, PA 15213
jeiheec@andrew.cmu.edu

Moukhik Misra
Carnegie Mellon University
Pittsburgh, PA 15213
moukhikm@andrew.cmu.edu

Aarya Makwana
Carnegie Mellon University
Pittsburgh, PA 15217
amakwana@andrew.cmu.edu

Tong Jiao
Carnegie Mellon University
Pittsburgh, PA 15213
tongjiao@andrew.cmu.edu

Abstract

Identical speech reconstruction is a scarcely explored problem in the field of speech processing. Speech recognition and speech synthesis are currently more popular avenues of research. Our motivation is that if a model can reconstruct an identical speech only with features extracted from it, these features are shown to be effective in representing this speech signal. In this paper, we derive and improve upon the state-of-the-art in both speech recognition and speech synthesis to present *SpeechPerfect*, an end-to-end speech reconstruction model that fundamentally fuses Automatic Speech Recognition (ASR) and speech synthesis to regenerate the input speech signal. On a high level, the system utilizes phoneme embeddings that are generated by state-of-the-art ASR, and speaker characteristics which are obtained by our proposed speaker embedding extractor module to reconstruct input speech. Our model leverages an encoder-decoder architecture with a variance adapter module. The variance adaptor module predicts the pitch, energy, and duration of input speech and in combination with the speaker embedding is capable of replicating the target speech given only a few seconds of inputs. Notably, this paper proposes a complete, automatic end-to-end speech recognition system, capable of replicating input speech without relying on human efforts. Our model achieves a Mean Opinion Score - Listening Quality Objective (MOS-LQO) score of 3.58 which surpasses the acceptable threshold for speech quality. [1]

1 Introduction

Speech recognition and synthesis are classical problems in the field of signal processing and computational linguistics. Speech recognition involves converting input speech to text and speech synthesis involves converting text to target speech. The complexities of speech recognition lie in dealing with the varied characteristics of human speech such as dealing with accents, speech pitch, quality, tonality, intonations and such. Traditional machine learning approaches to speech recognition have

[1] Division of work: Moukhik Misra (moukhikm): Baseline Creation, Model Creation, Model Evaluation, Documentation - Report, Literature Review; Jeihee Cho (jeiheec): Baseline Creation, Model Code Creation/Execution, Model Training, Documentation - Report, Literature Review, Presentation; Tong Jiao (tongjiao): Baseline Creation, Model Creation, Model Evaluation, Documentation - Report, Model Training; Aarya Makwana (amakwana): Model Creation, Documentation - Presentation, Documentation - Report, Literature Review;

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

included Hidden Markov Model (HMM) (Gales & Young, 2007) and Dynamic Time Warping (Amin & Mahmood, 2008). More recent approaches to Speech Recognition have involved using CNN's, RNN's and Attention and Transformer models. Newer approaches have even considered End-to-End Speech Recognition that removes the need for acoustic, pronunciation and language models and directly converts speech to text. Some of these advances in Speech Recognition are models such as *wav2vec* (Schneider et al., 2019) and its extension *wav2vec 2.0* (Baevski et al., 2020) and Conformer based models for Automatic Speech Recognition (ASR) (Gulati et al., 2020).

Speech synthesis on the other hand refers to the process of generating speech from text with the need for understanding context in text and grammar and converting the same to speech signals. Traditional speech synthesis is a complex and multi-step process with steps such as sound wave generation and phonetic analysis. Traditional approaches to speech synthesis include Parametric Speech Synthesis using HMMs (Zen, 2015) and Concatenative Speech Synthesis (Longster, 2003). Recent state-of-the-art approaches to speech synthesis are, however, all rooted in deep learning. Some of the best speech synthesis methods currently are *Tacotron2* (Shen et al., 2018) and *FastSpeech2* (Ren et al., 2022).

With the objective of simplifying language problems to simple text-to-text problems, (Raffel et al., 2023) came up with the T5 framework which stands for Text-To-Text Transfer Transformer. Adapting this to speech tasks, (Ao et al., 2022) developed *SpeechT5*, consisting of a shared encoder-decoder network capable of voice conversion, ASR, Text to Speech (TTS), and more. However, there exists a flaw in *SpeechT5*, it requires the use of external speaker embeddings for TTS which makes it difficult to reconstruct of original speech.

