

# Voice Cloning: Comprehensive Survey

Hussam Azzuni, and Abdulmotaleb El Saddik, *Fellow, IEEE*,

**Abstract**—Voice Cloning has rapidly advanced in today’s digital world, with many researchers and corporations working to improve these algorithms for various applications. This article aims to establish a standardized terminology for voice cloning and explore its different variations. It will cover speaker adaptation as the fundamental concept and then delve deeper into topics such as few-shot, zero-shot, and multilingual TTS within that context. Finally, we will explore the evaluation metrics commonly used in voice cloning research and related datasets. This survey compiles the available voice cloning algorithms to encourage research toward its generation and detection to limit its misuse.

**Index Terms**—Voice Cloning, Few-shot TTS, Zero-shot TTS, Speaker adaptation.

## I. INTRODUCTION

**V**OICE Cloning is the ability to replicate a person’s voice. Advancing these algorithms relies on enhancing the performance of Text-to-Speech (TTS) systems in various areas, including speech quality, naturalness, prosody, and timbre, ensuring the produced voice closely resembles the target speaker. The development of voice cloning can revolutionize many fields as we shift toward a more digital world. These applications include voice-over, personalized virtual assistants, and Human-Computer Interaction (HCI).

Voice cloning has been progressing rapidly, as demonstrated in Fig. 1; thus, establishing a comprehensive survey is crucial for further research in its generation and detection. The first step is reviewing previous literature that discussed deepfake generation and detection summarized in Table I. We explore these works to gain insights before building on more recent methods in the domain. Tan et al. [1] provided a comprehensive analysis of neural speech synthesis while delving slightly into more advanced topics, including speaker adaptation. Kadam et al. [2] surveyed methods from voice cloning and lip synchronization, including techniques such as concatenative TTS and parametric TTS. Masood et al. [3] explored deepfake generation and detection for visual and audio modalities. This survey includes visual manipulation such as face swapping, puppet mastery, lip-syncing, entire face synthesis, and facial attribute manipulation. It also focuses on TTS and voice conversion (VC) for audio deepfakes. Similarly, Khanjani et al. [4] explored audio, text, video, and image deepfakes while delving deeper into audio deepfake generation and detection, including both TTS and VC. More recently, Ramu et al. [5] explored voice cloning in the context of multilingual video dubbing. Our survey differs from the previous papers as we will investigate and provide an in-depth study of audio deepfake generation, specifically in TTS systems, commonly called voice cloning. This document includes establishing a

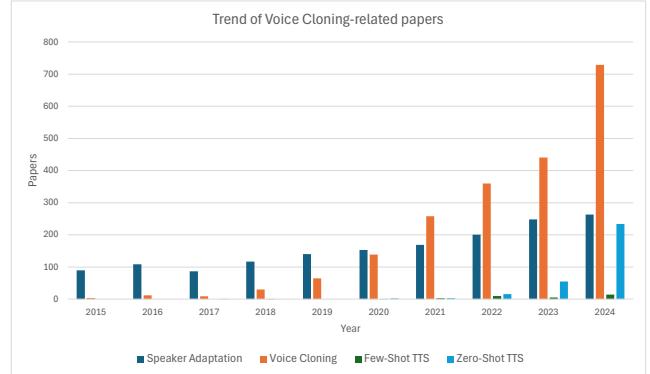


Fig. 1. Bar graph showing the trend in domains related to speaker adaptation.

standardized terminology while going into more advanced topics, including few-shot, zero-shot, and multilingual TTS.

Voice cloning definition has varied across research. It differs from speech synthesis as it aims to preserve the speaker’s characteristics. However, the amount of data required is in question. Some research refers to it similarly to speaker adaptation with no constraint on data availability [3], [6], [7]. On the other hand, others state that voice cloning requires the usage of limited amount of data to replicate the target speaker’s voice [8]–[14]. Nautilus [7] takes another approach in defining voice cloning, as the authors state that voice cloning is a big umbrella of any technology that can generate speech imitating a speaker’s characteristic. Cross-lingual voice cloning [15] differs from the previous definitions as it adapts known speakers to unknown languages, which is not within the scope of this survey. For this work, we propose establishing a standard definition for voice cloning as an extension of TTS speaker adaptation. It is essential to note that Table II will cover all the abbreviations within this document.

- 1) **Voice Cloning:** Replicating a specific person’s voice using a TTS system.
- 2) **Speaker Adaptation:** Fine-tuning a TTS model to replicate a specific user’s voice using limited data.
- 3) **Few-shot Voice Cloning:** This follows the central concept of speaker adaptation. However, the difference is the amount of data required. Thus, the reference audio can range from a few seconds to a maximum of 5 minutes, which is decided based on previous work, and anything more is challenging to obtain in real-life scenarios.
- 4) **Zero-shot Voice Cloning:** This fundamentally differs from the previous approaches as you do not need to fine-tune the TTS model. A specialized module, such as a speaker encoder, is required to use a short audio clip to generate speech with voice characteristics similar to the reference waveform.

Hussam Azzuni performed this work during his stay with the Computer Vision Department, MBZUAI, UAE

Abdulmotaleb El Saddik is with the University of Ottawa, Canada

TABLE I  
RELEVANT LITERATURE REVIEW.

Paper	Year	Conference/Journal	Scope
Ref. [5]   2024   International Conference on Wireless Communication Signal Processing and Networking		Multilingual video dubbing, focusing on voice cloning and machine translation	
Ref. [4]   2023		Frontiers in Big Data	Audio deepfake generation and detection
Ref. [3]   2023		Applied Intelligence	Audio and video deepfake generation and detection
Ref. [2]   2021		EAI Endorsed Transactions on Creative Technologies	Voice cloning and lip synchronization
Ref. [1]   2021		Arxiv	Neural TTS, including components and advanced topics

We follow a multi-stage process to include papers in this survey. First, we review the previous literature on TTS speaker adaptation and voice cloning and select the articles aligned with our scope. Second, we perform a comprehensive search for relevant English methodology papers using keywords such as "voice cloning," "few-shot voice cloning," "zero-shot voice cloning," and "speaker adaptation" to find relevant papers. We exclude voice conversion, singing voice synthesis (SVS), and speech enhancement (SE) papers. Additionally, we exclude 2024 non-peer-reviewed documents, as they may still be works in progress unless they have already had a significant impact.

This survey will comprehensively explore the current status of voice cloning, including speaker adaptation, few-shot, zero-shot, and multilingual voice cloning. Establishing a voice cloning reference paper is crucial for further development and exploring audio deepfake generation and its detection to limit its misuse in real-world scenarios. Section II provides a taxonomy of TTS system components. Understanding the main components is crucial for developing the basic knowledge of voice cloning within various models. Section III explains background information regarding the most common TTS algorithms as they are the basis of much recent work. In addition, information regarding pre-training and training algorithms is needed for an in-depth understanding of some contributions. Section IV explores voice cloning variants with additional subcategories based on researchers' contributions. Section V goes through speech-related datasets and explores the evaluation metrics used for TTS systems. Section VI goes through potential applications for voice cloning across various sectors and certain misuse cases to familiarize the reader with the dangers of voice cloning. Finally, Section VII will discuss future directions to ensure growth in voice cloning.

## II. TEXT-TO-SPEECH TAXONOMY

Voice Cloning is the process of adapting a TTS system to unseen speakers. Understanding its various categories is crucial in building on previous knowledge and making novel contributions. These systems follow a modular approach, starting from a text/phoneme until an audio waveform. TTS systems use various technologies such as Convolutional Neural Networks (CNN), Variational autoencoders (VAE), Generative Adversarial Networks (GAN), Flow-based generative models, Diffusion Models, and more recently, Codec-based networks. Here are the main three components.

- 1) Text-Analysis: This is the first module of a TTS system, known as the front-end, which generates linguistic features from a text input. This process includes text

TABLE II  
TABLE OF COMMON ABBREVIATIONS.

Abbreviation	Full Name		Abbreviation	Full Name
TTS system				
FastSpeech 1/2	FS/FS2	StyleTTS & StyleTTS2		STTS/STTS2
NaturalSpeech 2/3	NS2/NS3	Guided-TTS & Guided-TTS 2		GTTS/GTTS2
MegaTTS & MegaTTS2	MTTSM/MTTS2	HierSpeech & HierSpeech++		HS/HS++
YourTTS	YTTS	Content-Dependent Fine-Grained Speaker Embedding		CDFSE
StyleSpeech & Meta-StyleSpeech	SS/MSS	VoiceBox		VB
Fundamental Neural Network Concepts				
Variational Autoencoder	VAE	Convolutional Neural Networks		CNN
Deep Neural Network	DNN	Recurrent Neural Network		RNN
Bidirectional Long Short-Term memory	Bi-LSTM	Generative Adversarial Network		GAN
Sequence-to-Sequence	Seq2Seq	Denoising Diffusion Probabilistic Models		DDPM
Self-Supervised Learning	SSL	Model-Agnostic Meta-Learning		MAML
Language-Agnostic Meta-Learning	LAML	Autoregressive & Non-Autoregressive		AR/NAR
Language Preprocessing				
Byte-Pair encoding	BPE	Grapheme to Phoneme		G2P
International Phonetic Alphabet	IPA	-		-
Speech Representation				
Fundamental frequency	f0	Vector-Quantization VAE		VQ-VAE
Phoneme-Posteriorgram	PPG	Five-time sampling		FTS
Lookup Table	LUT	Learnable Dictionary Encoding		LDE
TTS and speech-related applications				
Human-Computer Interaction	HCI	Singing Voice Synthesis		SVS
Text to Speech	TTS	Few-Shot TTS & Zero-Shot TTS		FS-TTS/ZS-TTS
Language Identification	LID	Language Understanding		LU
Machine Translation	MT	Speech-to-Speech Translation		S2ST
Automatic Speech Recognition	ASR	Automatic Phoneme Recognition		APR
Speaker Verification	SV	Speaker Identification		SI
Voice Conversion	VC	Emotional Voice Conversion		EVC
Speech Emotion Recognition	SER	Speech Enhancement		SE
Speaker & Style Modeling				
Global Style Token	GST	Style-adaptive Layer Normalization		SALN
Adaptive Instance Normalization	AdaIN	Adaptive Layer Normalization		AdaLN
Conditional Layer Normalization	CLN	-		-
Language models in TTS systems				
Large Language Model	LLM	Neural Codec Language Model		NCLM
Evaluation				
Mean Opinion Score	MOS	Comparative MOS		CMOS
Speaker Verification Equal Error Rate	SV-EER	Word/Character Error Rate		WER/CER
Real-Time Factor	RTF	Speaker Embedding Cosine Similarity		SECS
Gross Pitch Error	GPE	Voicing Decision Error		VDE
F0 Frame Error	FFE	Mel-Cepstral Distortion		MCD
Mel-Frequency Cepstral Coefficient	MFCC	-		-

normalization, word segmentation if applicable, part-of-speech tagging, prosody prediction, and grapheme-to-phoneme (G2P) conversion [1].

