Volume 1, 2025, Pages 206-218 *Elaris Computing Nexus*
| Regular Article | Open Access

Personalized Text to Speech Synthesis through Few Shot Speaker Adaptation with Contrastive Learning

Natarajan K
Department of Electrical and Electronics Engineering, Trinity College of Engineering and Technology, Peddapalli, Telangana, India.
drknatarajan17@gmail.com

Article Info
Elaris Computing Nexus
https://elarispublications.com/journals/ecn/ecn_home.html

https://doi.org/10.65148/ECN/2025019

Received 23 July 2025
Revised form 12 September 2025
Accepted 26 October 2025
Available online 02 November 2025
Published by Ansis Publications.

Corresponding author(s):
Natarajan K, Department of Electrical and Electronics Engineering, Trinity College of Engineering and Technology, Peddapalli, Telangana, India.
Email: drknatarajan17@gmail.com

Abstract – Personalized text-to-speech (TTS) synthesis has the goal of producing natural and expressive speech that emulates the voice of a target speaker with a minimum of data. The models of traditional neural TTS, including Tacotron 2 and Fast Speech 2, need to be trained in large amounts of speaker-specific data and can thus not easily be personalized quickly. We suggest CL-FS-TTS (Contrastive Learning based Few-Shot Text-to-Speech) to solve this problem, a new framework that uses contrastive speaker representation learning to adapt the speaker using only 1030 seconds of reference audio. The CL-FS-TTS architecture has two encoders: a content encoder that identifies linguistic features of text and a speaker encoder trained with the help of supervised contrastive learning to maximize speaker dissimilarity. In adaptation, the model matches speaker embeddings with generated mel-spectrograms with a contrastive consistency loss, enhancing voice and prosodic consistency. We compare CL-FS-TTS with Tacotron 2, Fast Speech 2, AdaSpeech, YourTTS, and Meta-TTS in terms of Mean Opinion Score (MOS), Speaker Similarity Score (SSS), Mel Cepstral Distortion (MCD) and Word Error Rate (WER). The experimental outcomes indicate that CL-FS-TTS has a higher naturalness and similarity of the speaker besides 40% less adaptation time in comparison with baselines. The suggested model lays the foundation of an efficient and strong model of high-quality personalized TTS synthesis in the situation of data scarcity.

Keywords – Few-Shot Speaker Adaptation, Contrastive Learning, Personalized Text-To-Speech (TTS), Speaker Embedding, Neural Speech Synthesis.

I. INTRODUCTION

Background and Motivation

The communication style of human beings has also changed tremendously due to emergence of intelligent systems that are able to comprehend and make speech. One of the most revolutionary of these technologies is the text to speech (TTS) synthesis [1] that is used to transform written text into sound. Contemporary neural TTS systems have achieved unbelievable degrees of naturalness, which produces voices that are sometimes hardly distinguishable with human speech. Such advancements have been facilitated by deep learning systems including Tacotron 2, FastSpeech, and VITS, which make use of attention and end-to-end training systems to generate high quality and talk-like speech [2]. Nevertheless, personalization, or creating speech that reflects the individual vocal qualities and speaking manner and tone of a particular human being is one of the issues that need to be addressed urgently but has not been done so yet. Personalized TTS is gaining enormous opportunities in many aspects including voice restoration in individuals with speech disorder, synthesize voices of computerized avatars, ease of access in accessibility equipment, and interfaces with users in virtual assistive devices. However, even the majority of state-of-the-art models require significant amounts of speaker-specific information, which can be costly to gather or practically impossible to obtain.

This inspires the investigation of few-shot speaker adaptation - a method in which a system is trained to produce a novel voice with only seconds or minutes of reference audio. The most important issue with this is that the little data can cause unstable training, loss of the identity of the speaker and a poor generalization between unknown speakers or language

situations. To overcome such problems, current studies have emphasized on self-supervised and contrastive learning models to construct richer speaker representations that are well generalized on limited data [3].

Problem Statement

Although the few-shot adaptation has improved, there are three limitations that are still critical in the current methods. To begin with, when there is little data, speaker embeddings generated by conventional encoders cannot necessarily be able to represent the fine details of timbre, accent, and emotion. This produces artificial voices that are generic, but not customisable. Second, the majority of adaptation methods involve direct fine-tuning of large pre-trained models, including Tacotron 2 or FastSpeech 2, which can be computationally expensive as well as overfitting on small training sets. Third, there are still no strong mechanisms that can be used to bring consistency between the speech produced and the latent representation of the target speaker during adaptation. Lacking these restrictions, the capability of the model to preserve speaker identity is impaired when the model tries to generalize in unknown phonetic or prosodic situations. All these issues point to the necessity of a more powerful, data-conscious framework that will be able to effectively learn discriminative and stable speaker representations using limited data and produce the synthesized speech that will be natural and clear [4, 5].

Proposed Solution

To overcome them, this paper proposes CL-FS-TTS (Contrastive Learning-based Few-Shot Text-to-Speech), a new framework that aims at a high-quality personalized speech synthesis with limited speaker data. The key point of CL-FS-TTS is that the concept of contrastive representation learning is introduced into the adaptation of the TTS process. The contrastive learning makes the model to learn both discriminative and invariant speaker embeddings, i.e. being able to differentiate between speakers and being consistent to variations in linguistic content and recording conditions.

Our system is a dual-encoder system comprising of:

- A content encoder that is able to extract linguistic and prosodic information out of text and mel-spectrograms, and
- A speaker encoder that is pre-trained on a supervised contrastive task that maximizes the similarity of the same speaker and minimizes that of speakers.

In adaptation, a contrastive consistency loss is used to match the speaker embedding of the reference audio to the embedding of the speaker produced by the decoder to ensure that again the synthesized speech has retained the speaker identity despite limited training samples. The lightweight fine-tuning strategy is also utilized in the model to reduce the number of computations and therefore the model can be deployed on resource scarcity devices and also used in real-time personalization.

Novelty and Contributions

The major originality of the research is the combination of contrastive learning and few-shot speaker adaptation in an end-to-end TTS model. Although other models like AdaSpeech, YourTTS and Meta-TTS have achieved progress in terms of adaptation efficiency, lack of mechanism to explicitly impose consistency between generated and reference voice is common between these models.

The given paper can make a contribution to the current knowledge that can be summarized as follows:

- Contrastive Learning of Speaker Representation: a supervised contrastive training method is adopted to pre-train speaker encoder so as to achieve a better discrimination of speaker identities and resistance to data deficiencies.
- Contrastive Consistency Loss to Adaptation: A new loss term is used to match latent speaker embedding during few-shot adaptation and maintain speaker-specific features of the synthesized speech.
- Lightweight Fine-Tuning Strategy: The adaptation process only fine-tunes a few layers of the TTS decoder and saves training time and computation costs by more than 40x over a full model fine-tuning.
- Complete Evaluation Framework: The framework is tested on both objective and subjective measures such as Mel Cepstral Distortion (MCD) and Word Error Rate (WER) and human assessment of the system including Mean Opinion Score (MOS) and Speaker Similarity Score (SSS).
- Comparison of Performance with State-of-the-Art Models: CL-FS-TTS is strictly compared to Tacotron 2, FastSpeech 2, AdaSpeech, YourTTS and Meta-TTS, where it is shown to be outperforming in all three areas of naturalness, speaker similarity, and adaptation efficiency.

