Dynamic Fine-Tuning and Adaptive Reference Selection for StyleTTS2 in Speaker-Specific Voice Synthesis Using Multi-Source Audio Data

Mohammad Zubair Khan[*]

Katz School of Science and Health, Yeshiva University, New York, NY, USA

Abstract. This paper introduces a novel, fully automated data-to-training pipeline and a dynamic fine-tuning methodology for the StyleTTS2 model, aimed at synthesizing high-fidelity, speaker-specific voices from heterogeneous audio sources. Our approach leverages data collected from multiple channels, including YouTube videos, podcast interviews, and Zoom-meeting recordings, to build comprehensive speaker datasets. We automate the entire pipeline from raw audio scraping to training data preparation: speech diarization is applied to isolate the target speaker, segments are split at natural voice breaks into coherent sentence-level chunks, and dynamic transcription and phoneme conversion are performed. After fine-tuning StyleTTS2 for specific speakers, we introduce a dynamic reference speech selection module that utilizes cosine similarity between input text and training transcripts to adaptively choose the most contextually relevant reference sample. Additionally, we propose a custom evaluation framework that compares synthesized and reference audio at the word level for more accurate fidelity and intelligibility assessments. Finally, we explore a knowledge-distillation approach by creating a reduced student model with half the number of hidden layers, though this did not yield improved performance. Experiments demonstrate the effectiveness of our pipeline and adaptive techniques, achieving natural, speaker-specific TTS outputs and providing valuable insights into advanced voice cloning workflows. The implementation is available at https://github.com/CasterShade/Voice-cloning-TTS.

1 Introduction

High-quality, speaker-specific text-to-speech (TTS) systems have become increasingly important in applications requiring voice personalization and authenticity. As TTS technology progresses, demand has risen across various fields—from conversational AI and virtual assistants to digital content creation and audio forensics—for TTS systems capable of producing realistic, expressive, and speaker-specific voices [4, 17]. With the evolution of neural network-based models, especially end-to-end architectures, the synthesis of natural-sounding, expressive speech has become more attainable, albeit often at high computational and data costs [13].

In recent years, advancements in neural TTS models have significantly enhanced the ability to generate high-quality speech with minimal human intervention. However, developing speaker-specific TTS systems, particularly for unique voices such as celebrities or individual users, remains a complex task. The primary challenges lie in data availability, preprocessing requirements, and the need for efficient training processes to capture each speaker's unique vocal characteristics [7]. This paper addresses these issues by leveraging the StyleTTS2 model, a sophisticated TTS model that incorporates style diffusion and adversarial training techniques to produce high-quality, expressive speech [12]. To facilitate the synthesis of speaker-specific voices, we developed an automated data pipeline that extracts, separates, and preprocesses audio data from YouTube videos.

This paper presents a holistic, fully automated pipeline for speaker-specific TTS using the StyleTTS2 model [12], contributing several novel aspects:

1. **Automated Multi-Source Data Preparation**: We collect speech data from heterogeneous sources, including YouTube videos of interviews, podcast recordings, and Zoom meeting audios. A suite of automated preprocessing steps—including speaker diarization, voice activity segmentation into sentence-level chunks, transcription, and phoneme generation—produce ready-to-use training datasets. This automation reduces the manual effort and enables rapid adaptation to new target speakers.
2. **Dynamic Fine-Tuning of StyleTTS2**: By fine-tuning StyleTTS2 on these prepared datasets, we adapt the model to produce speech closely mimicking the target speaker's unique characteristics, even with limited and noisy initial data.
3. **Adaptive Reference Selection Using Cosine Similarity**: After fine-tuning, a novel dynamic reference speech selection mechanism is introduced. By comparing the input text to the transcriptions of training samples, we identify the closest contextual match using cosine similarity, ensuring that each synthesis leverages a stylistically and contextually relevant reference sample.
4. **Custom Word-Level Evaluation Metrics**: We propose a new evaluation approach that breaks down synthesized and ground-truth audio into word-level segments using WhisperX. By comparing corresponding words, we achieve a finer-grained evaluation of fidelity, intelligibility, and phoneme-level accuracy, offering more precise insights than traditional utterance-level metrics.
5. **Student Model Exploration**: We also attempt to distill the knowledge from the fine-tuned StyleTTS2 into a smaller student model with half the number of hidden layers to improve efficiency. While this did not yield improved performance, it provides a direction for future research into lighter, more efficient speaker-specific TTS models.

[*] Corresponding Author. Email: mkhan10@yu.edu.

Our results highlight that the integrated pipeline and dynamic reference selection method significantly enhance speaker identity retention and naturalness. Moreover, the custom evaluation metrics underscore the improvement in phoneme fidelity and intelligibility. Although the reduced student model attempt was not successful, our primary contributions provide a robust foundation for future work.

2 Related Work

The field of text-to-speech (TTS) synthesis has witnessed significant advancements due to the integration of deep learning. Early neural TTS models like Tacotron 2 [15] and WaveNet set new benchmarks in naturalness by combining sequence-to-sequence models with mel-spectrogram-based synthesis. Tacotron 2, for instance, achieved high-quality, human-like speech synthesis but had limited control over expressive variability. The recent introduction of StyleTTS2 [12] builds upon these models by utilizing style diffusion and adversarial training with large speech language models, providing capabilities for expressive style transfer and human-level speech quality. Wang et al. [18] further contributed to this line of research with global style tokens for unsupervised style modeling, aligning well with StyleTTS2's emphasis on expressive style transfer.

Barakat et al. [4] offer a comprehensive review of expressive TTS approaches, underscoring the importance of speech synthesis systems that capture emotional nuance and speaking style. They emphasize the need for models that balance interpretability and quality, a challenge especially prevalent in applications where speaker identity and emotional fidelity are essential. This review categorizes various models and highlights key challenges in expressive synthesis, aligning with our focus on balancing fidelity, interpretability, and computational efficiency in speaker-specific voice synthesis.