- 2) Acoustic Model: This module, called the parameter prediction module, generates acoustic features from linguistic features or text/phonemes input, focusing on voice characteristics like prosody and timbre. Prosody modeling contains elements such as the speech's rhythm, intonation, and stress to enable expressive speech [16], evaluated through fundamental frequency (f0), duration, and energy. On the other hand, timbre represents the unique qualities of every person's voice.
- 3) Vocoder: This module takes acoustic or linguistic features and uses them to generate an audio waveform.

## III. RELEVANT BACKGROUND INFORMATION

This section will discuss the basic principles of various topics to ensure the reader has the necessary knowledge to

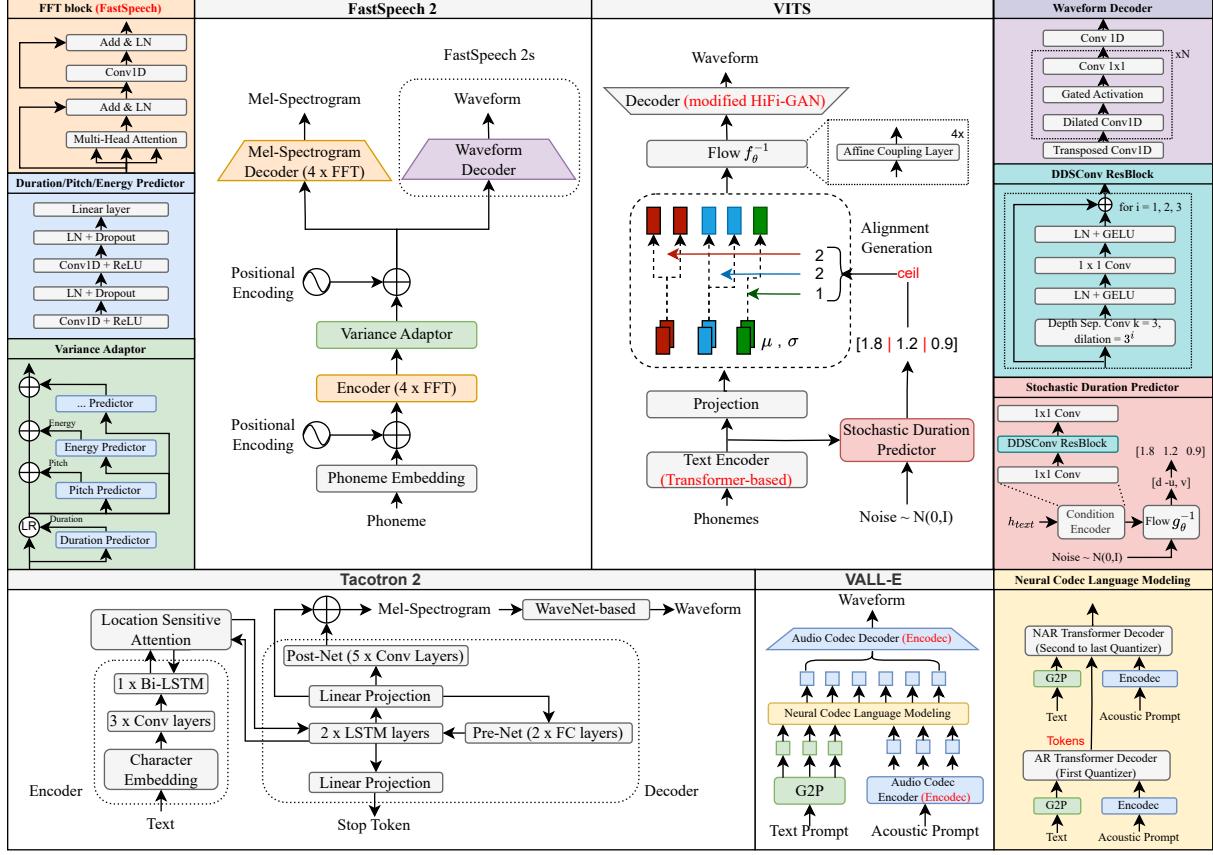


Fig. 2. Common TTS algorithms, adapted from Tacotron 2 [17], FastSpeech 2 [21], VITS [23], and VALL-E [24].

better understand authors' contributions to voice cloning and its variants. These topics include common TTS algorithms, pre-training algorithms, and training paradigms.

#### A. Common TTS algorithms

TTS algorithms can be categorized into various architectural classes. Fig. 2 shows the most common TTS models. Tacotron 2 [17] builds upon Tacotron [18] by introducing key modifications to ensure smoother training and better speech quality without relying on feature engineering. It is a sequence-to-sequence (Seq2Seq) model with an encoder-decoder architecture and an attention mechanism to map a text input sequence to a mel spectrogram sequence during acoustic modeling. Unlike Tacotron, which uses the Griffin-Lim algorithm [19], Tacotron 2 replaces it with an independently trained, modified WaveNet vocoder [20], leading to an improved synthesized speech quality. Similarly, FastSpeech 1/2 [21], [22] is a transformer-based Seq2Seq model that uses feed-forward network instead of an encoder-decoder architecture. This modification speeds up the generation process by adopting a non-autoregressive (NAR) generation methodology compared to the autoregressive (AR) Tacotron 2. FastSpeech 2 builds on top of its predecessor by removing the distillation process and training directly on ground-truth mel spectrogram and the introduction of conditional inputs to handle the one-to-many mapping problem; this problem occurs as words

are spoken in multiple ways from the same text. It also introduces FastSpeech 2s, the first attempt to generate a speech waveform in parallel through a GAN-based waveform decoder for end-to-end speech generation. VITS [23] is an end-to-end TTS architecture that utilizes VAE and flow-based modeling to unify acoustic and vocoder representations for speech synthesis. It enhances the speech's expressiveness by applying normalizing flows to the conditional prior distribution and improves speech quality using waveform-based adversarial training. It also tackles the one-to-many relationship problem by introducing a stochastic duration predictor, enabling speech generation with diverse rhythms. With the emergence of Large Language Models (LLMs), TTS algorithms started integrating language modeling to enhance performance. VALL-E [24] is a Neural Codec Language Model (NCLM) that regards TTS as a conditional language modeling task instead of a continuous signal regression task. Instead of obtaining an intermediate mel spectrogram, it uses discrete units based on phoneme and acoustic prompts. This model is capable of zero-shot TTS (ZS-TTS) using an acoustic prompt input to generate a waveform maintaining the speaker's emotion and voice characteristics. It consists of 4 components: a G2P module, an Audio Codec encoder and decoder following EnCodec [25], and an NCLM. The model is trained through in-context learning, similar to GPT-3 [26], enabling prompt-based approaches for ZS-TTS.

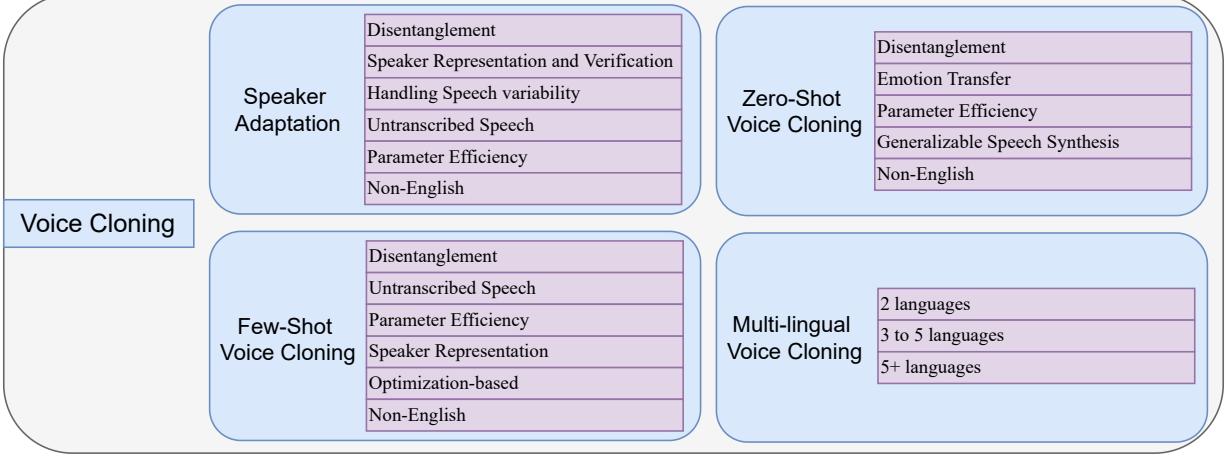


Fig. 3. Overview of the Survey paper.