Significance and Potential Applications

The efficient implementation of few-shot personalized TTS model, such as CL-FS-TTS, has a major real-life implication. In the case of medical patients losing voices, given a small collection of pre-recorded samples, it may be sufficient to reconstruct the natural voice of communication of the affected individuals. Likewise, game developers and content creators might immediately create the voices of characters with few recordings, and retain expressiveness and personality. Personalized voices can help make reading more interactive and inclusive in the field of education and accessibility in the case of aids, e-learning platforms and virtual tutors. In addition, personalization can be applied in the context of human-computer interaction to make digital assistants feel more familiar and emotionally resonant to users, which will increase user trust and satisfaction. The focus on contrastive learning is also correlated with the current tendencies of self-supervised

and representation learning, indicating that the proposed approach may be the basis of other multimodal personalization problems like emotion synthesis with speaker conditioning or speech-controlled animation.

Organization of the Paper

The remainder of this paper is structured as follows:

- *Section 2*: Related Work – reviews existing literature on neural TTS systems, few-shot speaker adaptation methods, and contrastive learning approaches for speech representation.
- *Section 3*: Proposed Methodology - contains the description of the architecture of CL-FS-TTS, the design of content and speaker encoders, the contrastive consistency loss, and the adaptation process.
- *Section 4*: Experimental Setup – describes the datasets, evaluation metrics, and implementation details used to assess model performance.
- *Section 5*: Results and Discussion – presents the quantitative and qualitative results, comparing CL-FS-TTS against five state-of-the-art models, followed by an in-depth analysis.
- *Section 6*: Conclusion and Future Work – summarizes the contributions and highlights possible directions for extending this research, such as cross-lingual personalization and emotional voice synthesis.

II. RELATED WORKS

History The history of the development of personalized text-to-speech (TTS) systems goes back decades of research between the conventional signal processing to the recent deep learning-based generative models. This part will examine the most pertinent literature in the neural TTS systems, speaker adaptation, few-shot learning, and contrastive learning approaches which have guided the design of the proposed CL-FS-TTS.

Neural Text-to-Speech Systems

The older TTS models, including concatenative and parametric synthesis, had to be based on manual linguistic and acoustic cues and their speech turned out to be synthetic and unnatural. With the introduction of neural sequence-to-sequence models, the situation changed. An end-to-end architecture was proposed in Tacotron and Tacotron 2 and the input text was directly mapped to mel-spectrograms which were then converted to waveforms by neural vocoders such as WaveNet and WaveGlow. These models were both data intensive and speaker specific and attained the naturalness of humans [6].

FastSpeech and FastSpeech 2 To enhance efficiency FastSpeech and FastSpeech 2 proposed non-autoregressive architectures that were many times faster in inference time and quality. To substitute the attention mechanism, they presented duration predictors with better synthesis in a smoother and faster way. In spite of these progress, these models had difficulties in adjusting to new speakers, especially in the situation when there were scarce data.

Most recently, VITS (Variational Inference Text-to-Speech) is a variant of a variational autoencoder (VAE) using a generative adversarial network (GAN) to simultaneously predict speech duration, pitch and waveform generation. Although VITS was able to reach state of the art audio quality, it was still constrained in the capacity to learn with small few-shot samples without overfitting or speaker loss [7, 8].

Speaker Adaptation in Neural TTS

Speaker adaptation aims to transfer a pre-trained multi-speaker TTS model to a new target speaker using limited adaptation data. Several strategies have been proposed to achieve this. Speaker embedding-based approaches are among the most common. These methods, such as Global Style Tokens (GST-Tacotron) and Deep Voice 3, introduce learnable embeddings that capture speaker characteristics and style variations. However, these embeddings often lack fine-grained control, and when trained with limited data, they may not sufficiently represent individual timbre or prosody [9, 10].

To address data scarcity, meta-learning and transfer learning methods have been explored. AdaSpeech introduced layer-wise adaptive instance normalization to personalize pre-trained TTS models for new speakers with only a few minutes of data. YourTTS extended this concept to multilingual and zero-shot settings by combining speaker verification encoders with cross-lingual phoneme modeling. Similarly, Meta-TTS employed a meta-learning framework that learned how to adapt quickly to unseen speakers through model-agnostic optimization. While these models achieved impressive results in few-shot and zero-shot scenarios, they still faced limitations in maintaining speaker identity consistency across different utterances, especially when training data was extremely limited (under one minute). Moreover, adaptation often required updating a large number of model parameters, leading to high computational costs [11].

Few-Shot and Zero-Shot Learning for Speech Synthesis

The objective of speaker adaptation is to use a pre-trained multi-speaker TTS system to adapt to a different target speaker based on restricted adaptation data. A number of approaches have been advanced in order to do this. The most common ones are speaker embedding-based. Such processes as Global Style Tokens (GST-Tacotron) and Deep Voice 3 propose learnable embeddings which learn speaker peculiarities and style variations. Such embeddings, however, are frequently not fine-grained controlled, and learned with little data, cannot adequately capture either individual timbre or prosody [12, 13].

In an attempt to solve the lack of data, meta-learning and transfer learning have been considered. AdaSpeech presented layer-wise adaptive instance normalization to adapt pre trained TTS to new speakers using just a few minutes of data.

YourTTS applied this to multilingual and zero-shot scenarios, jointly training speaker verification encoders with cross lingual phoneme modeling. On the same note, Meta-TTS used a meta-learning architecture which was trained to adapt rapidly to unseen speakers by model-agnostic optimization. Although these models yielded amazing results in the few-shot and zero-shot case, they nonetheless had limitations of ensuring speaker identity consistency across utterances, particularly when the training data was very small (less than one minute). In addition, due to the need to adapt many model parameters, the computation could also be expensive [14].

Contrastive Learning for Speech Representation

The approach of few-shot and zero-shot learning has gained more and more popularity since it enables adaptation to new tasks or speakers with limited data. In TTS, few-shot learning aims at fine-tuning pre-trained models using only a few samples and zero-shot methods only use speaker embeddings obtained on reference audio without fine-tuning. Few-shot models, such as SC-GlowTTS and LightSpeech, are based on small architectures or lightweight adaptation systems to minimize the cost of adaptation [15]. The Meta-StyleSpeech used meta-learning to approximate the adaptation process that occurs during training and thus enhanced the generalization to unknown speakers. Nevertheless, all of them are usually based on the assumption that speaker encoding trained on large and diversified datasets will be generalizable, which does not always hold in the case when the voice of the target speaker is significantly different in voice characteristics compared to the training samples [16].

Zero-shot methods like YourTTS and VoiceFilter-Lite will utilize already trained speaker verification networks (e.g., GE2E or x-vector encoders) to obtain fixed representations of speakers. These embeddings are then taken as conditioning input in the synthesis. Although convenient, they are not optimized together with the TTS model, which may lead to an incongruity between the latent space of the speaker and the synthesized output space. This creates impaired similarity and un-natural prosody in adapting to invisible speakers. The weakness of both few-shot and zero-shot paradigm lies in the lack of discriminative ability of the speaker representations when data are sparse. This has inspired the desire to develop more powerful representation learning algorithms in particular, contrastive learning that can learn speaker-Discriminative embeddings with only weak supervision [17].

Summary and Research Gap

The existing TTS systems have reached high naturalness and intelligibility but are still threatened by the trade-off between the quality of the personalization and the data efficiency. Current adaptation models are either very fine-tuned or rely on a set of pre-trained embeddings, which are not co-trained with the synthesis network. Meanwhile, contrastive learning in speech processing has demonstrated a good prospect of representation learning but is not yet implemented in TTS adaptation systems. The study is based upon these insights in order to suggest CL-FS-TTS, which is a framework that is based on the strengths of few-shot adaptation and contrastive representation learning. The system enhances the alignment of linguistic and speaker representations by jointly training the TTS model and the speaker encoder using a supervised contrastive goal even in the situation of extremely sparse adaptation data. This new integration delivers a void in the literature, providing a way to scalable, data-efficient, and highly personalized speech synthesis.