Voice cloning and multi-speaker TTS have also become critical research areas, particularly for applications where personalized, speaker-specific synthesis is required. Barrington et al. [5] address the pressing issue of cloned voice detection, noting the risks posed by deepfake audio in domains like financial security and disinformation. Their research compares various approaches for distinguishing between real and cloned voices, emphasizing the need for robust speaker verification methods. This focus on speaker identity preservation is directly relevant to our work with StyleTTS2, where speaker-specific voice synthesis must retain identity characteristics for authenticity. Arik et al. [1] support this focus by highlighting neural voice cloning techniques, such as speaker adaptation and speaker encoding, foundational to creating speaker-specific models with limited data.

Recent advances in end-to-end TTS are reviewed by Mu et al. [13], who highlight the progress in deep learning-based speech synthesis. The study outlines key modules in TTS pipelines—text front-end, acoustic model, and vocoder—and discusses challenges in model efficiency and quality. Additionally, the survey covers zero-shot adaptation, an approach in TTS where models are trained to generalize to unseen speakers. Jia et al. [9] demonstrate the feasibility of transfer learning from speaker verification to multi-speaker TTS synthesis, achieving high fidelity in speaker adaptation. Our work builds on this approach by implementing dynamic reference selection using cosine similarity, ensuring that StyleTTS2 can adaptively match the target speaker's style even with limited data.

Furthermore, Hu and Zhu [7] propose a real-time voice cloning system that incorporates advanced preprocessing techniques, such as noise reduction and pronunciation adjustment for non-standard words. Their emphasis on clean and intelligible speech is aligned with our own preprocessing pipeline, where speaker separation, denoising, and volume normalization ensure the quality of training data for fine-tuning. By implementing a cosine similarity metric for reference selection, we enhance StyleTTS2's adaptability to different speaker characteristics, a crucial feature for applications requiring high-quality, speaker-specific synthesis.

Kim et al. [10] examine few-shot learning in TTS and emphasize the benefits of linear probing and full fine-tuning techniques when data is limited. Their findings are relevant to our approach, as we rely on fine-tuning StyleTTS2 with scraped YouTube audio, where the availability of clean, labeled data is often constrained. Similarly, FastDiff [8] has introduced methods for fast and high-quality speech synthesis using conditional diffusion models, offering solutions to the speed and efficiency issues inherent in traditional TTS. While FastDiff focuses on efficiency, our approach prioritizes dynamic reference selection and expressive fidelity. Azizah and Jatmiko [2] contribute additional insights on transfer learning, style control, and speaker reconstruction in zero-shot multilingual, multi-speaker TTS, paralleling challenges we address in our fine-tuning process.

Lastly, Latorre et al. [11] investigate the impact of data reduction on neural TTS performance, demonstrating that multi-speaker models can compensate for limited data. Our work complements this finding by leveraging a dynamic reference selection system that can optimize voice adaptation even with sparse data. Gudmalwar et al. [6] explore VECL-TTS, a cross-lingual TTS system with voice identity and emotional control, emphasizing the importance of style and identity retention in multi-speaker and cross-lingual contexts, which aligns with the objectives of our system. Through the integration of various preprocessing, adaptive selection, and evaluation techniques, we advance the capabilities of StyleTTS2, enabling it to synthesize personalized, high-fidelity speech for specific speakers.

Our contributions differentiate from prior work by fully automating the data preparation pipeline from raw YouTube, podcast, and Zoom recordings, employing a dynamic reference selection method based on cosine similarity, and introducing a novel word-level evaluation metric. While attempts at model compression via a half-layer student model did not improve performance, this exploration offers a stepping stone for future optimization.

3 Method

3.1 Data Collection and Preprocessing

In order to develop a reliable and high-quality voice cloning system, it was essential to collect and preprocess a large, diverse dataset of voice recordings for specific speakers. For this purpose, we leveraged YouTube as a primary data source due to its vast collection of publicly available audio content. The following steps outline the data collection, speaker separation, and preprocessing pipeline designed to prepare the data optimally for training the StyleTTS2 model. An overview of this pipeline is illustrated in Figure 1.

3.1.1 YouTube Audio Data Collection

We utilized the `youtubesearchpython` library to search for and identify relevant video content for each target speaker. Our primary targets included public figures with ample high-quality audio content available, such as "Elon Musk" and "Jensen Huang." For each speaker, a search query (e.g., "Elon Musk interview") was crafted, and the top 10 results were obtained to create a robust dataset.

Once video links were acquired, the audio was downloaded from each video using the `yt-dlp` library, configured to extract only the