TABLE III  
SPEECH PRE-TRAINING ALGORITHMS ADAPTED FROM [32].

Model	Year	Input	Output	Downstream tasks
Wav2Vec 2.0 [27]	2020	Speech	Discrete units	ASR, APR
Hubert [28]	2021	Speech	Discrete Units	ASR
SLAM [29]	2021	Speech, <Text-Speech>	Discrete Units	ASR, ST, LU, TN
MAM [30]	2020	Speech	Spectrogram	ST
FAT-MLM [31]	2021	Speech, <Text-Speech>	Spectrogram	ASR, ST, MT
A <sup>3</sup> T [32]	2022	Speech, <Text-Speech>	Spectrogram	TTS, Speech Editing

### B. Speech Pre-training Algorithms

Understanding speech pre-training algorithms is essential, as some works utilize a pre-trained network for feature extraction. They can be categorized based on input, output, and downstream tasks, as shown in Table III. Wav2Vec 2.0 [27], Hubert [28], and SLAM [29] all outputs discrete speech units for speech understanding as observed from their downstream tasks. On the other hand, Masked Acoustic Modeling (MAM) [30] and FAT-MLM [31] outputs reconstructed spectrogram as part of their training. More recently, A<sup>3</sup>T [32] is a pre-training framework for direct application for Text-to-Speech systems as an alternative to Wav2Vec 2.0 and Hubert.

### C. Adversarial Training

Adversarial training is a technique to improve the robustness of deep learning models by exposing them to adversarial examples, which are modified inputs designed to cause incorrect predictions. One common adversarial training approach is through a GAN [33], where a generator and a discriminator are trained simultaneously. The generative model creates outputs that follow the distribution of the input dataset, while the discriminator tries to discern between the generated sample and the actual input dataset. TTS systems employ adversarial training to improve specific aspects, such as the speech's naturalness or the speaker similarity of the generated speech.

### D. Meta Learning

Meta-learning, or learning to learn, is a strategy to speed up the learning process from a small amount of data. Model-Agnostic Meta-Learning (MAML) [34] explores this strategy across different models and problem settings, such as few-shot image classification and few-shot regression. For voice cloning, It identifies an effective parameter initialization for accelerating speaker adaptation. It utilizes an outer loop to determine a meta-initialization and an inner loop to quickly adapt to a new task using minimal data. Language-Agnostic Meta-Learning (LAML) treats languages as independent tasks while building multilingual voice cloning systems [12].

## IV. VOICE CLONING

Voice cloning is a complex problem, as improving its performance requires tackling different aspects within TTS systems, such as speech naturalness, quality, expressiveness, and speaker similarity. Typically, these modifications occur in acoustic modeling and, in some instances, in the vocoder. We will explore the contributions, following Fig. 3, as understanding these methodologies will pave the way toward enhancing voice cloning and understanding their limitations.

### A. Speaker Adaptation

Speaker Adaptation relies on fine-tuning a TTS model to replicate the voice of a certain speaker. This process is one of the main principles behind voice cloning. The fine-tuning process, summarized in Fig. 4, can happen on the speaker encoder, a model component, or the entire model.

**Disentanglement:** Representation disentanglement is an active area of research exploring decomposing the TTS system's elements to achieve better control of voice characteristics. Daft-Exprt [35] introduced an acoustic model for cross-speaker prosody transfer based on FastSpeech 2. This model uses FiLM conditioning layers [36] to inject prosodic information across the model, encoding high-level and low-level prosodic information (pitch, loudness, duration). This model includes an adversarial training strategy through a

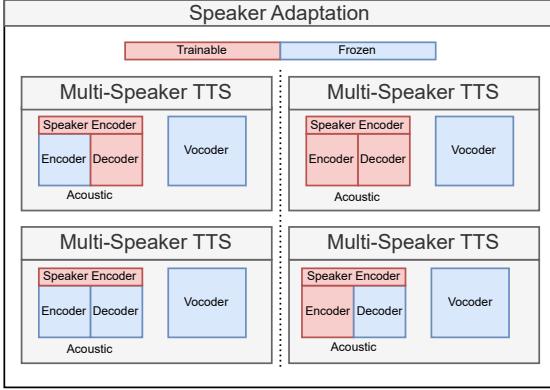


Fig. 4. Four variants for speaker adaptation of a Multi-Speaker TTS.

speaker classifier for speaker-prosody disentanglement. Ellinas et al. [37] proposed a phoneme-level prosody control for TTS systems using Tacotron as its base acoustic model. They introduce discrete labels, achieved through an unsupervised prosodic clustering process, to discretize prosodic features such as the fundamental frequency ( $f_0$ ) and phoneme duration. These features will be fed into a prosody encoder to obtain prosodic representation. Additionally, they use an adversarial speaker classifier for phoneme-speaker disentanglement. Du et al. [38] presented TN-VQ-TTS, a novel TTS model that leverages timbre-normalized vector-quantized (TN-VQ) acoustic features for style-timbre disentanglement in speaker adaptation with limited data. This architecture consists of two main components: a txt2vec module that predicts TN-VQ features from input phonemes and a vec2wav module that uses those features, auxiliary prosodic features, and a speaker embedding from a conditioned speaker encoder to generate a speaker-specific waveform. TN-VQ features are derived by performing dimensionality reduction of the VQ features via PCA, followed by timbre normalization. The resulting features are then quantized using K-means clustering for integration into a multi-speaker TTS model. For speaker adaptation, they use one of two approaches: fine-tuning free adaptation, where they use a pre-trained style embedding lookup table (LUT) as a codebook and choose the most suitable embedding, or weight fine-tuning adaptation, where it performs a weighted summation of all the embeddings within the table.

**Speaker Representation and Verification:** Speaker adaptation has been evolving over the years. Thus, researchers explore speaker representation and speaker-related tasks such as speaker verification to improve speaker similarity. Fan et al. [39] proposed a multi-speaker DNN that shares parameters across various speakers with a speaker-dependent regression layer. A new speaker is adapted by fixing the shared parameters and updating the regression layer. Zhao et al. [40] proposed a unified framework for speaker adaptation that includes a bidirectional Long Short-Term memory with a Recurrent Neural Network (Bi-LSTM-RNN) acoustic model. They explored speaker identity representations, such as i-

vector and speaker code, at the input layer of the framework to observe the effectiveness of different speaker representations. Doddipatla et al. [41] explored the usage of speaker identification vectors, called d-vectors instead of the previously used i-vectors, with linguistic features to train a multi-speaker DNN-based TTS model. More recent work started exploring the usage of speaker verification to enhance speaker adaptation performance. Cai et al. [42] proposed a method for improved knowledge transfer from a speaker verification network to a TTS framework, using a feedback constraint mechanism to facilitate voice cloning training. Their approach integrates a Tacotron-based model and a learnable dictionary encoding (LDE)-based verification model [43] to enhance speaker similarity for unseen speakers. Ruggiero et al. [44] proposed a multi-speaker TTS system consisting of three-component: a speaker encoder, a synthesizer, and a vocoder, following the general architecture of Jia et al. [45]. They explore various architectures for the speaker encoder and a novel transfer-learning technique that conditions the synthesizer on utterance embeddings instead of speaker embeddings. This approach enables the system to generalize to unseen speakers using only a short reference utterance. Nakai et al. [14] extends on GANSpeech [46] and FastSpeech 2 by introducing a novel multi-task adversarial training algorithm through discerning real from fake synthesized speech and performing speaker verification to ensure whether the speaker exists or is non-existent. Additionally, they introduce Adversarially Constrained Autoencoder Interpolation (ACAI), a regularization term to improve performance for unseen speakers. Chen et al. [47] tackled voice cloning by proposing a neural fusion architecture to include unit concatenation within a parametric TTS. This fusion architecture comprises a text encoder, acoustic decoder, and phoneme-level reference encoder. This phoneme-level encoder extracts the speakers' prosody to improve speaker similarity and uses auto-regressive distribution modeling and decoder refinement to tackle concatenation discontinuity. Wadoux et al. [48] explored the impact of phonetic content and sample duration on voice cloning. The system consists of a multi-speaker Tacotron 2 model and an x-vector speaker encoder. This exploration led to inconclusive results since the impact is shown in the speaker representation but not the generated waveform; these results showed improved speech quality and speaker similarity in very specific cases.

**Handling Speech Variability:** Speaker adaptation can face difficulties handling speech variability from different speaking styles, accents, and acoustic conditions. GMVAE-Tacotron [49] focused on speech variabilities, such as controlling speech attributes such as speaking style, accent, noise, and acoustic conditions. It introduced the VAE-based hierarchical generative TTS model, improving attribute control compared to the Global Style token (GST)-based approaches [50], [51]. They explored synthesizing speech for unseen speakers in clean and noisy environments from an unseen reference utterance. Similarly, Neekhara et al. [52] tackled the challenge of noisy speech in multi-speaker public datasets by proposing a transfer-learning guideline for adapting a single-speaker TTS model for a new speaker. More specifically, two fine-tuning methods are used to reduce the time and data required to adapt

the TTS model to a new voice. AdaSpeech [53] introduced a novel architecture, based on FastSpeech 2, for efficient and high-quality speaker adaptation in TTS systems. It models acoustic information at both the utterance and phoneme levels to handle acoustic variety and enable better generalization across various acoustic conditions. Additionally, it employs conditional layer normalization (CLN) in the mel-spectrogram decoder, improving the efficiency during speaker adaptation by only fine-tuning the CLN parameters and speaker embeddings during speaker adaptation. AdaSpeech 3 [54] builds upon AdaSpeech but focuses on synthesizing spontaneous speech, similar to podcasts and conversation, instead of reading-style speech. It utilizes the already available reading-style TTS and fine-tunes it for spontaneous TTS. Adapting to spontaneous TTS includes inserting filled pauses (FP) through an FP predictor, a mixture of experts (MoE)-based duration predictor, and a pitch predictor for rhythm adaptation. Finally, the authors fine-tune specific decoder parameters for target timbre speaker adaptation. Mandeel et al. [11] concentrated on improving prosody control, especially in interrogative sentences. This work explores speaker adaptation using FastSpeech 2 with a HiFi-GAN neural vocoder, in which they demonstrated their performance on two HiFi-based adaptation datasets [55] containing interrogative and declarative sentences.