To the best of our knowledge, not too many attempts have been made to improve end-to-end speech reconstruction i.e., replication of target input speech. To pave the way for advancements in the field, we aim to reconstruct the input speech and replicate it perfectly in a complete end-to-end system. We achieve this by using a multi-step process that utilizes an encoder-decoder network with the fundamental architecture closely resembling that of a fused ASR and TTS system. In an age where video conferences and large multimedia content exchange are commonplace, efficiently transmitting high-fidelity speech over networks remains a challenge. Our end-to-end speech reconstruction model facilitates faster and more compact transmission of audio files by extracting compact representations that can accurately resynthesize the original speech at the receiving end.

In our project, we present an enhanced end-to-end fused speech recognition and speech synthesis model, *SpeechPerfect*, building upon the FastSpeech2 framework. The novel aspect of our architecture is the integration of speaker-specific information in the decoding process, enabling the generation of speech that closely matches the input. Traditional models often concentrate on either generating speech from text or modifying speech style, which typically falls short in replicating an input speech's exact characteristics. Our model addresses this by incorporating an additional Characteristics Extractor module, employing a ConvGRU model for extracting unique speaker embeddings from raw audio waveforms. This module, in tandem with a Variance Adaptor, enriches the phoneme hidden sequence with specific attributes like pitch, energy, and speaker characteristics, leading to a more accurate reproduction of the input speech. The overall architecture and its key components, including a phoneme embedding layer, ASR module, encoder-decoder structure, and a MelGAN-based vocoder, collectively contribute to the model's ability to reconstruct an identical waveform from the given speech input. This approach signifies a substantial advancement in speech synthesis, focusing on preserving the unique attributes of the input speech.

We perform our training and evaluations using the VCTK dataset. The qualitative aspect of our evaluation involves listening tests to assess the naturalness and speaker similarity of the generated samples. For quantitative evaluation, we employ the VisQOL metric, generating MOS-LQO scores to objectively measure speech quality. Our model's loss function, focuses on pitch, energy, and duration and is instrumental in enhancing speech naturalness and accuracy. Through our evaluations, we aim to demonstrate *SpeechPerfect*'s ability in preserving the unique characteristics of the input speech thereby enabling effective reconstruction of input speech.

The main contribution can summarized as follows:

- We introduce a speech reconstruction module that requires no other inputs other than speech and text.
- We leverage the effective speaker characteristics extractor module with an encoder-decoder network and variance adaptor to enhance the reconstruction performance.

- We will conduct extensive experiments using VCTK dataset which consists of multi-accented utterances of 109 native English speakers.

2 Related Work

Enhancing speech synthesis performance dragged the attention of the researchers as it can be applied to diverse downstream tasks such as speech enhancement, voice conversion, and more. The foundation of our paper is grounded in an investigation of papers related to ASR and TTS.

Automatic Speech Recognition Traditionally, researchers focus on generating speech representations for an automatic transcript or phoneme generation. The authors of *wav2vec* (Schneider et al., 2019) suggested an unsupervised speech representation learning method using a contrastive predictive coding objective. It uses a convolutional neural network that takes raw audio as input and computes generate representation to give input as a speech recognition module. Moreover, building upon this foundation, *wav2vec 2.0* (Baevski et al., 2020) and *ReVISE* (Hsu et al., 2022) learn powerful speech encodings from raw audio in a completely unsupervised manner. Specifically, *ReVISE* contains a shared speech encoder, separate audio and visual decoders, and a contrastive loss between the two modality decoders to align the learned representations.

Text-to-Speech To generate realistic speech that goes along with the given text, researchers focus on producing diverse styles of speech by predicting the characteristics of speech. In *FastSpeech* (Ren et al., 2019) and *FastSpeech2* (Ren et al., 2022), they employ a variance predictor that predicts energy, pitch, and duration, which is trained by comparing the ground truth values with mean square error (MSE) loss. With predictor coupled with the phoneme embeddings, the researchers proposed a successful framework that generates speech with the given transcript.

SpeechT5 For ultimate general speech *SpeechT5* (Ao et al., 2022), a unified encoder-decoder model pre-trained on large amounts of unlabeled speech and text data. *SpeechT5* converts spoken language tasks like ASR, TTS, and speech translation into a speech-to-speech framework. It uses a shared encoder-decoder backbone and modal-specific pre/post-nets. The model is pre-trained using a denoising sequence-to-sequence method and a novel cross-modal vector quantization approach to align textual and acoustic representations. However, this work failed to provide a framework for the reconstruction of speech as they focused on extracting general speech or text representation for downstream tasks such as TTS or ASR. Furthermore, an examination of the influence of speaker embedding on the vocoder was not undertaken, thereby resulting in a lack of comprehensive integration within the entirety of the end-to-end system for speech generation.