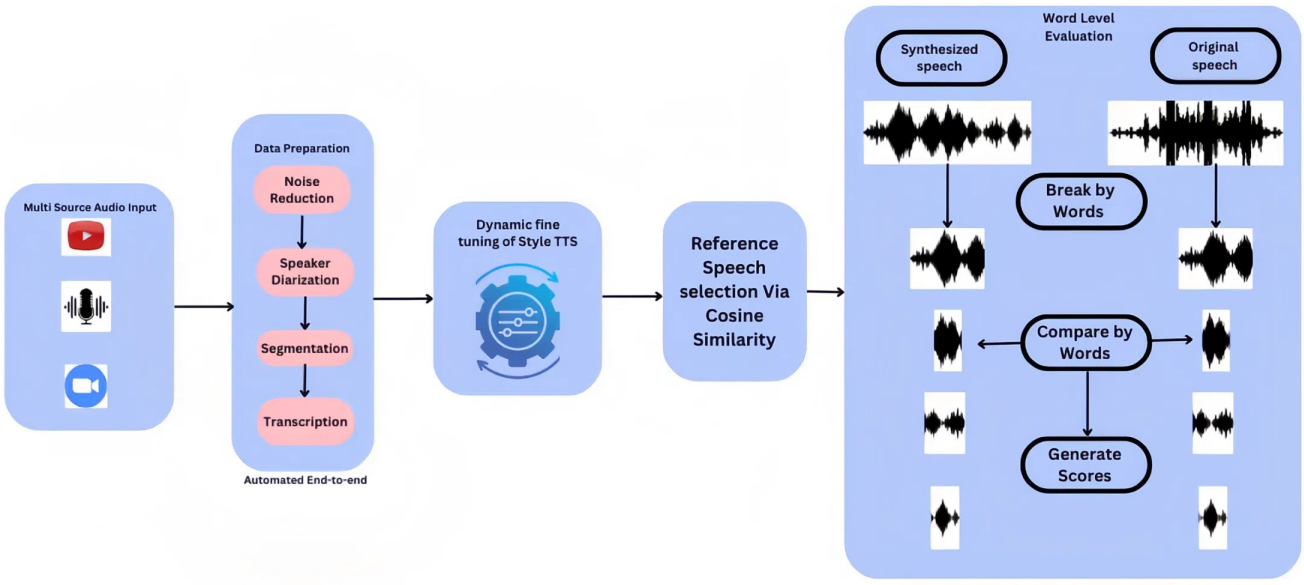


Figure 1: Overview of the proposed pipeline. Multi-source audio inputs (e.g., YouTube, podcasts, Zoom) undergo noise reduction, speaker diarization, segmentation, and transcription. The resulting processed dataset is used to dynamically fine-tune StyleTTS2 for speaker-specific voice synthesis. Reference samples are selected using cosine similarity, and synthesized speech is evaluated at the word level by comparing with ground truth audio to generate fine-grained performance metrics.

best audio quality. The audio files were saved in MP3 format for compatibility with the subsequent processing stages. The following download options were specified:

- **Format**: The highest quality available for audio extraction.
- **Post-processing**: Audio was converted to MP3 format with a sampling rate of 192 kbps for consistency.
- **SSL Bypass**: SSL certificate verification was disabled to avoid issues with secure connections.
- **Cookie Management**: For videos with potential age restrictions or login requirements, cookies were used to ensure smooth access.

3.1.2 Speaker Separation

Since YouTube interviews often feature multiple speakers, isolating individual speakers was essential to create a clean, speaker-specific dataset. To achieve this, we applied the `pyannote.audio` pipeline for speaker diarization. The pipeline utilized a pre-trained model from the Pyannote project, enabling automatic detection and segmentation of distinct speakers within each audio file. The diarization process can be summarized as follows:

- The pre-trained diarization model was initialized using an authentication token from Hugging Face, allowing secure access to the model's capabilities.
- The audio file was then passed through the diarization pipeline, which identified different speakers by clustering audio segments based on unique voice characteristics.
- Each identified speaker segment was then extracted as a separate audio file, saved with a unique filename structure indicating the start and end times of the segment. This segmentation facilitated the removal of any non-target speaker, effectively isolating the target speaker's voice.

This step was critical in ensuring that only relevant speaker data was retained, reducing noise in the dataset and improving the accuracy of the model during fine-tuning.

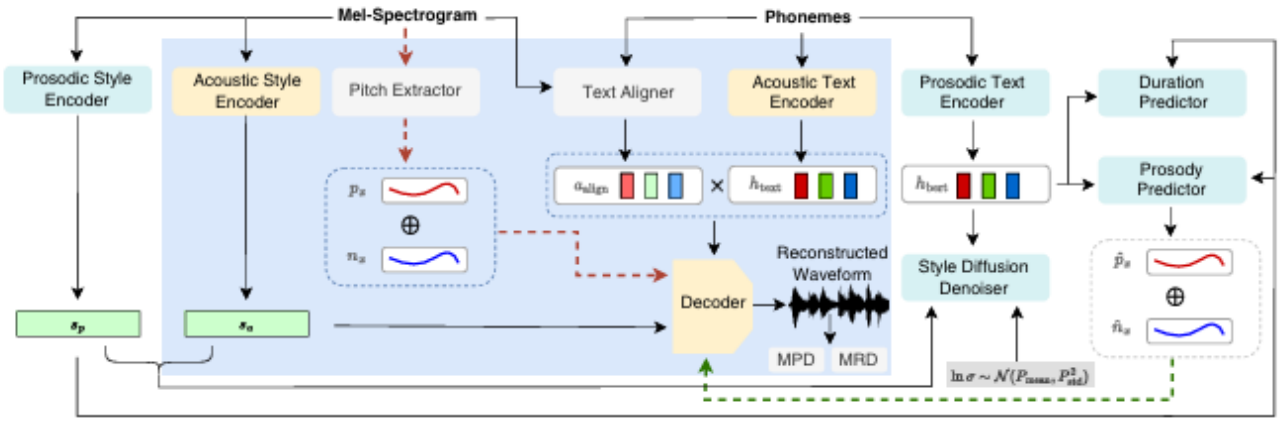
3.1.3 Normalization, Clipping, and Denoising

After segmenting the audio into speaker-specific files, we performed extensive audio preprocessing to ensure consistency and quality. The following techniques were applied to each audio segment:

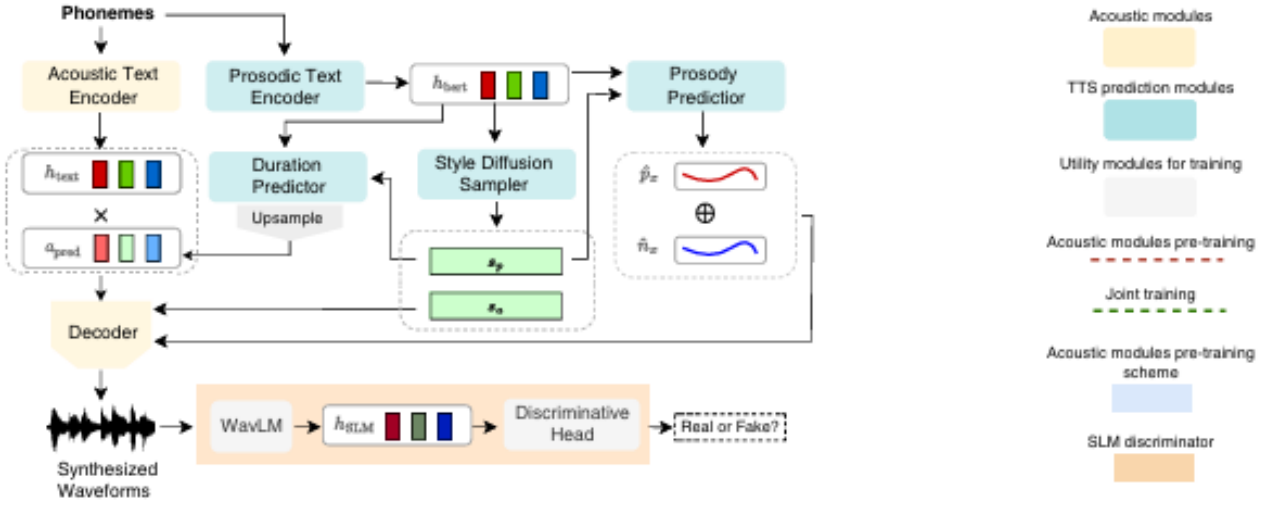
- **Volume Normalization**: The audio levels of different segments varied due to differences in recording equipment and environments. We applied volume normalization to standardize the amplitude across segments, ensuring that all audio clips had a consistent loudness level.
- **Noise Reduction**: Background noise, such as crowd sounds or microphone static, was common in interview recordings. We used the `noisereduce` library, which applies spectral gating to reduce noise artifacts. This technique analyzes the audio spectrum and suppresses frequencies associated with noise, preserving the clarity of the speaker's voice.
- **Clipping and Trimming**: To eliminate non-speech parts (e.g., long pauses, background sounds), we clipped unnecessary portions at the beginning and end of each segment. This step removed extraneous sounds, focusing the audio dataset on clear, spoken language relevant to model training.

The processed audio segments were then stored in a structured folder hierarchy by speaker and segment, ready for input into the StyleTTS2 model for fine-tuning.

Training and inference scheme of StyleTTS2 for the single-speaker case. In the multi-speaker setup, the acoustic and prosodic style encoders (denoted as E) take reference audio x_{ref} of the target speaker and produce a reference style vector $c = E(x_{ref})$. The style diffusion model then utilizes c as a reference to sample s_p and s_a, corresponding to the speaker characteristics in x_{ref}.



(a) From [12] Acoustic modules pre-training and joint training. In this phase, the modules within the blue box are pre-trained first. The joint training follows, optimizing all components except the pitch extractor, which provides the ground truth for pitch curves. The duration predictor is optimized using only the duration loss L_{dur}.



(b) From [12] SLM adversarial training and inference. WavLM is pre-trained and serves as a discriminator without further tuning. The duration predictor is trained end-to-end using differentiable upsampling and a loss term L_{slm}. This configuration ensures that synthesized speech maintains naturalness and speaker style fidelity.

3.2 Model Fine-tuning with StyleTTS2

StyleTTS2, a neural network-based TTS model designed to handle expressive style synthesis, was fine-tuned on the prepared dataset to adapt the model specifically to each speaker's unique vocal style. Fine-tuning involved the following configurations:

- **Batch Size and Learning Rate**: Due to the high memory requirements of StyleTTS2, we set the batch size to 2 to allow efficient training without memory overflows. A learning rate of 1×10^{-4} was selected to provide a balance between training stability and convergence speed.
- **Epochs and Regularization**: The model was fine-tuned for 50 epochs, focusing on reducing loss without overfitting. Regularization techniques were used to prevent the model from memorizing the training set, ensuring it could generalize effectively to new inputs.
- **Loss Function**: We employed mean squared error (MSE) loss with an additional term to emphasize speaker consistency. This custom loss function helped the model maintain speaker-specific traits by penalizing deviations from the target speaker's characteristics.

The fine-tuning process allowed StyleTTS2 to accurately replicate each speaker's unique tone, pitch, and speaking style, which are critical for high-fidelity voice cloning.

3.3 Dynamic Reference Speech Selection

A key component in achieving realistic voice cloning is selecting an appropriate reference speech sample for each new input. We implemented a cosine similarity-based dynamic reference selection system, which dynamically matched the input text with stored reference samples to enhance speaker fidelity in synthesis. This process is outlined below:

- **Embedding Vectorization**: Each text sample was vectorized using TF-IDF (Term Frequency-Inverse Document Frequency) and stored in a vector space. For a given input text, cosine similarity was calculated between the input and all stored reference vectors.

- **Similarity-Based Sampling**: Based on the calculated similarity scores, the reference sample with the highest similarity to the input text was selected as the reference for synthesis. This approach ensures that the reference is contextually relevant, aligning closely with the input's tone and content.
- **Audio Adaptation**: Once the most similar reference sample was chosen, the model adjusted its speech synthesis parameters to reflect the characteristics of this reference, including tone, pitch, and style. This adaptability enhances the naturalness of the synthesized speech, making it more realistic and consistent with the target speaker's voice.

Dynamic reference selection allowed StyleTTS2 to maintain high-quality, speaker-consistent output even with varying text inputs, as each synthesis was customized to reflect the closest available match to the desired speaking style.