**Untranscribed Speech:** The abundance of untranscribed speech online can be a useful tool for improving voice cloning systems. Inoue et al. [56] proposed a semi-supervised speaker adaptation method utilizing a pre-trained ASR system to obtain text data from available untranscribed speech recordings and then use it to train a TTS model. This method can simplify speaker adaptation through an end-to-end ASR/TTS system. Zhang et al. [57] introduced a novel unsupervised pre-training mechanism that uses VQ-VAE [58] to extract linguistic units directly from untranscribed speech and then use <linguistic unit, speech> pairs to pre-train Tacotron. For speaker adaptation, we use a few <text, audio> pairs for fine-tuning to target speakers. AdaSpeech 2 [59] is an adaptive TTS that uses untranscribed speech for speaker adaptation. It follows the basic structure of AdaSpeech and utilizes a mel-spectrogram encoder for speech reconstruction, constrained by an L2 loss with the phoneme encoder. For speaker adaptation, untranscribed speech is used to fine-tune the TTS decoder. Klapsas et al. [60] explored BYOL [61], a self-supervised framework for obtaining speech representation. Then, they extended the augmentations typically applied through training to capture the speakers' identity and ensure the features are noise-resistant. These pre-trained weights are used as utterance-level embeddings and inputs to a Non-Attentive Tacotron-based architecture [62] to be used for unseen speaker adaptation.

**Parameter Efficiency:** Personalized TTS focuses on reducing TTS models' complexity to ensure usability in less computational devices. Inoue et al. [63] investigated the roles of various modules in Transformer-TTS [64] during fine-tuning to understand its effectiveness in handling speech deletion and/or repetition. This includes freezing specific modules, including character embedding, encoder, probability output layer, and choosing to include Binary Cross Entropy (BCE) loss in the fine-tuning process. The authors concluded that fine-tuning

the TTS model while freezing character embeddings and the probability output layer reduces speech repetition and deletion. AdaVITS [65] is a lightweight VITS-based TTS for low-resource speaker adaptation. They replace the upsampling-based decoder with an Inverse short-time Fourier transform (iSTFT)-based decoder. They also introduce NanoFlow to reduce the prior encoder parameters and replace scaled-dot attention with linear attention to reduce the text encoder complexity. Phonetic posterogram (PPG) is used as the frame-level linguistic feature for supervision as the model becomes less stable. HyperTTS [66] proposed a hypernetwork-based approach for TTS speaker adaptation, integrating a small learnable network into an adapted FastSpeech 2 backbone. Instead of fine-tuning the entire model, the hypernetwork dynamically generates adapter parameters conditioned on speaker embeddings derived from a speaker verification model trained with generalized end-to-end loss (GE2E) [67].

**Non-English TTS:** Voice cloning works mainly focus on English speaker adaptation, even though there are many languages worldwide. VStyclone [68], a real-time Mandarin voice cloning algorithm, consists of three main components: an efficient tone extractor, a transformer-based style synthesizer with an embedded style extraction module, and a GAN-based vocoder. The tone extractor uses self-attention to capture speaker-specific information efficiently instead of GE2E loss optimized with a modified softmax loss. Cheng et al. [69] proposed using prosodic features for Mandarin speaker adaptation. These features are obtained at the utterance/speaker level and then integrated into existing speaker-encoding or speaker-embedding adaptation frameworks such as Tacotron 2.

### B. Few-shot Voice Cloning

Few-Shot TTS (FS-TTS), namely Few-Shot Voice Cloning, focuses on speaker adaptation through fine-tuning with minimal data, which is crucial in tackling practical voice cloning cases. Fig. 5 shows some few-shot TTS algorithms.

**Disentanglement:** Researchers actively study methods for obtaining better control of speech elements, including content-speaker disentanglement, to improve few-shot learning performance. Wang et al. [70] proposed a few-shot speaker adaptation method through spoken content and voice factorization to tackle overfitting due to the complexity of acoustic models. They decompose the acoustic model into a spoken content predictor that predicts PPG to constrain the acoustic model space. Then, we use it as an input to a voice conversion module to extract the acoustic features. Additionally, they proposed speaker adaptation strategies for data with or without transcription. Attentron [71] is a Tacotron2-based novel architecture for few-shot voice cloning, where it utilizes two encoders to improve unseen speaker similarity and speech quality. This architecture includes a fine-grained encoder that extracts style information from multiple reference audio files, and the coarse-grained encoder generates global embeddings of the target speaker, which helps stabilize the output. It follows a warm-start training method, similar to [72], to reduce training time. Lu et al. [73] proposed an end-to-end unified model for voice cloning and conversion as they share a similar

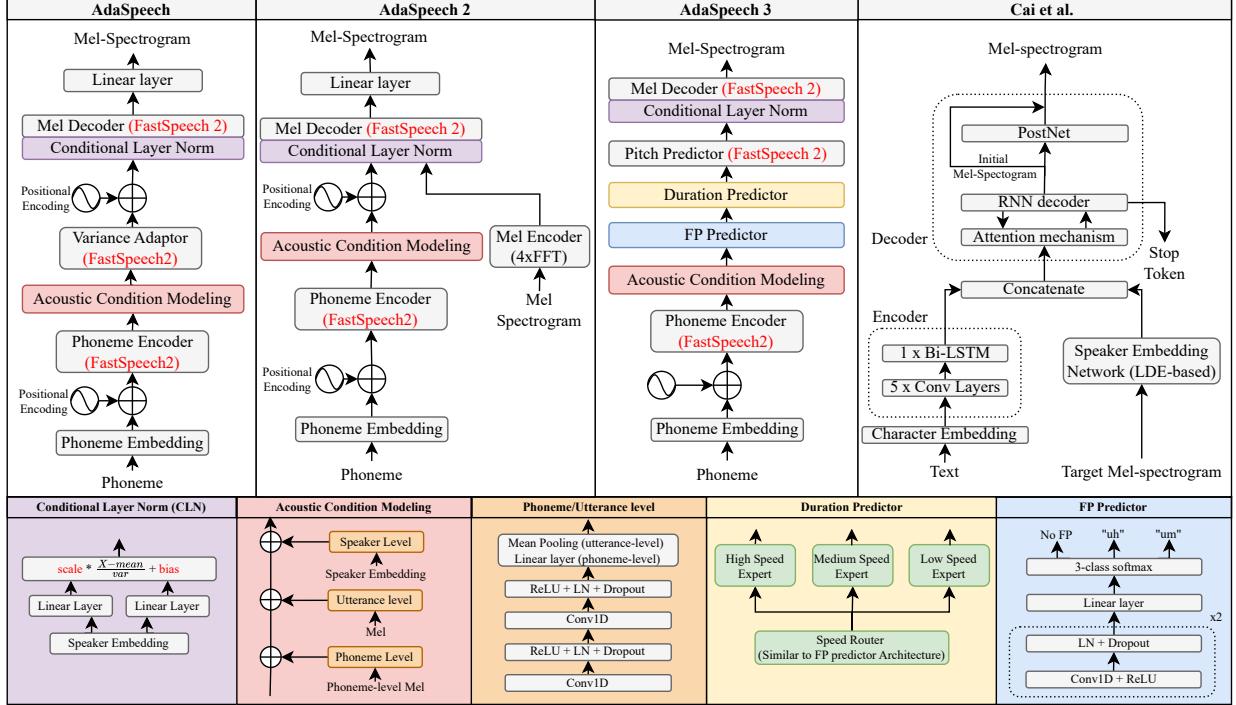


Fig. 5. Few-Shot TTS algorithms, adapted from AdaSpeech [53], AdaSpeech 2 [59], AdaSpeech 3 [54], and Cai et al. [42].

information flow. The authors proposed extending VAE-based speech disentanglement [74], [75] by introducing learnable text-aware prior, enhancing content-speaker disentanglement by providing more accurate and speaker-independent content representation than the initially used fixed Gaussian prior.