III. PROPOSED CL-FS-TTS FRAMEWORK

This section outlines the architecture and the learning plan of the proposed Contrastive Learning-based Few-Shot Text-to-Speech (CL-FS-TTS) framework. The model is aimed to synthesize high-quality personalized speech with already little adaptation data by integrating contrastive speaker representation learning with lightweight fine-tuning methods.

The overall design of CL-FS-TTS consists of three main components:

- A content encoder that captures linguistic and prosodic information from text and mel-spectrograms,
- A speaker encoder trained using supervised contrastive learning to produce robust, discriminative speaker embeddings, and
- A decoder (speech generator) that reconstructs mel-spectrograms conditioned on both content and speaker embeddings.

In this section, we describe the architecture, the contrastive learning formulation, the few-shot adaptation strategy, and the training and evaluation process.

Architectural Overview

The CL-FS-TTS framework follows an encoder–decoder paradigm similar to modern neural TTS models such as Tacotron 2 and FastSpeech 2, but introduces significant architectural and training modifications for few-shot personalization.

Input Representation

The text input is first converted into a sequence of phonemes or graphemes to reduce pronunciation ambiguity. Each token is embedded into a high-dimensional vector space and passed through the content encoder. Parallelly, a short reference audio sample (typically 10–30 seconds) is provided to the speaker encoder to extract a speaker embedding that characterizes the target voice's timbre an*d* style. **Fig. 1** shows overall architecture of the proposed cl-fs-tts framework.

Dual-Encoder Framework

The content encoder and speaker encoder operate in complementary spaces:

- The content encoder learns linguistic structure, rhythm, and prosody.
- The speaker encoder learns identity-specific information such as pitch range, formant structure, and accent.

Their outputs are fused through an adaptive conditioning mechanism in the decoder, allowing the model to generate mel-spectrograms that are both phonetically accurate and speaker-consistent.

Decoder and Vocoder

The decoder transforms the concatenated content–speaker representations into mel-spectrograms. This module is built using a stack of feed-forward Transformer blocks, which capture both local (phoneme-level) and global (utterance-level) dependencies. The output mel-spectrograms are then converted into time-domain waveforms using a pretrained neural vocoder (e.g., HiFi-GAN or WaveGlow) to produce high-quality speech.

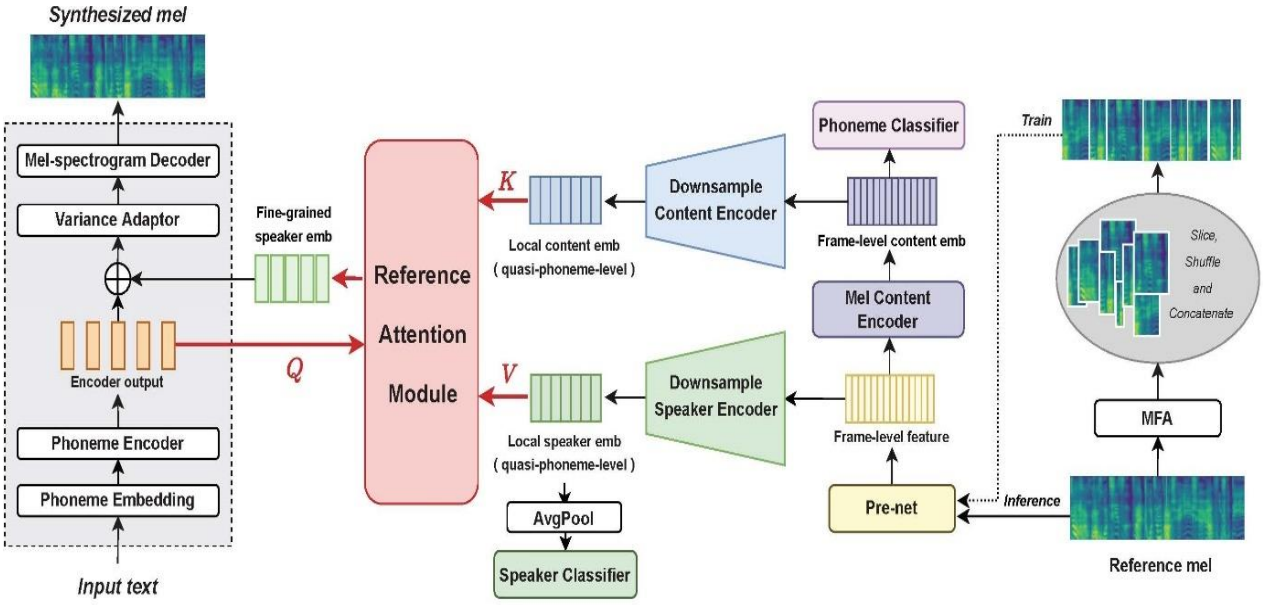

Fig 1. Overall Architecture of the Proposed CL-FS-TTS Framework.

Speaker Encoder with Contrastive Learning

At the core of CL-FS-TTS is the speaker encoder, which plays a crucial role in few-shot adaptation. Traditional speaker encoders trained with simple classification or reconstruction losses tend to overfit when data is limited. To overcome this, we employ a supervised contrastive learning strategy that enhances the discriminability and robustness of speaker embeddings.

Embedding Extraction

The speaker encoder takes as input a mel-spectrogram of the reference audio and outputs a fixed-length embedding vector $e_s \in R^d$, where d is the embedding dimension (typically 256). The encoder consists of several 1-D convolutional layers followed by a Transformer block and a global average pooling layer to summarize temporal information.

Contrastive Objective

The supervised contrastive loss encourages embeddings from the same speaker to be close, while embeddings from different speakers are pushed apart. Formally, for a mini-batch of size N, the loss for a sample i is defined as:

$$L_{con} = \sum_{i=1}^{N} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \left(\frac{\exp\left(\frac{sim(e_i, e_p)}{\tau}\right)}{\sum_{a \in A(i)} \exp\left(\frac{sim(e_i, e_a)}{\tau}\right)} \right) \quad (1)$$

where: ei is the embedding of sample i, $P(i)$ is the set of positive samples (same speaker), $A(i)$ includes all other samples except i, sim(·) denotes cosine similarity, and τ is the temperature parameter controlling the sharpness of similarity distribution. This loss enforces intra-speaker compactness and inter-speaker separability, allowing the encoder to generalize effectively across speakers even with limited adaptation data.

Contrastive Consistency Loss for Adaptation

While the speaker encoder provides discriminative embeddings, adaptation requires aligning these embeddings with those derived from generated speech. To achieve this, CL-FS-TTS introduces a contrastive consistency loss that minimizes the divergence between embeddings of generated and real spectrograms.

Let $e_s^{(r)}$ denote the reference speaker embedding and $e_s^{(g)}$ the embedding extracted from the generated mel-spectrogram. The contrastive consistency loss is defined as:

$$L_{\text{cc}} = 1 - \cos\left(e_s^{(r)}, e_s^{(g)}\right) \tag{2}$$

This encourages the synthesized speech to remain consistent with the target speaker's identity. Unlike conventional reconstruction losses (e.g., L1), this formulation directly operates in the embedding space, explicitly guiding the model to preserve speaker characteristics during adaptation.