The aforementioned studies require additional components or datasets besides the input speech to reconstruct the target speech. To tackle this challenge, we introduce *SpeechPerfect*, an advanced end-to-end speech reconstruction framework designed to account for the inherent characteristics of both speech and transcript automatically. This is achieved through the integration of innovative ASR and TTS modules within the proposed framework.

This paper is organized as follows: After describing the baseline in Sec. 3, we present our system model in Sec. 4.

3 Baseline

We introduce a baseline that combines *FastSpeech2* with speaker table-based speaker embedding for speaker-targeted speech generation. With this baseline, we can show the efficacy of our proposed framework in relation to the established methodology.

Original *FastSpeech2* does not accept either speech embedding or speaker embedding, while we found github repository that combined table-based speaker embedding with *FastSpeech2* module ga642381 (2022). As the original *FastSpeech2* does not consider the speaker information, it just predicts the speech based on the given input text, and the loss function guides to the real speech. However, modification of *FastSpeech2* takes the speaker into account that the model can learn the characteristics of the speech in a better way. We decided to use this modified *FastSpeech2* as a baseline due to this reason. The overall architecture of the baseline is provided in Fig. 1.

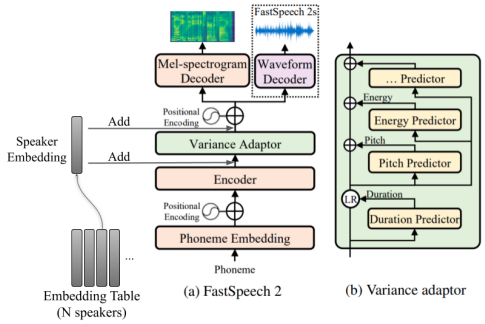


Figure 1: The overall architecture of the baseline.

4 SpeechPerfect

On top of the illustrated baseline, we added a component to provide information about the speaker to the decoder to generate speech that is identical to the given speech. The existing baselines normally focus on either generating speech with given text by predicting characteristics of speech or modifying the style of the speech, which eventually fails to generate speech identical to the given input. Therefore, by extracting additional features from the speech, we aim to reproduce the entered speech.

4.1 Model Overview

The overall architecture of our model is shown in Fig. 2. Our model is an improvement of FastSpeech2. The model accepts a piece of speech waveform as an input and aims to reconstruct an identical waveform as an output. First, the phoneme embedding layer takes in a phoneme sequence, which is extracted from the original input waveform by the Automatic Speech Recognition (ASR) module. The phoneme sequence is then converted to a phoneme hidden sequence by the encoder, after applying positional encoding to it. Next, the variance adaptor adds different speaker-specific information to the phoneme hidden sequence, such as speaker embedding (from the characteristics extractor), duration, pitch, and energy. Finally, the decoder converts the hidden sequence with added information into a mel-spectrogram sequence, and MelGAN module transfers that into the reconstructed output.

4.2 Model components description

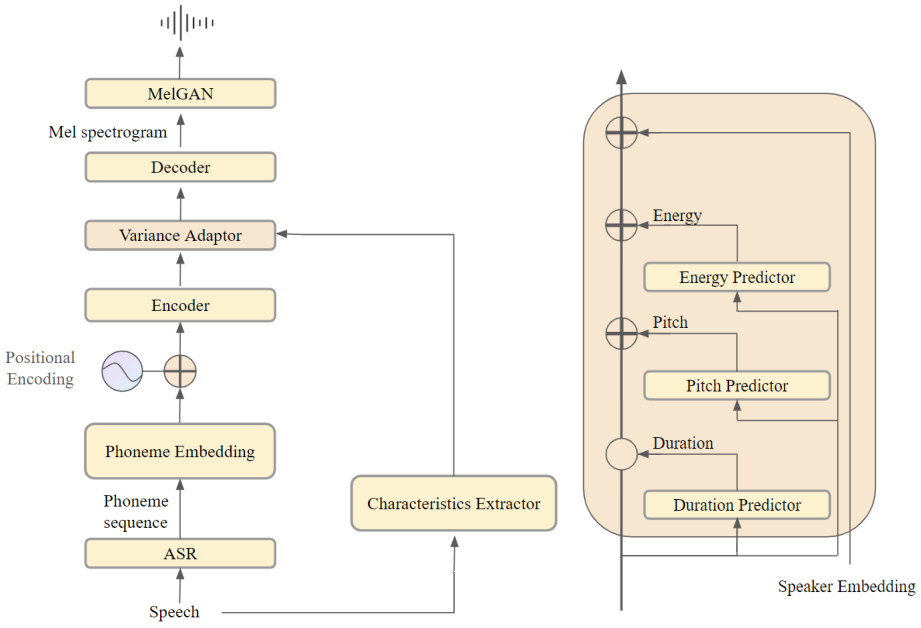
ASR *wav2vec* (Schneider et al., 2019) helps us to get a phoneme sequence $(x_1, \ldots, x_n)$ from the original input.