**Untranscribed Speech:** Few-shot TTS systems require text-speech pairs to perform fine-tuning to clone a speaker's voice. Researchers propose voice cloning algorithms without requiring transcribed speech. Nautilus [7] is a multi-speaker TTS/VC system with the ability of speaker adaptation using untranscribed speech through a learned linguistic latent embedding (LLE). The process of voice cloning inference without transcribed speech follows these steps: they first adapt the speech decoder and vocoder to the target's voice speech. Then, both are welded together by joint fine-tuning and, finally, inference. Similarly, Sadekova et al. [76] built a multimodal TTS/VC system capable of performing voice cloning and any-to-any voice conversion using untranscribed data. This system consists of three modules: a text encoder, a mel decoder, and a diffusion-based shared decoder. Zhang et al. [77] also proposed a multimodal learning approach capable of both TTS and VC. They extend Tacotron 2 by introducing an unsupervised speech representation module, VQ-VAE, to extract linguistic units given speech input. Then, a Tacotron 2-based model generates a mel spectrogram conditioned on these unsupervised units for VC or phonemes for TTS to generate a waveform through a vocoder. On the contrary, other works focus on uni-modal TTS algorithms, HierSpeech [78] builds on top of VITS by employing self-supervised learning to reduce the text-speech information gap. To improve linguistic

feature extraction, they utilize XLS-R [79], a pre-trained Wav2Vec 2.0 on large-scale cross-lingual speech dataset. They further extend this architecture by introducing HierSpeech-U, which performs speaker adaptation with untranscribed speech. Guided-TTS 2 [80] builds on top of Guided-TTS [81] and enables adaptive TTS using untranscribed data. Guided-TTS proposed a diffusion-based TTS model that uses an unconditional Denoising Diffusion Probabilistic Model (DDPM) [82], phoneme classifier, duration predictor, and a speaker encoder to remove the reliance on text-speech alignment by using untranscribed data. Additionally, it introduces a norm-based scaling method to improve pronunciation. Then, Guided-TTS 2 extends this work by introducing a speaker-conditional diffusion model to enable few-shot adaptation and zero-shot TTS for target speakers. This model fine-tunes the pre-trained speaker conditional diffusion model. Due to the difficulty of training its unconditional generative model, Kim et al. [83] developed a method to fine-tune an adaptive diffusion-based TTS model, UnitSpeech, with limited untranscribed data. It utilizes a unit encoder that generates a discretized representation from untranscribed speech using HuBERT [28]. The output representation is then used for speaker adaptation, improving speech quality and speaker similarity.

**Parameter Efficiency:** Personalized TTS involves researchers' contributions toward reducing models' sizes and improving their speed. For example, Voiceloop [84] proposed a unique approach that deals with samples in the wild without requiring phoneme alignment or linguistic features. This work is based on a novel shifting buffer working memory that estimates the attention, computes output, and updates the

buffer. This architecture can fit new speakers post-training by only updating the speaker embeddings, and it can be run almost in real-time on a CPU. Similarly, Kons et al. [85] proposed a lightweight TTS model that consists of three modules: prosody prediction, acoustic feature prediction, and an LPCNet vocoder. This system can run faster than real-time on a standard CPU. Morioka et al. [86] proposed a parameter-efficient few-shot speaker adaptation method. It uses a PnG NAT model [62], [87] as its backbone model and then includes trainable lightweight modules called residual adapters for each target speaker while keeping the backbone architecture frozen. Similarly, Hsieh et al. [88] proposed parameter-efficient adapter modules, using FastPitch [89] as a baseline TTS model through only fine-tuning those adapters to a target speaker during speaker adaptation. Huang et al. [13] proposed trainable structured pruning for voice cloning, utilizing FastSpeech 2 as its backbone, to compress the voice cloning model up to 7 times while maintaining its performance. Chen et al. [90] introduced diffusion-based backbones for adaptive TTS called Diffusion Transformer (DiT). They utilize adaptive layer normalization (adaLN), initially introduced for image generation [91], and made it compatible by receiving a sequence as a condition instead of class embeddings. In addition, they use a generator-based diffusion method to enable faster and higher-quality speech synthesis, demonstrating superior performance compared to transformer-based solutions. Recently, Wang et al. [92] proposed a universal speaker-adaptive TTS (USAT) that unifies few-shot and zero-shot strategies. The framework introduces a timbre converter consisting of two discriminators and a Memory-Augmented VAE (MA-VAE) to generate a timbre-invariant phoneme representation to enable disentanglement during pre-training. USAT demonstrates high parameter efficiency as it requires adapting fewer parameters while maintaining FS-TTS performance.

**Speaker Representation:** Improving speaker representation extraction is crucial to enhance the overall performance of TTS systems, specifically speaker similarity, for unseen speaker adaptation. Deng et al. [93] explored the multi-speaker latent space to enhance few-shot speaker adaptation using just a few minutes of data or improving voice quality by leveraging data from multiple speakers. GC-TTS [94] proposed a Tacotron 2 extension by introducing a neural speaker encoder and a classification layer. The authors utilized two geometric constraints, initially proposed in [95] for natural images, to improve speaker representation. This system employs a multi-stage training and fine-tuning approach while exploring freezing-specific modules for a more efficient fine-tuning process. Lee et al. [96] proposed a novel few-shot adaptation algorithm that utilizes a speaker embedding LUT for training a multi-speaker Tacotron 2 TTS system. They introduce a GAN-based initial embedding predictor trained on speaker embeddings using adversarial training. This approach speeds up the adaptation process by generating the initial embeddings suitable for unseen speaker adaptation. Voice Filter [97] takes another approach by shifting the voice speaker adaptation task into two modules: a high-quality TTS system followed by a voice conversion system as a post-processing step. This split leads to requiring less amount of data for speaker adaptation by

limiting the problem's complexity. Xue et al. [98] proposed a three-module separately trained TTS architecture consisting of an ECAPA-TDNN speaker encoder [99], FastSpeech 2 acoustic model, and a HiFi-GAN vocoder. Similarly, TDNN-VITS [100] adopts the same speaker encoder for the VITS model. It introduces a multi-angle feature fusion approach, which improves speech quality with less data rather than simply concatenating speaker embeddings after its text embeddings.

**Optimization-based:** Improving FS-TTS performance is achievable through developing novel architectures or even components to enhance its performance. Additionally, exploring optimization techniques such as meta-learning and Bayesian optimization is an exciting approach to enhance performance with minor architectural modifications. Chen et al. [101] introduced a meta-learning approach applied to a multi-speaker model with a shared conditional WaveNet core. Their training mechanism focuses on having two sets of parameters: task-dependent and independent. During training, they train all but discard the dependent ones for deployment; this leads to only optimizing for the task-dependent parameters for faster adaptation to new tasks. Moss et al. [102] proposed Bayesian Optimization for Fine Tuning Neural Text to Speech (BOFFIN TTS), in which achieving convincing results requires an adaptation strategy. This strategy includes fine-tuning adaptation hyper-parameters and adding two new ones, considering the target speaker's acoustic, phonetic, and corpus properties to achieve optimal results. Min et al. [103] proposed StyleSpeech, a FastSpeech2-based model capable of synthesizing speech in the style of a target speaker using a single reference speech sample. The authors introduce style-adaptive layer normalization (SALN), which aligns the gain and bias of the input text features based on the style vector. Building on this, they introduce Meta-StyleSpeech by incorporating style prototypes and episodic meta-learning to enhance the model's performance to unseen speakers. Unlike previous works, Meta-Voice [104] explores MAML and meta-transfer for faster few-shot speaker adaptation. This architecture, based on FastSpeech 2, consists of four modules: a text encoder, a variance adapter, a mel decoder, and a style encoder. They also use and extend SALN to handle multiple speakers and prosodies. In addition, it uses domain adversarial training for speaker-prosody disentanglement, which is useful for cross-speaker style transfer. Similarly, Meta-TTS [12] proposed applying MAML to speed up few-shot speaker adaptation using FastSpeech 2 as its base TTS architecture. The authors achieve this by designing a meta-task, where each task randomly chooses a speaker and takes ten utterances from that speaker, 5 for the support set and 5 for the query. This training consists of an inner loop and an outer loop, where the inner loop optimizes specific modules using multiple gradient descents (inner loop update), and the outer loop uses second-order gradient computation for the loss (Meta update gradient).

**Non-English FS-TTS:** Mandarin is the second language with voice cloning publications. Cong et al. [9] performed speech-noise disentanglement to improve performance with noisy audio target samples by introducing domain adversarial training. Zheng et al. [105] used a Tacotron2-based architecture and proposed various extensions, such as a novel

TABLE IV

COMPARISON OF FEW-SHOT TTS ALGORITHMS COVERING NATURALNESS (NAT), SPEECH QUALITY (QUAL), AND SPEAKER SIMILARITY (SIM). FOR MULTIPLE DATASETS, RESULTS ARE AVERAGED. PAPERS ARE ARRANGED IN ASCENDING ORDER BY PUBLICATION DATE.