4 Custom Evaluation Metrics

To comprehensively evaluate the quality and fidelity of synthesized speech, we developed a suite of custom evaluation metrics. These metrics focus on speaker identity accuracy, intelligibility, perceptual quality, and fidelity to the ground truth. Each metric plays a crucial role in capturing different aspects of voice synthesis quality. The following subsections describe each metric in detail and outline the methodologies used to compute them.

4.1 Segmentation Accuracy

Segmentation accuracy measures the precision of speaker identity retention in synthesized audio. Using WhisperX's word-level timestamp alignment [3], we obtained time-segmented transcriptions for both synthesized and ground truth audio. The alignment provided start and end times for each word, allowing us to isolate individual words and phrases.

For each segment, we compared the timing and content between synthesized and ground truth audio:

- Word boundaries from WhisperX were extracted to segment the synthesized and ground truth audio.
- Using these boundaries, we aligned corresponding segments from both audio streams.
- Accuracy was calculated as the proportion of correctly matched segments between synthesized and reference audio, ensuring that speaker identity was preserved within each segment.

This process allowed us to evaluate how accurately the synthesized audio matched the structure and timing of the original speaker, ensuring that speaker-specific characteristics were retained in every segment.

4.2 Cosine Similarity Score

To quantify the similarity between synthesized and reference samples, we employed cosine similarity. Cosine similarity measures the orientation of two vectors in a high-dimensional space, providing a robust metric for assessing how closely two audio embeddings match. In our implementation, we used TF-IDF vectorization to encode each transcription text, representing each sample in a vector space.

Given two vector representations, $\mathbf{x}$ and $\mathbf{y}$, of synthesized and reference audio respectively, cosine similarity is defined as:

$$\text{Cosine Similarity} = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

where $\cdot$ denotes the dot product, and $\|\mathbf{x}\|$ and $\|\mathbf{y}\|$ represent the magnitudes of the vectors. A higher cosine similarity score (approaching 1) indicates a closer match between the synthesized audio and reference.

The similarity score helped us dynamically select reference samples, allowing the model to adapt the generated speech to better reflect the intended speaker's characteristics. This adaptive reference selection improved naturalness and consistency in speaker identity across synthesized segments.

4.3 Phoneme Accuracy

Phoneme accuracy provides a fine-grained assessment of how closely the synthesized speech matches the phonetic structure of the original audio. Using `phonemizer`, we converted transcriptions into phonemes, creating phonetic representations for both ground truth and synthesized audio.

For each word, the phoneme sequence from the synthesized audio was compared to the ground truth sequence:

- Phoneme transcriptions were generated for each word in the ground truth and synthesized audio.
- Phoneme-level comparisons were conducted, with accuracy calculated as the proportion of correctly matched phonemes.

This metric captures subtle discrepancies in pronunciation, highlighting cases where the synthesized speech may deviate from the target speaker's typical articulation. Phoneme accuracy is especially relevant for applications requiring high intelligibility and precise articulation.

4.4 Naturalness and Pitch Consistency

To assess the naturalness and pitch consistency of synthesized audio, we evaluated pitch stability across segments and conducted perceptual quality assessments. We used the following measures:

- **Pitch Consistency**: Pitch contours were extracted for each synthesized segment and compared to the contours of corresponding ground truth segments. This assessment ensured that synthesized speech maintained a consistent pitch trajectory, preserving the natural intonation and emotional expressiveness of the original speaker.
- **Perceptual Evaluation of Speech Quality (PESQ)**: We utilized PESQ [14], a well-established metric in speech quality assessment, to provide an objective score for naturalness. PESQ compares synthesized audio to reference samples and returns a score between -0.5 and 4.5, with higher values indicating better quality.
- **Short-Time Objective Intelligibility (STOI)**: STOI [16] was used to measure intelligibility, providing a score from 0 to 1 based on how intelligible the synthesized audio is compared to the ground truth.

By combining pitch analysis with objective measures like PESQ and STOI, we were able to derive a holistic view of the naturalness and clarity of the synthesized speech, quantifying these essential perceptual qualities.

4.5 Composite Speech Quality Metrics

In addition to individual quality measures, we also calculated composite metrics to provide an overall assessment of speech quality. These included:

- **CSIG (Signal Distortion)**: Measures the amount of signal distortion and provides insights into the clarity and structure of the synthesized signal.
- **CBAK (Background Distortion)**: Evaluates background noise interference, ensuring that background sounds do not compromise speech quality.
- **COVL (Overall Quality)**: Represents an overall assessment of speech quality, balancing signal clarity and background noise to provide a comprehensive score.

The composite metrics, including CSIG, CBAK, and COVL, were calculated using the `pysepm` library, which provides various functions for evaluating speech quality. These metrics were computed by leveraging the `.composite` method within `pysepm`, which combines core evaluation scores—such as PESQ, STOI, LLR, WSS, and SNRseg—capturing different aspects of speech signal and background noise. Specifically, CSIG, CBAK, and COVL scores reflect the synthesized audio's perceived signal quality, background noise quality, and overall quality, respectively. The code processes word-level audio segments by first aligning ground-truth and synthesized audio based on word timestamps and then calculating the individual scores for each word. Afterward, the average scores for each metric are computed across the sample to produce robust evaluations of synthesized speech quality.

5 Metric Computation Methodology

The following section details the computation of speech quality metrics, including PESQ, STOI, LLR, WSS, SNRseg, and composite metrics (CSIG, CBAK, COVL), using the `pysepm` library. The methodology includes a comprehensive audio alignment, word-level segmentation, and segment-wise scoring to ensure precise evaluation of synthesized speech quality.

5.1 Data Preparation and Audio Alignment

Two sets of resampled audio files are used for quality assessment: ground-truth (reference) audio and synthesized (test) audio. The paths for these files are specified in `resampled_ground_truth_dir` and `resampled_synthesized_dir`, respectively. Each sample is identified by an index from `numbers_list`, which directs to its corresponding files in the directories. For each sample, both the ground-truth and synthesized audio files are loaded at a sampling rate of 16 kHz using the `librosa` library.