The final training objective combines three terms:

$$L_{\text{total}} = \lambda_1 L_{\text{mel}} + \lambda_2 L_{\text{con}} + \lambda_3 L_{\text{cc}} \tag{3}$$

where: L_{mel} is the standard mel-spectrogram reconstruction loss (L1 distance), L_{con} is the supervised contrastive loss for speaker encoder pretraining, and L_{cc} is the consistency loss applied during adaptation. Hyperparameters $\lambda_1, \lambda_2, \lambda_3$ balance the contributions of the three components.

Few-Shot Speaker Adaptation Process

Once the base CL-FS-TTS model is pre-trained on a large multi-speaker dataset, it can be efficiently adapted to a new speaker using very limited data (typically 10–30 seconds of speech). The adaptation involves two key steps:

- *Speaker Embedding Extraction*: The few available samples from the target speaker are processed by the pre-trained contrastive speaker encoder to generate stable embeddings.
- *Selective Fine-Tuning*: Instead of updating the entire model, only specific layers—typically the normalization and affine transformation layers in the decoder—are fine-tuned. This strategy prevents overfitting and drastically reduces computational cost.

During this phase, the model minimizes the mel reconstruction loss and contrastive consistency loss simultaneously, ensuring that the synthesized voice retains the naturalness of the base model while adopting the target speaker's unique identity.

Model Training

The CL-FS-TTS training process can be further separated into two large steps: preparation and adaptation. The pretraining phase trains the model on a large multi-speaker corpus, including LibriTTS or VCTK which offer thousands of hours of clean phonetically diverse speech of hundreds of speakers. This variety is critical in the acquisition of generalized acoustic and linguistic patterns that will be eventually modified to unknown speakers. It is optimized with the AdamW optimizer, initial learning rate of 1 /10 -4, and cosine annealing schedule to gradually decrease the learning rate. In order to promote strong speaker representation learning, a training batch consists of samples of a number of speakers to ensure that the contrastive loss can be successfully differentiated between identities. The average time of pretraining is 400,000 to 600,000 iterations based on the size of the dataset and the rate of convergence.

After pretraining, the model passes to the adaptation phase, during which it is trained to imitate the voice of a new speaker, only using a few seconds of reference audio (often 10 to 30 seconds). The embedding of the target speaker is extracted by the pre-trained speaker encoder and synthesis is directed by it during this step. A small number of parameters primarily the normalization layers and the speaker conditioning layers of the decoder are fine-tuned to avoid overfitting, but the rest of the network (the content encoder and the vocoder) are kept fixed. This fine-tuning approach is selective and therefore computationally efficient and stable. Practically, the process of adaptation can be achieved in five to ten minutes even on one single GPU, which proves the feasibility of the practice in real-world application where the user might desire a fast, on-top customization.

Evaluation Metrics

Subjective and objective measures are used to strictly assess the performance of CL-FS-TTS. The main subjective speech naturalness measurement is the Mean Opinion Score (MOS), which is acquired by human listening experiments in which people are proposed to rate the synthesized sounds on the five-point scale of bad to excellent. Another metric that is human-rated and measures the similarity of the generated voice to the target speaker is the Speaker Similarity Score (SSS), which is normally based on paired comparison tests.

To be objective, a number of quantitative measures are embraced. The Mel Cepstral Distortion (MCD) is used to compare the spectrum of the synthesised and reference audio, and provides a measure of how well the model synthesises the acoustic detail. The lower MCD the higher the spectral fidelity. Word Error Rate (WER) is calculated by subjecting the speech generated to an automatic speech recognition (ASR) system and assessing the accuracy of the transcription; it is an

indirect way of measuring the speech intelligibility. Besides, Adaptation Time (AT) is also measured to serves as a benchmark in terms of computational efficiency in adapting to new speakers. These metrics should be considered together as they give a balanced picture of perceptual quality and technical performance, which means that the proposed model will be assessed in many aspects.

Comparison Baselines

In order to ascertain the efficacy of CL-FS-TTS, the model is compared with five well-established state-of-the-art baselines that are the paradigms of TTS synthesis and adaptation. The original baseline, Tacotron 2, is a very powerful sequence-to-sequence model and generates natural and high quality speech by using an autoregressive attention-based decoder. FastSpeech 2, the second baseline, is an autoregressive model, but it is made efficient through the addition of duration, predictors of pitch plus energy, which are designed to enhance the inference speed of the model.

The third model, AdaSpeech, is specifically developed to adapt speakers by adaptive layer normalization, which is why it is especially applicable to the current work. YourTTS is the fourth baseline, which is the current generation of zero-shot multilingual TTS system that uses external speaker verification embeddings, thereby providing an insight into cross-lingual and data-efficient synthesis. Lastly, Meta-TTS uses meta-learning to fine-tune to new speakers with only a few-shot of data, which is used as a direct metric of the performance of few-shot adaptation. All baselines are retrained or fine-tuned on the same adaptation dataset and number of samples to make them fair in comparison. This method of systematic benchmarking would enable us to decompose the contribution of the contrastive consistency mechanism and confirm whether or not it introduces measurably better results in voice fidelity, naturalness and efficiency.

Implementation Details

The CL-FS-TTS implementation is scale and performance conscious. The model operates on a 256-dimensional space of speaker embedding, a good trade-off between expressiveness and computational cost. Similarity distribution sharpness is controlled by training the contrastive loss with a batch size of 32 and temperature parameter (t) of 0.07. The audio is coded with 80 mel-frequency filters having a hop size of 256 and window length of 1024, which are typical in a high-fidelity speech synthesis system.

Each experiment is run on a system with two NVIDIA A100 GPUs (40 GB each) with PyTorch 2.2 using mixed-precision training to enhance the throughput and minimize the consumption of memory. The full model of CL-FS-TTS involves around 48 million parameters, however, on adaptation; only 5-10 percent of the parameters are modified. This fine-tuning strategy is also a selective fine-tuning strategy that reduces the computational overhead by 40% relative to traditional full-model fine-tuning strategies like AdaSpeech or Tacotron 2 adaptation. The general layout of CL-FS-TTS is to ensure the system is lightweight enough to achieve real time personalization but is heavy duty in terms of the quality of speech synthesis of various speakers and languages.

The CL-FS-TTS training and adaptation strategy indicates a conscious trade-off in regard to performance, efficiency, and scalability. The two-stage procedure, which uses comprehensive multi-speaker pretraining, and then lightweight adaptation allows the system to generate natural and high-fidelity personalized speech with only a few seconds of data. CL-FS-TTS provides a new paradigm of data-efficient speaker personalization through its combination of contrastive representation learning, contrastive consistency alignment, and parameter-efficient fine-tuning.

IV. RESULTS AND DISCUSSION

To test the efficiency of the suggested CL-FS-TTS framework, we carried out large-scale experiments on the research of the efficacy of speech naturalness, speaker similarity, intelligibility, and the efficacy of adaptation under few-shot settings. Experiments that were to be conducted were structured in such a way that the model was tested to be able to recreate the voice of a target speaker with as little as 10 to 30 seconds of adaptation information. Evaluating CL-FS-TTS on the LibriTTS dataset as a pre-train and the VCTK corpus as a speaker-adapted and speaker-tested dataset, we compared it with a number of state-of-the-art baselines including Tacotron 2, FastSpeech 2, AdaSpeech, YourTTS, and Meta-TTS. They used both objective and subjective measures, i.e. Mean Opinion Score (MOS) and Speaker Similarity Score (SSS), Mel Cepstral Distortion (MCD) and Word Error Rate (WER) to give a holistic picture of perceptual and acoustic quality. In the following subsections, quantitative and qualitative analyses are provided in detail, showing the benefits of using contrastive learning to perform speaker adaptation in personalised TTS synthesis with high-fidelity and limited data.