Paper	Test set, additional details	TTS Evaluation
Naturalness and Speech Quality		
Attentron [71]	VCTK (10 utt/spk, 8 spks)	Nat (NMOS) ↑: Attentron (3.97) > Cooper et al. [72] (3.91) > GMVAE-Tacotron (3.88) Qual (MCD) ↓: Attentron (11.67) > GMVAE-Tacotron (13.11) > Cooper et al. (13.50)
Meta-StyleSpeech (MSS) [103]	VCTK (1 utt/spk, 108 spks)	Nat (NMOS) ↑: MSS (3.82) > SS (3.77) > MS-FS2 (d-vector) (3.74) > MS-FS2 (3.69) > GMVAE [49] (3.15) Qual (MCD) ↓: MSS (4.95) > MS-FS2 (4.97) > SS (5.01) > MS-FS2 (d-vector) (5.03) > GMVAE (5.54) Qual (WER) ↓: MSS (16.79) > MS-FS2 (17.35) > SS (17.51) > MS-FS2 (d-vector) (17.55) > GMVAE (23.86)
GC-TTS [94]	VCTK (80 utt, 16 spks) [Subjective]; VCTK (800 utt, 16 spks) [Objective] *5 minutes fine-tuning	Nat (NMOS) ↑: Cooper et al. [72] (3.86) > Tacotron 2 (3.84) > Jia et al. [45] (3.83) > GC-TTS* (3.81) > Tacotron 2* (3.74) Qual (MCD) ↓: Tacotron 2 (4.87) > GC-TTS* (5.35) > Tacotron 2* (5.38) > Cooper et al. (5.77) > Cooper et al. (5.82)
Meta-TTS [12]	LibriTTS (1 utt/spk, 30 speakers); VCTK (1 utt/spk, 80 spks)	Nat (NMOS) ↑: Meta-TTS (3.20) > MS-FS2 (3.16) [emb table] — MS-FS2 (3.72) > Meta-TTS (2.70) [Shared emb]
Sadekova et al. [76]	VCTK (5 utt/spk, 25 spks)	Nat (NMOS) ↑: Tacotron-SMA [121] (4.23) > Proposed (4.18) > FS (3.98) > Grad-TTS (3.94) > SS (3.77)
UnitSpeech [83]	LibriTTS test-clean (5 utt/spk, 10 spks)	Nat (NMOS) ↑: GTTS2 [80] (4.16) > UnitSpeech (4.13) > YTTS (3.57) Qual (CER) ↓: GTTS2 (0.84) > UnitSpeech (1.75) > YTTS (2.38)
TDNN-VITS [100]	VCTK (300 utt, 10 spk)	Nat (NMOS) ↑: TDNN-VITS (3.72) > FS2 (3.61) > Tacotron 2 (3.54)
VAE-TP [73]	LibriTTS test set (50 utt/spk, 11 spks)	Nat (NMOS) ↑: CDFSE (3.82) > VAE-TP (3.81) > SC-GlowTTS (3.56) Qual (CER) ↓: CDFSE (1.96) > VAE-TP (2.82) > SC-GlowTTS (6.99)
USAT [92]	ESLTTS (10 utt/spk, 30 speakers) — *VITS is fully fine-tuned	Nat (NMOS) ↑: USAT (3.84) = UnitSpeech (3.84) > VITS (3.80) Qual (WER) ↓: UnitSpeech (10.30) > USAT (11.00) > VITS (14.60)
Speaker Similarity		
Attentron [71]	VCTK (10 utt/spk, 8 spks)	Sim (SMOS) ↑: Attentron (3.57) > GMVAE (3.27) > Cooper et al. (3.17) Sim (SECS) ↑: Attentron (0.79) > Cooper et al. (0.71) > GMVAE (0.70)
Meta-StyleSpeech (MSS) [103]	VCTK (1 utt/spk, 108 spks)	Sim (SMOS) ↑: MSS (3.81) > SS (3.46) > MS-FS2 (3.36) > GMVAE (3.11) > MS-FS2 (d-vector) (2.12) Sim (SECS) ↑: MSS (0.82) > SS (0.80) > MS-FS2 (0.77) = GMVAE (0.77) > MS-FS2 (d-vector) (0.62)
GC-TTS [94]	VCTK (80 utt, 16 spks) [Subjective]; VCTK (800 utt, 16 spks) [Objective] *5 minutes fine-tuning	Sim (SMOS) ↑: Tacotron 2 (3.60) > GC-TTS* (3.45) > Tacotron 2* (3.34) > Cooper et al. (3.19) > Jia et al. (3.15) Sim (SECS) ↑: Tacotron 2 (0.89) > GC-TTS* (0.84) > Tacotron 2* (0.82) > Cooper et al. (0.80) > Jia et al. (0.78)
Meta-TTS [12]	LibriTTS (1 utt/spk, 30 speakers); VCTK (1 utt/spk, 80 spks)	Sim (SMOS) ↑: Meta-TTS (2.96) > MS-FS2 (1.55) [emb table] — Meta-TTS (3.06) > MS-FS2 (1.33) [Shared emb]
Sadekova et al. [76]	VCTK (5 utt/spk, 25 spks)	Sim (SMOS) ↑: Tacotron-SMA (4.31) > Grad-TTS (4.19) > Proposed (4.15) > FS (4.02) > SS (3.96)
UnitSpeech [83]	LibriTTS test-clean (5 utt/spk, 10 spks)	Sim (SMOS) ↑: GTTS2 (3.90) = UnitSpeech (3.90) > YTTS (3.34) Sim (SECS) ↑: GTTS2 (0.94) = UnitSpeech (0.94) > YTTS (0.87)
TDNN-VITS [100]	VCTK (300 utt, 10 spk)	Sim (SMOS) ↑: TDNN-VITS (3.85) > FS2 (3.73) > Tacotron 2 (3.65)
VAE-TP [73]	LibriTTS test set (50 utt/spk, 11 spks)	Sim (SMOS) ↑: VAE-TP (3.66) > CDFSE (3.60) > SC-GlowTTS (3.38) Sim (SECS) ↑: VAE-TP (0.78) > CDFSE (0.76) > SC-GlowTTS (0.72)
USAT [92]	ESLTTS (10 utt/spk, 30 speakers) — *VITS is fully fine-tuned	Sim (SMOS) ↑: UnitSpeech (3.75) > USAT (3.74) > VITS (3.56) Sim (SECS) ↑: UnitSpeech (0.83) = USAT (0.83) > VITS (0.80)

speaker encoder with a random sampling mechanism. They also used a speaker classifier, and text encoder adversarial training for content-speaker disentanglement. Finally, they also proposed multi-alignment guided attention for text-audio pairs for model training stability. Wang et al. [106] proposed a speaker adaptation framework that disentangles prosody and voice characteristics. They achieved this by introducing a prosody control attention mechanism and an LPCNet vocoder to ensure computational efficiency. In addition, they proposed a pruning strategy to remove mismatching text-audio pairs, ensuring a more stable system. Fu et al. [107] proposed separate speaker and prosody modeling based on a reference audio and PPG using a multi-head attention process, allowing more control of prosody duration and speaker timbre. In addition, the decoder is factored into an average-net decoder to control prosody by generating speaker-independent acoustic features and an adaptation-net decoder to generate acoustic features conditioned on speaker embeddings for speaker timbre imitation. Dian [108] is a speech synthesis system consisting of an attention-free acoustic model with a duration model. The attention was replaced by inputting the duration directly into the acoustic model, eliminating the skipping and repeating challenge and improving speech intelligibility. Chien et al. [109] investigated various speaker representations, including pre-trained and learnable ones. This study concluded that speaker embeddings pre-trained through voice conversion achieved the highest performance compared to d/x-vectors. Additionally, learnable speaker representations, such as speaker embedding LUT and GST, can improve speaker similarity. He et al. [110] proposed AdaptiveFormer, a FastSpeech2-based FS-TTS

model, with two contributions: proposing a novel adaptive attention mechanism that employs speaker embeddings, modified calculations for Q, K and V, and CLN compared of the traditional multi-head self-attention component. They also introduce a conditional adapter to predict rhythm information (duration, pitch, energy). UNet-TTS [111] is a one-shot TTS algorithm replicating unseen speakers’ voices and styles. It adopts a UNet-based architecture, initially proposed for image segmentation [112], to efficiently extract hierarchical speaker and utterance level features from a reference audio. In addition, the authors introduced a content loss after the style encoder to enhance content-style disentanglement. Unlike Tacotron baselines using GST or speaker embedding obtained from a speaker verification task, this method enables fine-grained prediction of acoustic features and better performance on an unseen emotional corpus. MSDTRON [113] proposed a multi-speaker TTS system to tackle speaking style diversity. This paper investigated effective approaches toward making use of characteristic information. This includes explicitly characterizing the harmonic structure of speech through an excitation spectrogram and a conditional gated LSTM (CGLSTM) to improve voice control without affecting speech quality. LightClone [114] focused on speeding up speaker adaptation and inference time by introducing a speaker-guided subnet selection process. They use FastSpeech 2 as the baseline model, and during adaptation, only one subnet of the source model is selected for the target speaker. Unlike previous work, AdaVocoder [115] approached voice cloning by proposing a custom voice vocoder with a cross-domain consistency loss to tackle overfitting in few-shot voice cloning.