Phoneme Embedding This module converts a sequence of phonemes $(x_1, \ldots, x_n)$ into a sequence of vectors with dimension $d = 256$ and then applies positional encoding to this sequence to get a phoneme hidden sequence $(y_1, \ldots, y_n)$, with $y_i \in \mathbb{R}^d$. The positional encoding used is the same as that used by Transformer (Vaswani et al., 2023).

Encoder and decoder Encoder and decoder architecture in our model is similar to that in *FastSpeech2* (Ren et al., 2022), they comprise a stack of feed-forward Transformer blocks and a 1D-convolution layer. After the encoder module, the sequence length is unchanged and each element in the sequence has $d = 256$ dimensions. The decoder module converts the input sequence to a mel spectrogram.

Characteristics Extractor (Speaker Embedding Extractor) To produce a piece of speech that is as similar as possible to the input, the model needs to capture some information that is specific to the input. This module and the Variance Adaptor module aim to capture this kind of information. This module focuses on information that is not specific to each phoneme in the input sequence, while the Variance Adaptor calculates features both input-specific and phoneme-specific and aggregates them with results from this module.

The Characteristics Extractor module calculates a 256-d speaker embedding and passes it to Variance Adaptor for future use. It does so using a ConvGRU model that extracts speaker embeddinggs from



(a) *SpeechPerfect* (b) Variance Adaptor

Figure 2: The overall architecture of *SpeechPerfect*. Subfigure (b) expands the structure of the variance adaptor.

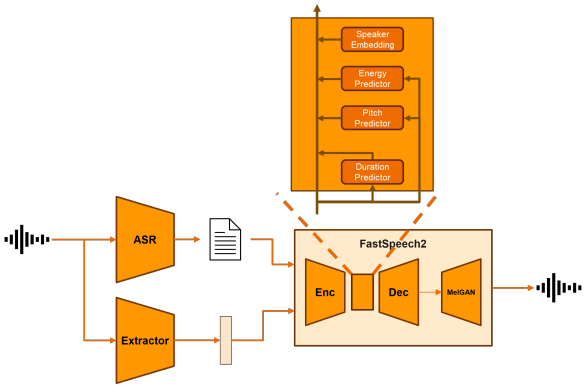


Figure 3: The overall flow of *SpeechPerfect*.

raw unonormalized waveforms. The convolutional layers which are the initial layers are responsible for extracting feature representations from the raw audio waveform. Convolutional layers are effective in handling spatial hierarchies and local patterns in data, making them suitable for initial feature extraction in audio processing.

Following the convolutional layers, the network employs GRU layers, specifically three layers with 786 hidden units each. GRUs are a type of recurrent neural network (RNN) that are efficient in handling sequential data, like audio. They are known for their ability to capture dependencies over different time scales, making them ideal for processing the features extracted by the convolutional layers.

This combination of convolutional layers for feature extraction and GRU layers for sequential data processing makes the ConvGRU model particularly adept at generating speaker embeddings from raw audio waveforms. These embeddings are compact, 256-dimensional vectors that effectively capture the unique characteristics of a speaker's voice.

Variance Adaptor The goal of the variance adaptor is to aggregate the characteristics that are specific to a certain piece of speech, including pitch, energy, duration, and speaker-specific information. By adding this information to the phoneme hidden sequence, the decoder is capable of generating a representation that is specific to a certain input and making it as similar to the input as possible. The original variance adaptor in our baseline contains three predictors capturing three kinds of information: duration, pitch, and energy for each phoneme z_i in the phoneme hidden sequence. Each predictor is made up of two 1D-convolution layers and a linear layer.