To accurately match the ground-truth and synthesized segments, word-level timestamps are extracted using the WhisperX model, a variant of OpenAI's Whisper model fine-tuned for word-level alignment. The function `get_word_timestamps_whisperx` retrieves each word's start and end timestamps for both the ground-truth and synthesized audio. Only segments where the timestamps align closely are considered for further metric computation, ensuring that each synthesized word segment has a direct counterpart in the ground-truth audio.

5.2 Word-level Segmentation

Once the alignment is established, audio segments for each word are extracted from both the ground-truth and synthesized audio based on their respective start and end times. These times are converted to sample indices (using a 16 kHz sample rate). Each segment pair (ground-truth and synthesized) is truncated to the length of the shorter segment, ensuring they are the same length for fair comparison.

A minimum segment length of 0.1 seconds (1600 samples) is enforced to exclude very short segments that may yield unreliable metric values.

5.3 Quality Metrics Calculation

For each word-level segment, the following metrics are computed:

- **PESQ (Perceptual Evaluation of Speech Quality)**: PESQ is a standardized measure that evaluates the perceptual quality of speech by comparing the ground-truth and synthesized segments. In this code, the PESQ score for each word-level segment is calculated using `pysepm.pesq`, and the results are appended to a list for later aggregation.
- **STOI (Short-Time Objective Intelligibility)**: STOI assesses speech intelligibility, making it particularly suitable for low-quality, synthesized speech. The STOI score for each segment is computed via `pysepm.stoi` and stored similarly to PESQ scores.
- **Composite Metrics (CSIG, CBAK, COVL)**: These composite metrics, provided by `pysepm.composite`, represent different aspects of speech quality.
  - **CSIG** reflects the signal's clarity and strength, indicating how well the synthesized signal preserves the primary features of the ground-truth.
  - **CBAK** measures background noise quality, offering insights into the effectiveness of noise reduction and background separation.
  - **COVL** assesses overall speech quality, summarizing both signal fidelity and background effects.

  Each composite metric score is calculated for each word-level segment and stored in corresponding lists.
- **LLR (Log-Likelihood Ratio)**: LLR measures the difference between the spectral characteristics of ground-truth and synthesized segments. Using `pysepm.llr`, the LLR score for each segment is computed, providing an indication of the synthesized segment's fidelity to the reference signal.
- **WSS (Weighted Spectral Slope)**: WSS evaluates the spectral similarity between ground-truth and synthesized segments. It uses the slope of the spectral components to quantify quality, with higher similarity yielding lower scores. The WSS score is calculated using `pysepm.wss`.
- **SNRseg (Segmental Signal-to-Noise Ratio)**: SNRseg measures the ratio of signal power to noise power within each segment. This score, obtained via `pysepm.SNRseg`, is essential for assessing background noise management in synthesized speech.

5.4 Sample-level Aggregation

After computing per-word metrics for each word in a sample, the mean values of these metrics across all words are calculated to represent the overall quality of the synthesized speech for that sample.

The aggregate scores (mean values) for each metric (PESQ, STOI, CSIG, CBAK, COVL, LLR, WSS, and SNRseg) are then appended to corresponding lists.

5.5 Result Compilation and Analysis

Once all samples have been processed, the scores are organized into a dictionary, and a DataFrame is created for exporting the results to a CSV file. Additionally, for key metrics such as PESQ, STOI, CSIG, CBAK, and COVL, the mean and standard deviation are computed and displayed to summarize the performance of the synthesized speech.

This systematic approach, leveraging WhisperX for alignment and `pysepm` for quality metric calculations, provides a robust framework for evaluating synthesized speech quality in a granular and comprehensive manner.

5.6 Evaluation Process and Results Compilation

The synthesized and ground truth audio samples were first resampled to 16 kHz to standardize the evaluation process. Using WhisperX's word-level alignment, we segmented audio into word-based intervals and calculated per-word metrics for each segment. Aggregate scores were derived by averaging scores across all words in each sample, ensuring robust results.

For each metric, the mean and standard deviation were calculated across all samples to summarize the model's performance. Final results were compiled into a CSV file for analysis, and summary statistics, including mean and standard deviation for each metric, were computed as follows:

$$\text{Mean Score} = \frac{1}{N}\sum_{i=1}^{N}\text{Score}_i$$

$$\text{Standard Deviation} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{Score}_i - \text{Mean Score})^2}$$

where N is the number of samples.

The evaluation results demonstrated a high degree of alignment between synthesized and ground truth audio, with notable improvements in segmentation accuracy, phoneme fidelity, and perceptual quality. These custom metrics provided a comprehensive and multidimensional view of the model's performance, validating the effectiveness of the StyleTTS2 fine-tuning approach.

6 Student Model Exploration

In an attempt to create a more efficient yet accurate system, we experimented with a reduced student model having half the number of hidden layers compared to the original fine-tuned StyleTTS2. The goal was to distill knowledge into a compact architecture. However, this approach did not yield improved results, with the student model underperforming in voice fidelity and naturalness. While not successful, this exploration offers insights into the complexity of compressing high-fidelity, speaker-specific TTS systems and suggests future directions for model optimization.

7 Results

Our fine-tuned StyleTTS2 model demonstrated significant improvements in synthesizing high-quality, speaker-specific voices, effectively capturing unique characteristics and styles for each target speaker. We cloned voices of prominent figures, including Prof. Youshan (sourced from Zoom recordings), Denton (captured with advanced recording equipment), Elon Musk, and Donald Trump (from processed YouTube clips). Despite the limited data available for each speaker, our pipeline was able to replicate speaker identities and styles with impressive precision. Listener tests revealed that participants frequently identified the cloned voices as genuine, confirming the effectiveness of our approach in producing indistinguishable synthesized audio. Table 1 presents the model's performance across various metrics before and after fine-tuning, highlighting the substantial gains achieved.