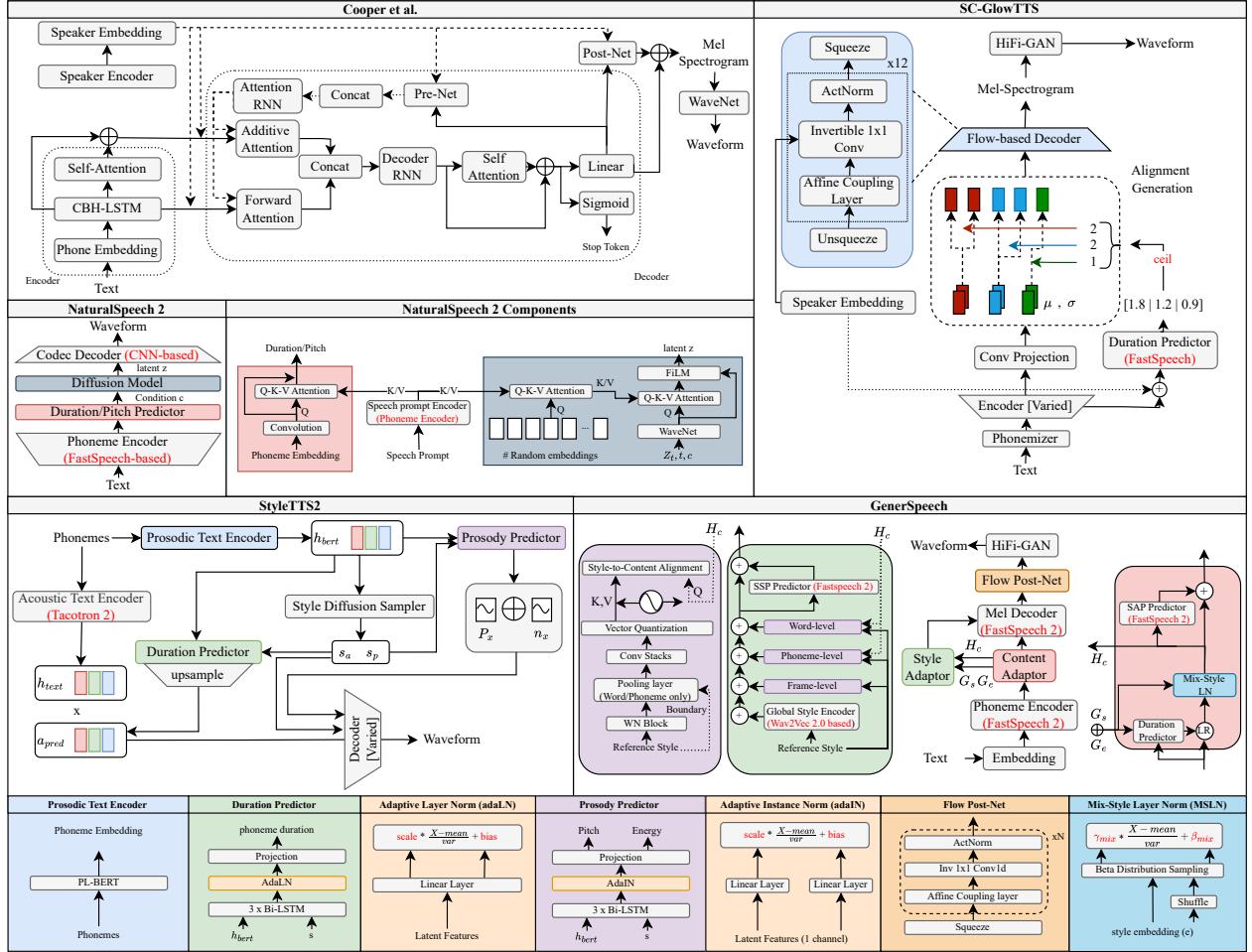


Fig. 6. Zero-shot TTS algorithms, adapted from SC-GlowTTS [155], StyleTTS 2 [147], NaturalSpeech 2 [202], GenerSpeech [122], and Cooper et al. [72]

**Summary:** Few-shot TTS algorithms have been rapidly evolving, where research focuses on novel components, architectures, or training strategies to enhance the overall performance. Table IV shows a performance comparison of few-shot TTS algorithms in naturalness, speech quality, and speaker similarity. Word Error Rate (WER) and Character Error Rate (CER) are objective metrics useful to ascertain speech quality by testing it using a pre-trained Automatic Speech Recognition (ASR) system. This system can vary, such as using Whisper ASR [116], CTC-based conformer [117] available in Nemo toolkit [118]. The same goes for Speaker Embeddings Cosine Similarity (SECS), as the speaker encoder used for reporting results varies. This includes X-vector [119], GE2E [67] obtained through the Resemblyzer package, and TitaNet-L [120].

### C. Zero-shot Voice Cloning

Zero-shot TTS (ZS-TTS) algorithms aim to generate speech that closely resembles a target speaker without fine-tuning the TTS model. Unlike speaker adaptation, ZS-TTS architectures require the inclusion of specialized modules, such as a speaker encoder, to capture and replicate the target speaker's voice during inference. Fig. 6 illustrates some ZS-TTS algorithms.

**Disentanglement:** Representation disentanglement is an exciting approach as it provides more control over the target voice's attributes. GenerSpeech [122] decomposes style-agnostic (content) from style-specific (speaker identity, emotion, prosody) information while papers [123]–[127] focus on timbre-content disentanglement. GenerSpeech is a FastSpeech2-based TTS model that tackles disentanglement by introducing multiple modules, such as a multi-level style adapter with a Wav2Vec 2.0 encoder. In addition, it uses a mix-style layer normalization (MSLN) to eliminate style from content representation. GZS-TV [123] proposed a generalizable zero-shot TTS and VC model that uses a disentangled representation learning mechanism to separate phoneme from timbre embeddings to reduce phoneme leakage into speaker embeddings. The authors built an ECAPA-TDNN-based speaker encoder [99] and leveraged VAE to enhance its performance. UTTS [124] is an unsupervised TTS acoustic model training scheme that does not require text-audio pairs. They employ a conditional disentangled sequential VAE (C-DSVAE) as the model's backbone for timbre-content disentanglement. Jeon et al. [125] proposed a FastSpeech-based zero-shot TTS algorithm that utilizes a negation feature

learning paradigm that decouples speaker information from linguistic content by subtracting content representations from the audio embedding. This approach reduces content leakage into speaker representation, leading to more accurate speaker identity representation. They also introduce multi-stream transformers to improve robustness to unseen speakers. Similarly, MRMI-TTS [126] is a Fastspeech2-based framework for zero-shot TTS that reduces content leakage in speaker embeddings using a multi-reference audio approach. It employs a speaker-content disentangling module (SCDM) comprising a separate content and speaker encoder. The multi-reference audio selection is based on Jaccard similarity to ensure that the content is varied and that mutual information minimizes the entanglement between speaker and content embeddings. Choi et al. [127] proposed a novel speaker conditioning technique using a variable length embedding sequence and enhancing an affine coupling function for flow-based ZS-TTS architectures. This method builds upon the VITS model as its base architecture. They tackle timbre-content disentanglement by using different references and target speech from the speaker. Unlike previous methods, Ren et al. [128] proposed TiCodec, a neural speech codec that enhances efficiency by separating time-variant and invariant speech features. By encoding time-invariant features separately, it reduces the number of tokens needed while maintaining performance in ZS-TTS.

Other researchers take a different approach by emphasizing finer control over speech synthesis. Paul et al. [129] proposed a multi-speaker and multi-style TTS framework capable of generating speech with the speaker’s characteristics given a reference signal. They proposed a novel Renyi Divergence-based disentangled representation framework to disentangle content, style, and speaker information, ensuring reduced content and style leakage. Kumar et al. [130] proposed a novel zero-shot TTS approach that uses a normalization architecture with a NAR transformer-based architecture. They proposed two normalization techniques to improve learning prosodic properties, such as energy and f0, in a disentangled manner. Fujita et al. [131] proposed a zero-shot TTS model conditioned on speech embeddings derived from a self-supervised learning (SSL) model. They also introduced the separate conditioning of acoustic features and the phoneme duration predictor, effectively disentangling rhythm and acoustic features. Lee et al. [132] disentangled timbre (intra-speaker consistent component) from cadence (inter-utterance variate component) through a hierarchical structure to model each independently, and then used variance-invariance covariance regularization (VICReg) on cadence embeddings and a speaker ID loss to ensure stable timbre embeddings for the speaker. HierSpeech++ [133] is a fully parallel zero-shot TTS and VC framework built on top of HierSpeech [78]. They proposed multiple improvements, such as the Text-to-Vec (TTV) module that generates semantic and f0 representation from text and prosody prompts. It also introduced an efficient speech super-resolution (SpeechSR) component for efficient up-sampling of high-resolution audio waveforms. HierSpeech++ enables fine-grained control over speech synthesis by disentangling speaker-related and speaker-agnostic representation, including an encoder with f0 control for manual control of pitch contour.

Mega-TTS [134] is a novel large-scale zero-shot TTS system that enhances speech generation by considering intrinsic speech properties. It achieves this by disentangling speech into four components: content, timbre, prosody, and phase, using mel spectrogram as the intermediate representation instead of codec-based features. The model employs a timbre encoder as a global representation since timbre changes slowly over time, while a VQGAN-based acoustic model generates a spectrogram and a prosody large language model (P-LLM) to capture prosody distribution, effectively modeling both local and long-range dependencies. Mega-TTS 2 [135] builds on top of its previous work by proposing a multi-sentence prompting method to extract diverse speaker information. The authors designed a compressive acoustic autoencoder for timbre-prosody disentanglement, achieving more control while maintaining the speaker’s identity. They also introduced a multi-reference timbre encoder (MRTE) and a P-LLM for fine-grained timbre information extraction from multi-reference prompts.

Codec-based TTS systems use a hierarchical approach to obtain semantic tokens followed by acoustic tokens. Even though they are not fully independent, they aim to decompose speech into semantic (linguistic and para-linguistic) and acoustic (non-linguistic, e.g., speaker identity) information. Papers [136]–[141] are codec-based TTS algorithms that disentangle semantic and acoustic features. Kharitonov et al. [136] introduced SPEAR-TTS, a multi-speaker TTS system that formulates speech synthesis as two seq2seq tasks: converting text into high-level semantic tokens (Reading) and then mapping semantic tokens into low-level acoustic tokens (Speaking). This decomposition enables training the reading stage with text-audio pairs while the speaking module leverages self-supervised audio models obtained from speech-only data. VecTok Speech [137] is a multi-task generation framework capable of zero-shot VC, TTS, and Speech-to-Speech Translation (S2ST). It introduced a novel speech codec with an encoder-decoder architecture, where the encoder extracts speech vectors and semantic tokens. Moreover, an autoregressive language model predicts semantic tokens, followed by byte-pair encoding (BPE), to reduce exposure bias and improve context coverage. The Semantic tokens remove speaker-related information while preserving its content. Similarly, Xue et al. [138] proposed a novel codec-based TTS algorithm that improves context utilization and handles longer speech prompts. They introduced multi-modal context-enhanced Qformer (MMCE-Qformer), building on Q-former [142], to utilize multi-modal context information. They also leveraged a pre-trained LLM to predict semantic tokens and use the SoundStorm model [143] to predict acoustic tokens, thereby improving speech quality and speaker similarity. Lee et al. [139] introduced a novel two-stage TTS framework. The text goes through an interpreting module based on Token Transducer++, a modified version of Token Transducer [144], to obtain semantic tokens and a speaking module based on a conformer-based architecture with a Grouped Masked Language Model (G-MLM) [145], to convert those tokens into acoustic ones. Finally, a codec decoder converts them into a speech waveform. Lei et al. [140] proposed a novel TTS algorithm using multi-scale acoustic prompts to better capture the timbre and speaking style of

the target speaker. Additionally, they use a speaker-aware text encoder and a modified VALL-E acoustic decoder to improve speaker similarity while maintaining high-quality speech. NaturalSpeech 3 [141] is a novel diffusion-based method that factorizes speech into various attributes (content, prosody, timbre, acoustic details) by introducing a novel factorized diffusion model and a neural codec with factorized vector quantization (FVQ). These contributions improved the speech's naturalness, intelligibility, prosody, and speaker similarity.