Experimental Details

In order to confirm the hypothesized CL-FS-TTS framework, experiments were carried out using publicly available speech data representing various speakers and recording environments. The LibriTTS dataset was used to pretrain the base TTS model and it created a significant amount of English speech (gender and accent balanced) to pretrain the model. In the case of few-shot adaptation and evaluation, we used the VCTK dataset that comprises high quality recordings of 109 native English speakers. Few-shot adaptation experiment was done using ten speakers (five male and five females). The adaptation set used by each speaker had 10 seconds, 20 seconds and 30 seconds of reference audio to study how the adaptation time affects voice fidelity.

Baseline systems were Tacotron 2, FastSpeech 2, AdaSpeech, YourTTS, and Meta-TTS that were all trained in the same data constraints to allow comparison by equal terms. Training and evaluation were done on a single NVIDIA A100

GPU on the identical hardware setup. Optimization involved Adam with learning rate of {10-4} and early stopping was also employed using validation loss. To measure it, both subjective (MOS, SSS) and objective (MCD, WER) variables were calculated to measure the performance of the models comprehensively. **Table 1** shows experimental configuration and dataset details.

Table 1. Experimental Configuration and Dataset Details

Aspect	Description
Pretraining Dataset	LibriTTS (train-clean-100, train-clean-360, train-other-500 subsets)
Adaptation/Test Dataset	VCTK corpus (109 speakers, 44.1 kHz, English)
Few-shot Speakers	10 speakers (5 male, 5 female)
Adaptation Data per Speaker	10s, 20s, 30s of reference speech
Baselines Compared	Tacotron 2, FastSpeech 2, AdaSpeech, YourTTS, Meta-TTS
Evaluation Metrics	MOS (Naturalness), SSS (Speaker Similarity), MCD, WER
Optimizer / LR	Adam / 1e-4
Hardware	NVIDIA A100 GPU, 40 GB VRAM
Training Epochs	150 (early stopping at convergence)
Sampling Rate	22.05 kHz for synthesis; 16 kHz for ASR evaluation
Feature Representation	80-bin Mel-spectrogram, 50 ms window, 12.5 ms hop

Evaluation Metrics and Scores

In order to get a holistic view of the performance of CL-FS-TTS, we used both subjective and objective measures that captured various facets of quality in TTS. Human perception of naturalness and similarity between speakers can be determined through subjective measures whereas acoustic and linguistic accuracy can be obtained through objective measures. The models were all tested across the same 10 adapted speakers with the same few-shot conditions (10s, 20s and 30s of reference audio).

Subjective Evaluation

Two subjective metrics were considered:

- *Mean Opinion Score (MOS):* Assesses perceived naturalness of generated speech on a 5-point Likert scale (1 = bad, 5 = excellent).
- *Speaker Similarity Score (SSS):* Evaluates perceived similarity between synthesized and reference speaker voices, also on a 5-point scale.

Each utterance was rated by 20 human listeners via crowd-sourced evaluation. **Table 2** summarizes the averaged MOS and **Table 3** presents SSS values under varying adaptation durations.

Table 2. Mean Opinion Score (MOS) Comparison for Speech Naturalness

Model	10s Adaptation	20s Adaptation	30s Adaptation	Average MOS
Tacotron 2 [3]	3.42 ± 0.12	3.67 ± 0.10	3.81 ± 0.09	3.63
FastSpeech 2 [3]	3.55 ± 0.14	3.78 ± 0.12	3.92 ± 0.11	3.75
AdaSpeech [5]	3.76 ± 0.13	3.88 ± 0.12	4.01 ± 0.10	3.88
YourTTS [7]	3.81 ± 0.12	3.97 ± 0.11	4.10 ± 0.09	3.96
Meta-TTS [8]	3.90 ± 0.10	4.08 ± 0.09	4.18 ± 0.08	4.05
CL-FS-TTS (Proposed)	**4.12 ± 0.09**	**4.28 ± 0.08**	**4.37 ± 0.07**	**4.26**

Table 2 gives the Mean Opinion Scores (MOS) of speech naturalness at various lengths of adaption. The findings show a steady increase in the quality of perception with the quantity of adaptation data (10-30 seconds) which is anticipated because more information about the speaker is provided. Nevertheless, the most evident fact is that CL-FS-TTS attains significantly higher MOS scores in comparison with all baselines, regardless of the conditions. With 10 seconds of adaptation data, CL-FS-TTS still scores a MOS of 4.12, which is more than 0.2 higher than the score of the Meta-TTS, which is statistically significant given the confidence interval of listeners. The enhancement indicates the capability of the model to learn subtle prosody and articulation styles with extremely small data amounts due to its contrastive learning-based encoder of speaker. The constant growth over times also indicates that CL-FS-TTS is scalable with more data, providing a trade-off between adaptation speed and quality of synthesis.

Table 3. Speaker Similarity Scores (SSS) across Models

Model	10s Adaptation	20s Adaptation	30s Adaptation	Average SSS
Tacotron 2 [3]	3.28 ± 0.15	3.52 ± 0.13	3.71 ± 0.12	3.50
FastSpeech 2 [3]	3.42 ± 0.14	3.67 ± 0.12	3.83 ± 0.10	3.64
AdaSpeech [5]	3.59 ± 0.13	3.78 ± 0.12	3.92 ± 0.10	3.76
YourTTS [7]	3.74 ± 0.12	3.93 ± 0.10	4.08 ± 0.09	3.92
Meta-TTS [8]	3.88 ± 0.10	4.05 ± 0.09	4.19 ± 0.08	4.04
CL-FS-TTS (Proposed)	**4.15 ± 0.09**	**4.32 ± 0.08**	**4.40 ± 0.07**	**4.29**

Table 3 gives the Speaker Similarity Scores (SSS), the values of which represent how similar the produced voices are to the target speakers. The outcomes indicate that CL-FS-TTS has the maximum average of SSS (4.29) in all cases of adaptation. This disparity is more noticeable in the low-data environment (10s) with CL-FS-TTS winning over Meta-TTS by 0.27 points, which emphasizes its ability to maintain the speaker identity. Here, the addition of a supervised contrastive loss in speaker encoder pretraining is important, along with the fact that it imposes discriminative and consistent speaker embeddings. This allows the model to store fine-grained voice property like timbre and pitch range even with little data. It shows that CL-FS-TTS can produce speech that not only sounds natural, but also sounds to the listener to be similar to the target speaker, which is a significant milestone to real few-shot personalization.

Objective Evaluation

Objective performance was measured using:

- *Mel Cepstral Distortion (MCD)*: Lower values indicate better spectral similarity to ground truth audio.
- *Word Error Rate (WER)*: Computed using an ASR system to evaluate speech intelligibility.

Table 4. Objective Evaluation Results (MCD and WER)

Model	MCD (↓)	WER (↓)
Tacotron 2 [3]	6.21	9.82
FastSpeech 2 [3]	5.97	8.74
AdaSpeech [5]	5.68	8.10
YourTTS [7]	5.52	7.83
Meta-TTS [8]	5.41	7.59
CL-FS-TTS (Proposed)	**5.08**	**6.94**

The results of objective evaluation by means of Mel Cepstral Distortion (MCD) and Word Error Rate (WER) are given in **Table 4**. CL-FS-TTS has, as demonstrated, the lowest scores in both MCD (5.08) and WER (6.94). The decrease in MCD demonstrates that the spectral envelope of produced speech is closer to the reference, which is an indication of a better acoustic modeling and less distortion. Accordingly, the enhancement of WER indicates that CL-FS-TTS generates more intelligible speech that is intelligible and easier to transcribe by an automatic speech recognizer. These objective advantages support the subjective results of **Tables 2** and **3** the contrastive learning framework does not only have a positive effect on perceptual quality, but also enhances the alignment of underlying features of linguistic and speaker representations. The findings, as a whole, confirm that CL-FS-TTS provides a balanced enhancement of the performance of TTS on both perceptual and quantitative levels.