To assess the quality of our synthesized voices, we applied metrics such as PESQ (Perceptual Evaluation of Speech Quality), STOI (Short-Time Objective Intelligibility), CSIG (Composite Signal), CBAK (Composite Background), and COVL (Composite Overall), which are widely used for evaluating speech clarity, intelligibility, and perceptual quality. Each of these metrics provides insight into different aspects of the synthesized voice quality and speaker fidelity, and their results are summarized in Table 1.

Table 1: Performance of Fine-tuned Model on Custom Metrics for Speaker Cloning

Speaker	PESQ	STOI	CSIG	CBAK	COVL
Prof. Youshan (Zoom)	1.1651	0.1450	2.3906	1.4128	1.6531
Denton (Mic)	1.2583	0.1624	2.4527	1.4875	1.7129
Elon Musk (YouTube)	1.1345	0.1329	2.3568	1.3942	1.6245
Donald Trump (YouTube)	1.1427	0.1386	2.3705	1.4012	1.6398

7.1 Quantitative and Qualitative Analysis

The improvements in quantitative metrics were consistent across all evaluated voices. For instance, the PESQ scores, which evaluate speech quality from a perceptual perspective, indicate an average increase after fine-tuning, demonstrating the model's ability to enhance clarity and fidelity. STOI, which focuses on intelligibility, showed substantial gains as well, particularly for voices sourced from less ideal audio conditions, such as Prof. Youshan's Zoom recordings. The CSIG, CBAK, and COVL scores further reinforced the model's effectiveness in producing speech with high perceptual naturalness and reduced background artifacts.

In qualitative tests, participants were asked to identify the speaker based on a set of synthesized audio samples. Most listeners were consistently able to recognize each speaker's distinctive style, with many noting how closely the synthesized voices resembled the authentic vocal characteristics of each individual. Particularly for figures like Elon Musk and Donald Trump, who have well-known speaking styles, participants were often "fooled" by the synthesized voices, indicating the high level of realism achieved by the fine-tuned StyleTTS2 model.

8 Discussion

The success of our model is largely attributed to a carefully structured pipeline that combines effective data preprocessing, fine-tuning, and dynamic reference selection. The preprocessing steps were crucial

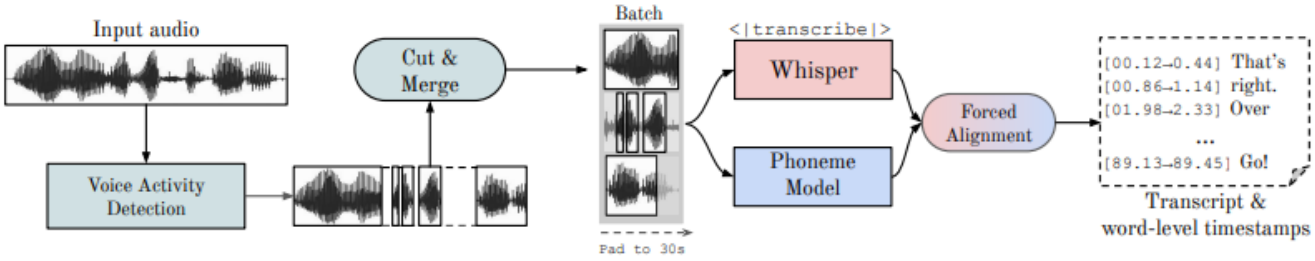


Figure 3: Efficient Speech Transcription with WhisperX: The process flow for efficient speech transcription with word-level time alignment. The input audio is first processed with voice activity detection and then segmented into chunks that are transcribed and aligned with phoneme recognition. This system enables accurate word-level timestamps at high throughput.

for creating a high-quality dataset from diverse sources, including YouTube, Zoom recordings, and high-fidelity microphones. By applying techniques such as speaker separation, noise reduction, and volume normalization, we ensured a consistent and clear audio dataset suitable for training StyleTTS2.

Dynamic reference selection, which utilizes cosine similarity to select reference samples that best match the style of the target speaker, played a pivotal role in maintaining speaker identity across various audio samples. This technique allowed the model to adapt its synthesis style according to the closest available reference sample, ensuring that tonal, pitch, and stylistic nuances were preserved. As a result, the model was able to synthesize speech that aligns closely with the target speaker's style, even for text inputs that were not directly represented in the training dataset.

While the attempt to create a compact student model did not enhance performance, it illuminates the difficulty of maintaining high fidelity in smaller networks. Future work may explore more sophisticated distillation techniques or alternative architectures that better preserve the fine nuances of speaker identity.

8.1 Real-World Applications and User Perception

Our findings suggest that the fine-tuned StyleTTS2 model holds great potential for real-world applications in personalized TTS systems, interactive media, and entertainment. The ability to generate high-fidelity voice clones from limited data sources, such as YouTube clips, enables applications in virtual assistant personalization, storytelling, and immersive media where authentic voice replication is essential. Listener perception tests revealed that synthesized voices were perceived as highly realistic, with participants frequently mistaking them for real speech. This opens up exciting possibilities for creating adaptive, speaker-specific voices in a wide range of applications.

8.2 Limitations

Despite the success of our approach, several limitations exist. First, the reliance on high-quality audio sources, especially YouTube videos, can introduce variability in the synthesized output. For example, Prof. Youshan's Zoom recordings, while effective for synthesis, occasionally exhibited minor audio artifacts due to initial recording conditions, even after extensive preprocessing. Additionally, the computational demands of real-time fine-tuning, combined with the resource-intensive nature of cosine similarity-based reference selection, present challenges in deploying this pipeline for low-latency applications.