The input to duration predictor is the phoneme hidden sequence $(y_1, \ldots, y_n)$, and the duration predictor predicts how many mel frames correspond to each frame. After knowing that, we expand the phoneme hidden sequence to a sequence $(z_1, \ldots, z_m)$ with a length m equal to the number of frames of the input to prepare for the next two predictors. The pitch predictor and energy predictor take the expanded sequence as input, and predict 1) pitch F_0 for each frame, and 2) L2-norm of the amplitude of each Short-Time Fourier Transform (STFT) frame correspondingly. Predicted results are embedded into a $d = 256$ dimension vector and added to corresponding z_i in the expanded sequence. The result from the characteristics extractor is added to every z_i. In sum, each element in the expanded sequence is calculated by

$$z_i = y_{phoneme} + p_i + e_i + s$$

Where $y_{phoneme}$ is the corresponding phoneme of z_i in the phoneme hidden sequence, p_i is the result from pitch predictor for the i^{th} frame, e_i is the result from energy predictor for the i^{th} frame and s is speaker embedding from the characteristics extractor.

MelGAN To convert the predicted mel spectrogram created by the decoder module into an actual speech waveform, we used a vocoder based on MelGAN (Ao et al., 2022).

5 Evaluation Plan

We compare speech reconstructed by SpeechPerfect against the original FastSpeech2 (Ren et al., 2022). Since multiple speakers are involved in our dataset and Ren et al do not mention how to handle multiple speakers, we give the ground truth of the speaker identification to the baseline model during training (our model uses speaker embedding from the Characteristics Extractor to distinguish speakers). As a result, the performance of the baseline on speakers that are seen in the training set is expected to be better than that of our model.

5.1 Dataset

In our training and evaluation, we utilized the CSTR VCTK Corpus dataset. The dataset is a comprehensive collection of speech data from 109 English speakers, each showcasing a unique accent. The dataset encompasses approximately 400 sentences per speaker, drawn from a variety of sources including newspapers, the Rainbow Passage, and an elicitation paragraph from the speech accent archive. This diversity in content ensures a broad representation of linguistic and phonetic elements, enabling us to accurately test and validate our model. We utilize 43908 training instances and 108 test instances for training and evaluating our model. There are very few samples in the test set because testing relies on an external ASR module and it is time-consuming to do this task on the validation after each epoch.

5.2 Qualitative Analysis

We listened to speech samples generated by FastSpeech2 and SpeechPerfect. Qualitatively, the audio from our model sounds more natural in terms of tone and rhythm compared to FastSpeech2. However, both models sound similar to the original speaker (FastSpeech2 outputs are more similar), indicating that solely relying on speaker embeddings is insufficient for high-fidelity reconstruction. This motivates future work to capture more features besides speaker embedding to reconstruct the audio better.

5.3 Quantitative Metrics

First, we quantify speech reconstruction performance using the Visual Speech Quality Objective Listener (VisQOL (Hines et al., 2015)) metric. VisQOL compares the original input audio against

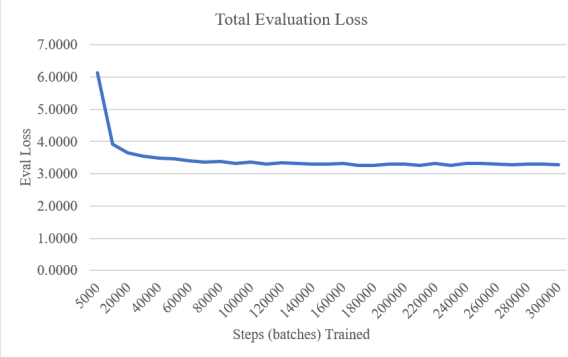


Figure 4: Total evaluation loss

the reconstructed degraded audio and outputs a Mean Opinion Score - Listening Quality Objective (MOS-LQO) from 1 (worst) to 5 (best).

We compute MOS-LQO on the VCTK test set for SpeechPerfect. Higher scores indicate better preservation of the input speech qualities like naturalness, speaker similarity, and intelligibility.

5.4 Loss Function Analysis

The loss function of SpeechPerfect primarily focuses on three key aspects: pitch, energy, and duration. These elements are essential in determining the naturalness and accuracy of the generated speech. The mathematical formulae for the loss function of SpeechPerfect can be expressed as:

1. **Pitch Loss (L_{pitch})**:

$$L_{pitch} = MSE(Pitch_{predicted}, Pitch_{actual})$$

2. **Energy Loss (L_{energy})**:

$$L_{energy} = MSE(Energy_{predicted}, Energy_{actual})$$

3. **Duration Loss ($L_{duration}$)**:

$$L_{duration} = MSE(Duration_{predicted}, Duration_{actual})$$

The total loss (L_{total}) is a sum of these components:

$$L_{total} = L_{pitch} + L_{energy} + L_{duration}$$

The quantitative and qualitative evaluations analyze

- baselines like FastSpeech2 versus end-to-end models, and
- the role of speaker embeddings.