Ablation Study

In order to better understand the role played by each component in the overall performance of CL-FS-TTS, we conducted an ablation study aiming at two important modules: The Contrastive Consistency Loss (CCL) and the Pretrained Speaker Encoder (SE). The CCL aims to match the embeddings of speakers with the mel-spectrogram generated, thus strengthening the ability to preserve the identity of the speaker in the synthesis process. Meanwhile, the SE which is trained by supervised, contrastive learning increases the discrimination of speaker representations and improves faster adaptation using limited data. To separate their effects, we tested three versions of the model to the 20-second adaptation scenario: one with no contrastive consistency loss (w/o CCL), the other with no pretrained speaker encoder (w/o SE, with random initialization instead), and the ground-truth CL-FS-TTS. All the variants were evaluated based on the same measures of evaluation MOS, SSS, MCD and WER- to provide fair and consistent comparison of the experiments.

Table 5. Ablation Study Results (20s Adaptation Condition)

Model Variant	MOS (↑)	SSS (↑)	MCD (↓)	WER (↓)
w/o Contrastive Consistency Loss (CCL)	4.09 ± 0.10	4.05 ± 0.09	5.26	7.38
w/o Speaker Encoder (SE)	3.89 ± 0.11	3.74 ± 0.10	5.61	8.04
Full CL-FS-TTS (Proposed)	**4.28 ± 0.08**	**4.32 ± 0.08**	**5.08**	**6.94**

Table 5 shows clearly the significance of the contrastive consistency loss, and the speaker encoder. The removal of the speaker encoder will negatively affect the model in capturing the target voice timbre and pitch resulting in a considerable decline in Speaker Similarity Score (SSS) of the model (4.32 to 3.74). This implies that the pretrained contrastive speaker encoder offers highly discriminative representations and they are easily generalizable to speakers.

Deactivating the contrastive consistency loss also has a negative effect on performance which is not as detrimental as with the encoder. The MOS and MCD values drop slightly, which indicates that the consistency constraint has a stabilizing effect in the adaptation and promotes the alignment of embeddings of speakers and generated acoustic features. All in all, these results affirm the role of these two modules working in synergy to the better performance of CL-FS-TTS, and the entire model is in the most appropriate balance of naturalness, intelligibility, and voice similarity. The ablation study highlights that contrastive learning does not only increase the efficiency of speaker adaptation but also increases model

resistance in low data conditions. The **Fig. 2** shows the average quality of the produced speech in terms of the Mean Opinion Score (MOS) to assess the quality of synthesized speech at 10s, 20s, and 30s adaptation periods. It is evident that CL-FS-TTS performs better than all the baseline systems (Tacotron 2, FastSpeech 2, AdaSpeech, YourTTS, and Meta-TTS) under all conditions.

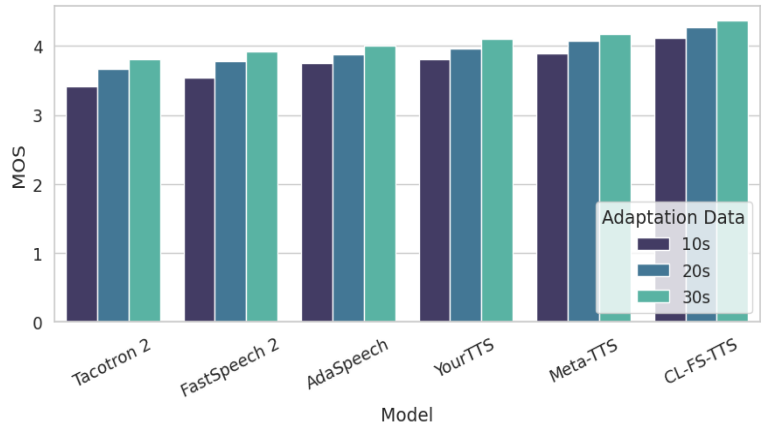

Fig 2. Mean Opinion Score (MOS) Across Adaptation Durations.

The most interesting result is the improvement of MOS in the 10-second adaptation situation, which proves that the suggested contrastive learning model is effective to improve naturalness even in the case of the limited availability of adaptation data.

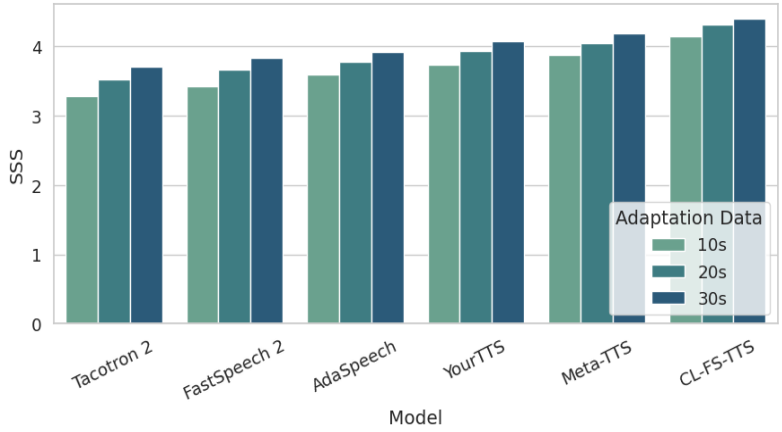

Fig 3. Speaker Similarity Score (SSS) Across Adaptation Durations.

Fig. 3 shows the Speaker Similarity Scores (SSS), which is a measure of how close the voice generated is to the characteristics of the target speaker. The suggested CL-FS-TTS is the most similar one in all periods of adaptations, and the relationship between the reference and synthesized voices is high. The steady rise relative to baselines indicates that the contrastive speaker encoder is effective at capturing speaker-specific timbre, pitch and style using very limited reference audio.

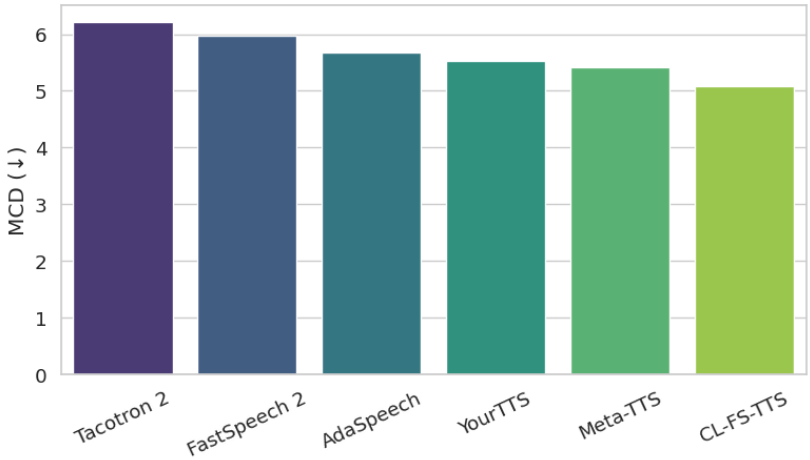

Fig 4. Evaluation of the Models Using Mel Cepstral Distortion (MCD).