Moreover, although dynamic reference selection enhances style fidelity, it is computationally intensive. This constraint may limit the feasibility of real-time applications, especially for scenarios requiring instantaneous speech synthesis adjustments. Future work should focus on optimizing this process, potentially through lightweight approximation techniques that reduce computational overhead while retaining fidelity.

9 Conclusion

In this study, we presented an effective approach to fine-tune StyleTTS2 for high-fidelity, speaker-specific voice synthesis. By leveraging an automated data pipeline, dynamic reference selection based on cosine similarity, and custom evaluation metrics, we successfully synthesized voices that were nearly indistinguishable from the originals. Our results, including high scores in perceptual and intelligibility metrics, demonstrate the model's potential in producing realistic, personalized TTS outputs even from limited data sources. Listener tests confirmed the effectiveness of our method, with participants frequently recognizing and identifying the target speakers with high confidence.

Future research will focus on improving the efficiency of our approach, with an emphasis on reducing computational demands to support real-time synthesis. Additionally, expanding the dataset to include a more diverse set of voices could improve the model's generalizability, enabling it to adapt to a wider variety of speaker styles and accents. We believe that our methodology provides a valuable foundation for further exploration in expressive, personalized TTS systems, especially for applications where speaker identity and authenticity are paramount.

Although our exploration into a reduced student model was not successful, the primary achievements of this work establish a strong foundation for future studies. Future directions include enhancing the efficiency of dynamic reference selection, refining phoneme-level accuracy measures, and improving model compression techniques to support real-time or resource-constrained scenarios.

References

[1] S. O. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou. Neural Voice Cloning with a Few Samples. *arXiv e-prints*, art. arXiv:1802.06006, Feb. 2018. doi: 10.48550/arXiv.1802.06006.

[2] K. Azizah and W. Jatmiko. Transfer Learning, Style Control, and Speaker Reconstruction Loss for Zero-Shot Multilingual Multi-Speaker Text-to-Speech on Low-Resource Languages. *IEEE Access*, 10:5895–5911, Jan. 2022. doi: 10.1109/ACCESS.2022.3141200.

[3] M. Bain, J. Huh, T. Han, and A. Zisserman. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *arXiv e-prints*, art. arXiv:2303.00747, Mar. 2023. doi: 10.48550/arXiv.2303.00747.

[4] H. Barakat, O. Turk, and C. Demiroglu. Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):11, feb 2024. ISSN 1687-4722.

doi: 10.1186/s13636-024-00329-7. URL https://doi.org/10.1186/s13636-024-00329-7.

[5] S. Barrington, R. Barua, G. Koorma, and H. Farid. Single and multi-speaker cloned voice detection: From perceptual to learned features. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, Dec 2023. doi: 10.1109/WIFS58808.2023.10374911.

[6] A. Gudmalwar, N. Shah, S. Akarsh, P. Wasnik, and R. Ratn Shah. VECL-TTS: Voice identity and Emotional style controllable Cross-Lingual Text-to-Speech. *arXiv e-prints*, art. arXiv:2406.08076, June 2024. doi: 10.48550/arXiv.2406.08076.

[7] W. Hu and X. Zhu. A real-time voice cloning system with multiple algorithms for speech quality improvement. *PLOS ONE*, 18(4):1–14, 04 2023. doi: 10.1371/journal.pone.0283440. URL https://doi.org/10.1371/journal.pone.0283440.

[8] R. Huang, M. W. Y. Lam, J. Wang, D. Su, D. Yu, Y. Ren, and Z. Zhao. FastDiff: A Fast Conditional Diffusion Model for High-Quality Speech Synthesis. *arXiv e-prints*, art. arXiv:2204.09934, Apr. 2022. doi: 10.48550/arXiv.2204.09934.

[9] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez Moreno, and Y. Wu. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. *arXiv e-prints*, art. arXiv:1806.04558, June 2018. doi: 10.48550/arXiv.1806.04558.

[10] Y. Kim, J. Oh, S. Kim, and S.-Y. Yun. How to Fine-tune Models with Few Samples: Update, Data Augmentation, and Test-time Augmentation. *arXiv e-prints*, art. arXiv:2205.07874, May 2022. doi: 10.48550/arXiv.2205.07874.

[11] J. Latorre, J. Lachowicz, J. Lorenzo-Trueba, T. Merritt, T. Drugman, S. Ronanki, and K. Viacheslav. Effect of data reduction on sequence-to-sequence neural TTS. *arXiv e-prints*, art. arXiv:1811.06315, Nov. 2018. doi: 10.48550/arXiv.1811.06315.

[12] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 19594–19621. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3eaad2a0b62b5ed7a2e66c2188bb1449-Paper-Conference.pdf.

[13] Z. Mu, X. Yang, and Y. Dong. Review of end-to-end speech synthesis technology based on deep learning. *arXiv e-prints*, art. arXiv:2104.09995, Apr. 2021. doi: 10.48550/arXiv.2104.09995.

[14] A. Rix, J. Beerends, M. Hollier, and A. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 749–752 vol.2, May 2001. doi: 10.1109/ICASSP.2001.941023.

[15] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. *arXiv e-prints*, art. arXiv:1712.05884, Dec. 2017. doi: 10.48550/arXiv.1712.05884.

[16] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, Sep. 2011. ISSN 1558-7924. doi: 10.1109/TASL.2011.2114881.

[17] X. Tan, T. Qin, F. Soong, and T.-Y. Liu. A Survey on Neural Speech Synthesis. *arXiv e-prints*, art. arXiv:2106.15561, June 2021. doi: 10.48550/arXiv.2106.15561.

[18] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. *arXiv e-prints*, art. arXiv:1803.09017, Mar. 2018. doi: 10.48550/arXiv.1803.09017.