This comprehensive assessment demonstrates SpeechPerfect's efficacy at reconstructing the input signal. Figure 3 demonstrates the total loss on the evaluation set. According to it, the model converges after 120000 batches. The evaluations in the next section is based on the checkpoint saved after 120000 batches.

6 Results and Discussion

6.1 Results

In the results and discussion section of our paper, we present a detailed analysis of the objective quality assessment for the speech synthesis model we developed. Our evaluation is based on a subset of 108 test samples derived from the VCTK corpus, which is a diverse speech dataset. These

samples are high-fidelity waveform files, each sampled at a frequency of 22kHz. The output from our proposed model yields two main components: the generated mel-spectrogram and the corresponding waveform audio file, also at a 22kHz sampling rate. Also, it should be noticed that the baseline we are comparing to has access to the ground truth of the speaker ID during training, so it is expected to perform better than our model since it gets all speaker-specific information and our goal is to achieve the same performance as the baseline. The same performance as the baseline means that the speaker embedding is as effective as the ground truth (speaker ID).

For objective quality evaluation, we employed the ViSQOL framework, which necessitates that both the ground truth and the synthetically generated wav files be downsampled to a 16kHz sampling rate. This preprocessing step aligns with the standard input requirements of the ViSQOL model and ensures compatibility with its internal metrics for assessing speech quality.

The Mean Opinion Score - Listening Quality Objective (MOS-LQO) is utilized as a benchmark to quantify the perceived audio quality. This metric provides a numerical indication of the quality, which is instrumental in comparing the synthesized speech against the ground truth. The MOS-LQO scores are designed to mimic the subjective judgments that would be made by human listeners.

Below, we provide a comprehensive overview of the MOS-LQO scores for a representative subset of 20 test samples from our analysis. These scores serve as an objective indicator of the performance of our speech synthesis model, offering insights into the fidelity and clarity of the generated audio in comparison to the original recordings. The results underscore the effectiveness of our model and shed light on areas where further refinement could be beneficial. This detailed examination not only benchmarks the current state of our system but also sets a precedent for subsequent improvement and optimization efforts in the domain of high-quality speech synthesis.

Ground Truth Audio File	Generated Audio File	MOSLQO
gt_audio/p234_349.wav	syn_audio/p234_349.wav	3.772115
gt_audio/p233_043.wav	syn_audio/p233_043.wav	4.058830
gt_audio/p275_030.wav	syn_audio/p275_030.wav	3.963712
gt_audio/p323_141.wav	syn_audio/p323_141.wav	3.387343
gt_audio/p276_131.wav	syn_audio/p276_131.wav	3.920014
gt_audio/p343_078.wav	syn_audio/p343_078.wav	3.666159
gt_audio/p333_223.wav	syn_audio/p333_223.wav	3.526254
gt_audio/p247_239.wav	syn_audio/p247_239.wav	3.957340
gt_audio/p336_164.wav	syn_audio/p336_164.wav	3.592371
gt_audio/p244_117.wav	syn_audio/p244_117.wav	4.288914
gt_audio/p231_433.wav	syn_audio/p231_433.wav	3.867531
gt_audio/p273_058.wav	syn_audio/p273_058.wav	3.652116
gt_audio/p288_355.wav	syn_audio/p288_355.wav	3.771710
gt_audio/p314_340.wav	syn_audio/p314_340.wav	3.976726
gt_audio/p306_330.wav	syn_audio/p306_330.wav	3.464264
gt_audio/p295_229.wav	syn_audio/p295_229.wav	3.981844
gt_audio/p278_029.wav	syn_audio/p278_029.wav	3.762528
gt_audio/p277_293.wav	syn_audio/p277_293.wav	4.046256
gt_audio/p252_043.wav	syn_audio/p252_043.wav	3.733566
gt_audio/p239_054.wav	syn_audio/p239_054.wav	3.892729

Table 1: MOSLQO scores for synthesized audio compared with ground truth

Metric	Score
MOSLQO	3.58
WER (ground truth)	0.09
WER (baseline)	0.14
WER (our model)	0.24

Table 2: Metric Scores

The interpretation and discussion of Mean Opinion Score - Listening Quality Objective (MOS-LQO) scores are an essential part of the evaluation of speech synthesis systems, as seen in the reference to

Table 1. MOS-LQO is a quantifiable metric derived from the ViSQOL—a model that predicts the subjective quality of audio samples based on human listening experiences. The scores range from 1 to 5, with higher scores indicating better perceived audio quality.