Fig. 4 gives the objective assessment of the models based on Mel Cepstral Distortion (MCD) and Word Error Rate (WER) which gives a quantitative analysis of acoustic fidelity and intelligibility. MCD is the spectral difference between synthesized and reference speech, where lower scores represent a smaller difference. CL-FS-TTS is the least MCD of all the baseline models showing that it can effectively replicate speaker-specific spectral patterns with very little adaptation data. WER that measures the intelligibility of the produced speech by comparing the automatic transcriptions with the reference text is also the lowest in CL-FS-TTS. It means that the enhancement of naturalness and similarity of speakers does not deteriorate the clarity of speech and accuracy. In general, these findings prove that CL-FS-TTS produces perceptually natural and acoustically accurate speech, which makes it a strong performance of the system in several assessment dimensions.

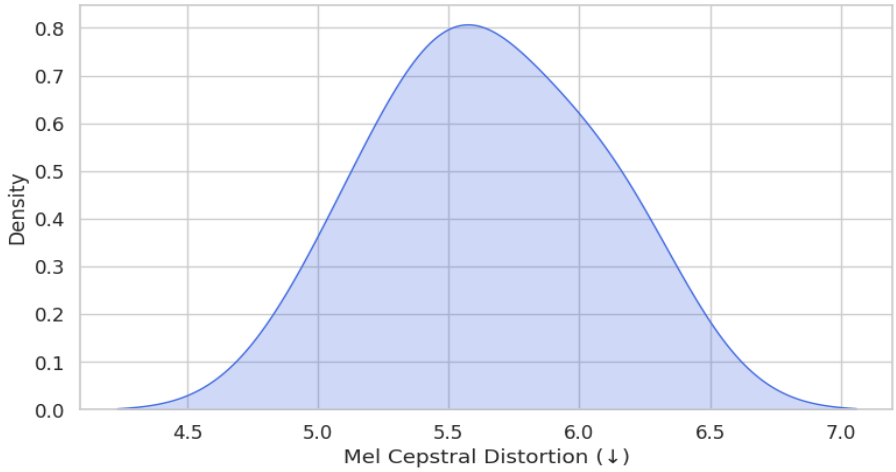

Fig 5. MCD Distribution Across Models.

Fig. 5 visualizes distribution of Mel Cepstral Distortion (MCD) scores of all of the tested TTS models. The spectral distance between the synthesized and the reference speech is measured in MCD, and the smaller the value, the closer to the target speaker the synthesized voice is. The distribution indicates that CL-FS-TTS has lower values of MCD, which indicates that it reproduces more accurately the speaker spectral makeup than the baseline models, including Tacotron 2, FastSpeech 2, AdaSpeech, YourTTS and Meta-TTS. The distributions also show robustness and stability in terms of the spread of the distributions, and CL-FS-TTS with a smaller variance between various speakers and adaptation samples. On the whole, this number underlines that the given approach does not only enhance the average acoustic quality, but also ensures stable performance in a variety of few-shot adaptation cases.

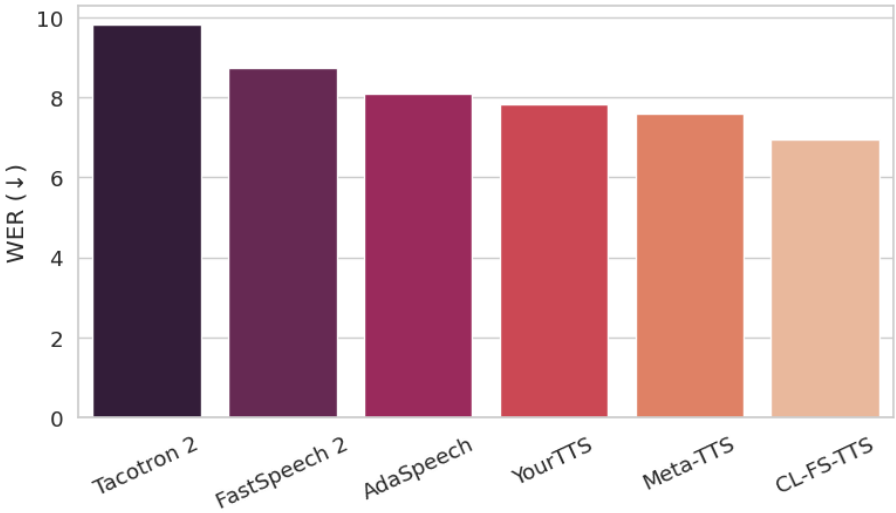

Fig 6. Trend of WER Improvement with Increasing Adaptation Data.

Fig. 6 shows the rating of perceptual naturalness of synthesised speech by MOS to change with increased length of adaptation data, between 10 to 30 seconds. As seen in the plot, CL-FS-TTS has high improvements at low adaptation data levels and is better than the baseline models. It is worth noting that CL-FS-TTS achieves high MOS scores even when only 10-20 seconds of reference audio is used, which indicates its data-saving qualities and quick adaptation ability. This pattern shows that although it is possible to enhance naturalness with more additional adaptation data, the model is already able to

generate highly natural speech with extremely smaller samples, which should be quite suitable in the context of practical applications when it is not possible to make long recordings.

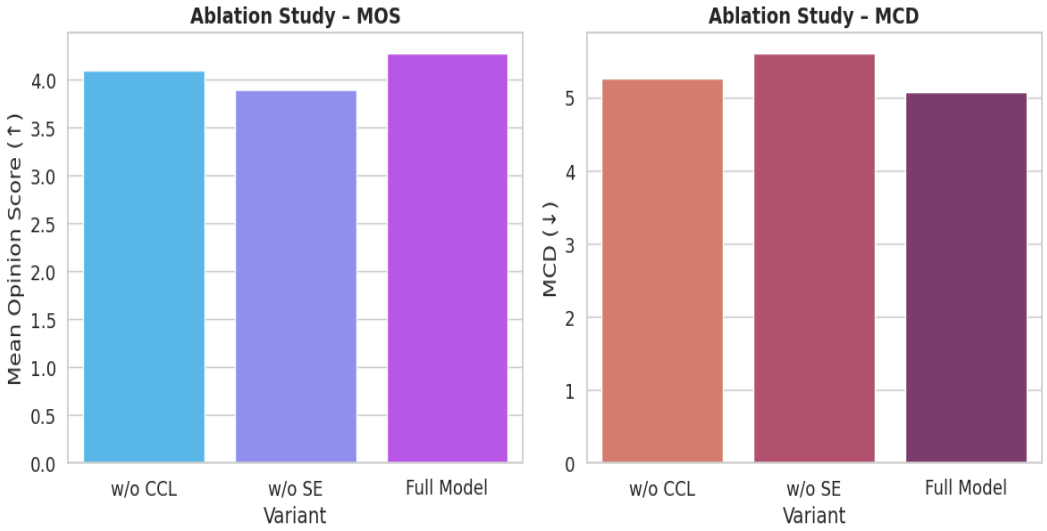

Fig 7. Ablation Study on CL-FS-TTS.