In the context of the results shown in Table 1, each score represents the quality assessment of an individual audio sample, with these scores reflecting the degree to which the synthesized speech approximates the original, or 'ground truth,' recordings. The average MOS-LQO score over all 108 training samples, as detailed in Table 2, stands at 3.58. This average is an important indicator, as it suggests that the overall perceived quality of the synthesized audio is consistently above the threshold of acceptability.

The benchmark for ViSQOL scores considered acceptable is typically above 3.0, which aligns with a fair quality of audio that most listeners would find satisfactory. Our model's average score exceeds this benchmark, which suggests that the synthesized speech is of a quality that would generally be regarded as good by listeners. This is an important achievement in the field of speech synthesis, as it not only validates the effectiveness of the model but also indicates that the synthesized speech is likely to be well-received by users in practical applications.

However, it is important to note that while the average score is above the acceptability threshold, individual variations still exist. Some samples score significantly higher, indicating excellent synthesis, while others might just cross the threshold, pointing toward potential areas for improvement. Detailed analysis of individual scores, particularly those that fall at the lower end of the spectrum, provide valuable insights into specific aspects of the synthesis process that may require further refinement.

Another measure of the quality of the generated speech is the word error rate (WER). It is defined as the ratio of errors in a transcript to the total number of words. We use Microsoft Azure's speech recognition service to transcribe the ground truth speech and generated speech. In Table 1, WER of ground truth audio, baseline's generated audio, and our model's generated audio are presented. Though being worse than the baseline is expected, the performance gap in WER between our model and baseline is relatively large, indicating that speaker embedding cannot fully capture all speaker-specific information, and using speaker embedding will result in a larger loss in the fidelity of the audio. A potential cause of the loss in fidelity may be that all information, including duration, pitch, energy, and speaker embedding are added to the intermediate representation. This will cause a mix of information and a better way to represent intermediate states inside the variance adaptor is possible.

In conclusion, the model demonstrates a commendable level of performance in generating speech that is perceived as high quality by the standards of objective measurement. This is indicative of its potential utility in real-world scenarios where synthesized speech is expected not just to be intelligible, but also to have a degree of naturalness and clarity that approaches that of natural human speech.

In Figure 3 (see 5), we present a comparative visualization of the mel spectrograms, illustrating both the ground truth and the generated output from our synthesized speech model. This side-by-side comparison effectively highlights the similarities and differences between the original audio's spectral features and those of the synthesized counterpart, offering a clear visual representation of the model's performance in replicating the intricate characteristics of human speech.

6.2 Discussion

A key limitation in our setup arises from ViSQOL's design, which accepts audio samples at a 16kHz sample rate, while our recordings are at a finer 22kHz. To align with ViSQOL, we downsample our audio, potentially losing some high-frequency details. This might result in lower MOS-LQO scores, as the downsampled speech may not fully capture the richness of the 22kHz originals. It's important to factor in this aspect when interpreting the MOS-LQO scores.

Additionally, the extraction of speaker embeddings in our model might lead to the loss of certain speech nuances, impacting the naturalness and fidelity of the generated audio. This could affect both objective measures like ViSQOL and the subjective perception of listeners. The choice of vocoder is also pivotal. Each vocoder has its strengths and weaknesses in replicating natural speech. We've used MelGAN, but we believe that switching to HiFi-GAN may yield better results

Our discussion acknowledges the technical and methodological factors that shape the quality of synthesized speech. By exploring these factors, we aim to refine our approach, ensuring that our model meets not just ViSQOL's objective standards but also the subjective expectations of listeners.

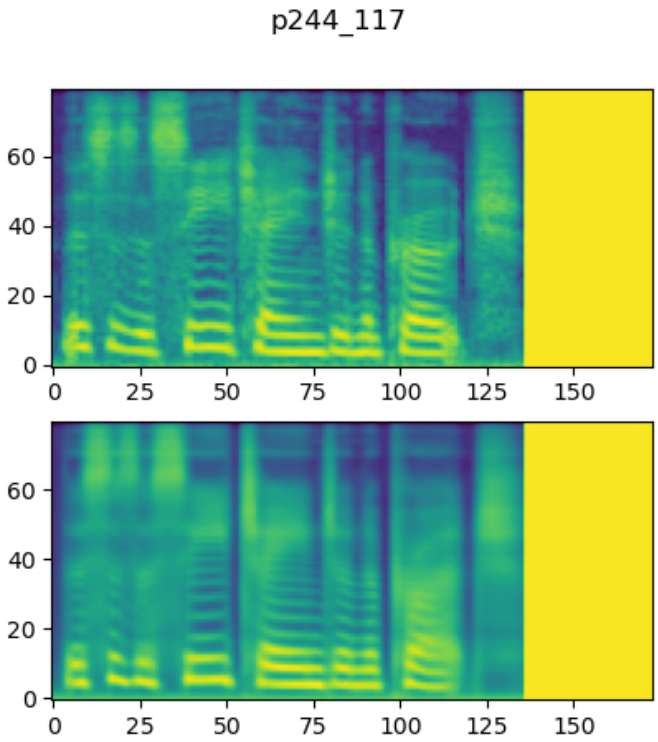


Figure 5: a) Top image: GT Spectrogram b) Bottom Image: Generated Spectrogram

7 Conclusion

In conclusion, *SpeechPerfect* effectively fuses the functionality of both ASR and TTS systems to present a functioning end-to-end speech reconstruction and synthesis model. We have successfully presented a novel approach to reconstructing target input speech. In our approach, we first extract speaker embeddings using a ConvGRU model. This, in combination with our modified FastSpeech2-based architecture and custom Variance Adaptor module, enables the effective generation of identical speech.

Our evaluations, utilizing the VCTK dataset, have provided insights into the model's ability to preserve the naturalness and speaker characteristics of the input speech. We achieve an average MOS-LQO score of 3.58, which meets the ViSQOL standard for acceptable speech quality. This system is capable of replicating input speech autonomously, without the need for human intervention. The implications of this work offer new perspectives and capabilities in the field of speech processing, and can set the stage for future advancements in the efficient reconstruction of speech.

8 Future Work

While this work demonstrates promising speech reconstruction capabilities, there are several fruitful avenues for further improvement:

- We can enrich the training data volume to boost zero-shot generalization performance across diverse unseen speakers.
- We can possibly explore the option of switching our vocoder to HiFi-GAN and methods of improving zero-shot performance.
- There also also a scope for conducting an in-depth analysis comparing the compact phoneme vector size against original waveform file size can reveal insights into bandwidth savings for practical applications.

- This model can enable speech compression for efficient transmission in diverse domains like media, customer service calls, video conferencing platforms, and more.
- Additional analysis can be carried out for reconstruction performance across various speech attributes like accent, pitch, tempo etc can uncover areas needing refinement.
- Moreover, extending and evaluating this approach by training the ASR module on non-English languages can expand the study's adoption.

In essence, while SpeechPerfect makes an important contribution demonstrating end-to-end speech reconstruction capability, exciting opportunities exist for both strengthening the model and expanding its impact across practical applications and languages. Targeted data expansion, comprehensive analytical evaluation, and cross-lingual support offer rich yet unexplored directions for future work.

References

Amin, T. B. and Mahmood, I. Speech recognition using dynamic time warping, 2008. URL https://api.semanticscholar.org/CorpusID:22123601.

Ao, J., Wang, R., Zhou, L., Wang, C., Ren, S., Wu, Y., Liu, S., Ko, T., Li, Q., Zhang, Y., Wei, Z., Qian, Y., Li, J., and Wei, F. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing, 2022.

Baevski, A., Zhou, H., rahman Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL https://api.semanticscholar.org/CorpusID:219966759.

ga642381. Fastspeech2, 2022. URL https://github.com/ga642381/FastSpeech2/tree/main.

Gales, M. J. F. and Young, S. J. The application of hidden markov models in speech recognition, 2007. URL https://api.semanticscholar.org/CorpusID:51039442.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. Conformer: Convolution-augmented transformer for speech recognition, 2020.

Hines, A., Skoglund, J., Kokaram, A., and Harte, N. Visqol: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015 (13):1–18, 2015.

Hsu, W.-N., Remez, T., Shi, B., Donley, J., and Adi, Y. Revise: Self-supervised speech resynthesis with visual input for universal and generalized speech enhancement, 2022.

Longster, J. Concatenative speech synthesis : a framework for reducing perceived distortion when using the td-psola algorithm, 2003. URL https://api.semanticscholar.org/CorpusID:46223196.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T. Fastspeech: Fast, robust and controllable text to speech. *CoRR*, abs/1905.09263, 2019. URL http://arxiv.org/abs/1905.09263.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. Fastspeech 2: Fast and high-quality end-to-end text to speech, 2022.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. wav2vec: Unsupervised pre-training for speech recognition. *CoRR*, abs/1904.05862, 2019. URL http://arxiv.org/abs/1904.05862.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.

Zen, H. Statistical parametric speech synthesis: from hmm to lstm-rnn, 2015. URL https://api.semanticscholar.org/CorpusID:36067578.