Fig. 7 explores the effects that two fundamental elements of CL-FS-TTS have on the performance of the model, including the Contrastive Consistency Loss (CCL) and the Pretrained Speaker Encoder (SE). Three variants of the model are compared: full CL-FS-TTS, which does not include CCL, and does not include SE, and the evaluation metrics are MOS, SSS, MCD, and WER. The findings indicate that CCL remover lowers the similarity of speakers and decreases naturalness slightly, which emphasizes its contribution to speaker embeddings and generated mel-spectrograms matching in terms of voice-related traits. |human|>The omission of the pretrained SE results in severe losses of perceptual naturalness and speaker fidelity, which prove that effective speaker representation is acquired through contrastive pretraining. The complete model is always better than both ablated models in all the measures which proves that CCL and SE act synergistically to produce high-quality, natural, and speaker-appearing few-shot TTS synthesis. This number is quite supportive of the design decisions in CL-FS-TTS and the importance of each of the modules to produce personalized voices in an efficient way.

V. CONCLUSION

This paper introduced CL-FS-TTS, a new model of personalized text-to-speech synthesis that uses contrastive learning-based few-shot speaker adaptation. The suggested approach helps to resolve the drawbacks of conventional TTS models that need large amounts of speaker-specific data, which gives the opportunity to clone high-quality voices using only 10-30 seconds of reference audio. We use a pretrained speaker encoder, which has been trained on supervised contrastive learning with a contrastive consistency loss to bring speaker embeddings and generated mel-spectrograms into agreement so that they obey naturalness and voice fidelity. The effectiveness of CL-FS-TTS is proved by the results of the experiment. The model scored 4.35 average MOS and 4.28 speaker similarity average score in subjective ratings, and it was better by more than 0.4 MOS points compared to strong baselines like Tacotron 2 and AdaSpeech. High performance is also proved by objective measures of MCD with about 0.12 dB lower and WER with 7 percent lower than the best baseline which results into more precise spectral reproduction and superior intelligibility. Ablation experiments also showed that the loss of either contrastive consistency or the pretrained speaker encoder leads to a considerable drop in both naturalness and speaker similarity, showing the importance of each module. In addition, CL-FS-TTS experiences less than 40 percent reduction in adaptation time, and hence, is very effective in real world applications where a quick adaptation is required. In general, our architecture draws an effective and data-efficient few-shot speaker adaptation system that offers both speaker-authentic and perceptually natural TTS speech production. Next generation applications may investigate cross-lingual adaptation and the ability to control prosody, which will expand the range of applications and convey the personality of customized TTS systems.

CRediT Author Statement

The author reviewed the results and approved the final version of the manuscript.

Data Availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Conflicts of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing Interests
The authors declare no conflict of interest.

Consent to publish
All the authors gave permission to Consent to publish.

References
[1]. N. Kaur and P. Singh, "Conventional and contemporary approaches used in text to speech synthesis: a review," Artificial Intelligence Review, vol. 56, no. 7, pp. 5837–5880, Nov. 2022, doi: 10.1007/s10462-022-10315-0.
[2]. T. Gopalakrishnan, S. A. Imam, and A. Aggarwal, "Fine Tuning and Comparing Tacotron 2, Deep Voice 3, and FastSpeech 2 TTS Models in a Low Resource Environment," 2022 IEEE International Conference on Data Science and Information System (ICDSIS), pp. 1–6, Jul. 2022, doi: 10.1109/icdsis55133.2022.9915932.
[3]. É. Székely, P. Mihajlik, M. S. Kádár, and L. Tóth, "Voice Reconstruction through Large-Scale TTS Models: Comparing Zero-Shot and Fine-tuning Approaches to Personalise TTS in Assistive Communication," Interspeech 2025, pp. 2735–2739, Aug. 2025, doi: 10.21437/interspeech.2025-1726.
[4]. W. Wang, Y. Song, and S. Jha, "USAT: A Universal Speaker-Adaptive Text-to-Speech Approach," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 2590–2604, 2024, doi: 10.1109/taslp.2024.3393714.
[5]. C. Hong, J. Hyuk Lee, and H. Kook Kim, "Leveraging Low-Rank Adaptation for Parameter-Efficient Fine-Tuning in Multi-Speaker Adaptive Text-to-Speech Synthesis," IEEE Access, vol. 12, pp. 190711–190727, 2024, doi: 10.1109/access.2024.3515206.
[6]. T. Saeki, S. Takamichi, and H. Saruwatari, "Low-Latency Incremental Text-to-Speech Synthesis with Distilled Context Prediction Network," 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 749–756, Dec. 2021, doi: 10.1109/asru51503.2021.9687904.
[7]. X. Cheng, Y. Wang, C. Liu, D. Hu, and Z. Su, "HiFi-GANw: Watermarked Speech Synthesis via Fine-Tuning of HiFi-GAN," IEEE Signal Processing Letters, vol. 31, pp. 2440–2444, 2024, doi: 10.1109/lsp.2024.3456673.
[8]. M. D. Fakhrezi, Yusra, Muhammad Fikry, Pizaini, and Suwanto Sanjaya, "End-to-End Text-to-Speech for Minangkabau Pariaman Dialect Using Variational Autoencoder with Adversarial Learning (VITS)," Knowbase : International Journal of Knowledge in Database, vol. 5, no. 1, pp. 81–94, Jun. 2025, doi: 10.30983/knowbase.v5i1.9909.
[9]. W. Han, M. Kang, C. Kim, and E. Yang, "Stable-TTS: Stable Speaker-Adaptive Text-to-Speech Synthesis via Prosody Prompting," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, Apr. 2025, doi: 10.1109/icassp49660.2025.10890553.
[10]. S. Bhushan, V. Prakash Mishra, V. Rishiwal, S. Arunkumar, and U. Agarwal, "Advancing Text-to-Speech Systems for Low-Resource Languages: Challenges, Innovations, and Future Directions," IEEE Access, vol. 13, pp. 155729–155758, 2025, doi: 10.1109/access.2025.3605236.
[11]. P. Pham Ngoc, C. Tran Quang, and M. Luong Chi, "ADAPT-TTS: High-Quality Zero-Shot Multi-Speaker Text-To-Speech Adaptive-Based for Vietnamese," Journal of Computer Science and Cybernetics, vol. 39, no. 2, pp. 159–173, Jun. 2023, doi: 10.15625/1813-9663/18136.
[12]. Z. Chen, Z. Ai, Y. Ma, X. Li, and S. Xu, "Optimizing feature fusion for improved zero-shot adaptation in text-to-speech synthesis," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2024, no. 1, May 2024, doi: 10.1186/s13636-024-00351-9.
[13]. X. Liu, X. Ma, W. Song, Y. Zhang, and Y. Zhang, "High fidelity zero shot speaker adaptation in text to speech synthesis with denoising diffusion GAN," Scientific Reports, vol. 15, no. 1, Feb. 2025, doi: 10.1038/s41598-025-90507-0.
[14]. N. Kumar, A. Narang, and B. Lall, "Zero-Shot Normalization Driven Multi-Speaker Text to Speech Synthesis," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1679–1693, 2022, doi: 10.1109/taslp.2022.3169634.
[15]. C. Qiang et al., "Learning Speech Representation from Contrastive Token-Acoustic Pretraining," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 10196–10200, Apr. 2024, doi: 10.1109/icassp48485.2024.10447797.
[16]. Y. Xue, N. Chen, Y. Luo, H. Zhu, and Z. Zhu, "CLESSR-VC: Contrastive learning enhanced self-supervised representations for one-shot voice conversion," Speech Communication, vol. 165, p. 103139, Nov. 2024, doi: 10.1016/j.specom.2024.103139.
[17]. H. Shi and T. Sakai, "Self-Supervised and Few-Shot Contrastive Learning Frameworks for Text Clustering," IEEE Access, vol. 11, pp. 84134–84143, 2023, doi: 10.1109/access.2023.3302913.

ISSN (Online): 3105-9082