**Emotion Transfer:** Maintaining the emotional characteristics of the reference audio and, in some instances, even controlling the emotional state of the synthesized speech is crucial toward expressive speech generation. StyleTTS [146] is a style-based generative modeling TTS algorithm that synthesizes speech with the prosody and emotion of the reference audio. It introduced a transferable Monotonic Aligner (TMA) for better text-speech alignment and duration-invariant data augmentation for improved duration prediction. In addition, it leveraged adaptive instance normalization (AdaIN) to incorporate style vectors, enabling natural and diverse speech generation. StyleTTS 2 [147] builds upon StyleTTS by proposing the usage of style diffusion models, adversarial training, and a pre-trained WavLM [148] during training, and a pre-trained PL-BERT [149] to enable natural and expressive speech. It performs like VALL-E on zero-shot speaker adaptation but uses 250 times less data, and can synthesize emotion-specific speech without a reference speech. DINO-VITS [150] introduced a data-efficient ZS-TTS system that utilizes a self-supervised loss to enhance robustness in noisy conditions. It captures intra-speaker variation in speaking style and emotion, proven through an emotion recognition classifier on speaker embeddings. This architecture consists of an mBART-based Text-to-Unit (T2U) module for feature extraction [151], a speaker encoder based on CAM++ [152] speaker verification model and a VITS-based Unit-to-Speech (U2S) module. DINO-VITS employs a multi-objective training approach integrating self-supervised DINO loss alongside speech synthesis objective losses to improve speaker representation learning and noise robustness. On the other hand, ZET-Speech [153] differs from previous models as it is an emotion-controllable TTS model requiring a neutral reference speech and an emotional label to generate emotion-specific audio. Its architecture is based on Grad-StyleSpeech and uses adversarial training and guidance methods for emotion-speaker disentanglement.

**Parameter Efficiency:** Ensuring the parameter efficiency of voice cloning algorithms can assist in easier deployment without requiring specialized hardware. Paul et al. [154] proposed a speaker-conditional WaveRNN (SC-WaveRNN) universal vocoder that efficiently performs speech generation of unseen speakers and acoustic conditions. It is a WaveRNN-variant that uses explicit speaker information, such as speaker embeddings, to improve performance without speaker adaptation or retraining the network, demonstrating enhanced speech naturalness and speaker similarity in zero-shot TTS. SC-GlowTTS [155] builds on GlowTTS [156] to achieve an efficient zero-shot multi-speaker TTS model. They use an external speaker encoder based on Angular prototypical loss [157], an adapted HiFi-GAN v2 vocoder, and explored the performance

while using Glow TTS's transformer-based encoder, residual dilated convolutional network [158] and a gated convolutional network [159]. Similarly, Ogun et al. [160] build on GlowTTS to improve utterance diversity by introducing an explicit stochastic pitch predictor and a flow-based duration predictor. Yoon et al. [161] introduced sparse attention, a transformer pruning method, to improve the TTS algorithm's generalizability. They selected StyleSpeech as the baseline model and then implemented sparse attention on the decoder only as it has more self-attention layers compared to the encoder, and applying it to the encoder degrades the performance. This cuts the redundant self-attention layer connections with weights below an automatically set threshold. Lux et al. [162] used FastSpeech 2 as the base TTS model and proposed an utterance-level normalization and utterance-level speaker embeddings. They also introduced a lightweight aligner for fine-grain prosodic feature extraction and fine-tuning within seconds. TacoLM [163] is a two-stage codec-based TTS model consisting of EnCodec [25] as an audio tokenizer and a BPE encoder to extract text representation. The authors improved inference efficiency and reduced model size by introducing two contributions: a gated attention mechanism as a replacement for regular multi-head attention used in transformer-based language models and a gated cross-attention layer to improve computational and storage efficiency. CM-TTS [164] is a generative novel architecture based on consistency models to reduce the steps required to achieve high-quality speech synthesis. The authors also introduced weighted samplers to ensure unbiased learning through the training process. Similarly, FlashSpeech [165] is a novel adversarial consistency training approach to train a latent consistency model from scratch without a pre-trained diffusion model. Additionally, it introduced a prosody generator to improve the diversity of the speech synthesized. This model achieves high-quality performance 20 times faster than other TTS systems. Fujita et al. [166] proposed a lightweight system using a mixture of adapters (MoA) to enhance speaker adaptation while maintaining high-quality speech synthesis with faster inference. These modules are integrated into the decoder and the variance adapter of a non-autoregressive TTS model such as FastSpeech 2. SimpleSpeech [167] is a NAR diffusion-based TTS system trained with speech-only data without alignment information and can generate speech directly from text. It proposed a speech-codec model based on scalar quantization that maps speech into a compact latent space, reducing complexity and increasing generation speed. Small-E [168] is a codec-based language model that relies on an encoder-decoder architecture in contrast to a decoder-only mechanism. It introduced Linear Casual Language Model (LCLM) blocks as a replacement for AR transformers, leading to efficient training. The authors also proposed position-aware cross-attention (PACA) tackling skipping & repeating errors. P-Flow [169] is a ZS-TTS model that consists of a speech-prompted text encoder within a NAR TTS model and a flow-based generative decoder for efficient speech synthesis. This text encoder enables maintaining the emotional prosody of the reference speech for synthesis.

**Generalizable Speech Synthesis:** The versatility of Speech synthesis models is crucial to handling various tasks such as

voice conversion, speech editing, and zero-shot TTS, or the diverse speaking styles. Glow-WaveGAN 2 [170] expands on Glow-WaveGAN [171] to enable zero-shot TTS and any-to-any VC. This architecture includes a Universal WaveGAN as the feature extractor and the vocoder, a multi-speaker GlowTTS, and an external speaker encoder. UnifySpeech [172] combines TTS and VC tasks into one framework by categorizing speech into three independent components: content, speaker, and prosody information. Everything in the model is shared except for the speech content extraction. Nansy++ [173] is a unified framework for various speech tasks such as VC, TTS, SVS, and voice design. Its backbone consists of an analyzer and synthesizer trained with speech-only data. Nansy++ analyzer consists of separate modules for pitch, linguistic, and timbre information that aim to disentangle the linguistic features from all speech representations. Ke et al. [174] proposed a FastSpeech2-based novel TTS method, Voicerifier, to improve zero-shot TTS performance across multiple accents. Voicerifier uses high-frequency mel spectrogram bins and two strategies to maintain timbre features while discarding content and rhythm. Then, they perform voice adaptation through Voice-Pin layer normalization (VPLN) to integrate timbre features with text features for better speaker adaptation. Zhang et al. [175] proposed a zero-shot multi-accent TTS model. They introduce an accent classifier on the Tacotron 2 framework to have more control of pronunciation variety, which is applicable for generating accented speech for unseen speakers. SpeechX [176] is a versatile speech generation model capable of zero-shot TTS for clean and noisy speech samples. They leverage NCLM by extending VALL-E with multi-task learning through task-dependent prompts. VoiceCraft [177] is a token-infilling NCLM capable of speech editing and ZS-TTS. It leveraged a transformer-decoder architecture and introduced a token rearrangement process to enable inter-sequence generation. Kanda et al. [178] proposed an Expressive Laughter-controllable ZS-TTS Engine (ELaTE) to integrate laughing sounds, built on top voicebox [179], a conditional flow-matching-based ZS-TTS, and condition it using a laugh detector through fine-tuning.

On a more general note, many researchers target improving speaker representation to enhance voice cloning performance. Lee et al. [180] proposed a Tacotron2-based ZS-TTS system utilizing two types of speaker embeddings. The first is extracted from a speaker encoder, while the second is obtained using a proposed attention-based VAE from an embedding dictionary. Multi-spectroGAN [181] is a multi-speaker TTS model fully trained through adversarial training. It employs an adversarial style combination to improve the generalization of unseen speaker styles. The model consists of a FastSpeech2-based generator and a frame-level conditional discriminator, which enhances speaker embeddings by distinguishing real mel spectrograms from generated ones. Choi et al. [182] proposed an adversarial speaker-consistency learning methodology to tackle the domain shift when generating speech for unseen speakers. This method utilizes external untranscribed speech during adversarial training to expose the model to a larger speaker pool. It is built on top of VITS architecture and demonstrated a performance increase in speech naturalness

and speaker similarity. Schnell et al. [183] proposed an all-pass wrap (APW) implementation, commonly used for vocal tract length normalization, capable of performing zero-shot speech synthesis. The idea is that APW is capable of generalizing to various speakers in its original task; thus, introducing an APW layer to an encoder-decoder TTS architecture should enhance the zero-shot adaptation, as was demonstrated in its result with an increase in speaker similarity while decreasing the audio quality. Lyu et al. [184] proposed a multi-scale zero-shot style transfer framework for ZS-TTS. They introduced a reference encoder with speaker-adaptive linear modulation (SALM), inspired by FiLM [36], to generate a scale and bias vector, improving adaptation performance to unseen speakers. They also designed a prosody extractor and predictor that leverages the reference audio and text to increase the diversity of the synthesized speech. Gorodetskiy et al. [185] proposed a novel attention-based TTS capable of replicating target voice and generalizable to long utterances. This architecture includes a speaker encoder, a synthesizer, and a universal vocoder. They proposed the usage of dynamic convolution attention with a modified tacotron2-based synthesizer. In addition, they conditioned both the synthesizer and vocoder on a pre-trained speaker encoder. Lee et al. [186] introduced a progressive VAE (PVAE) that learns speaking style gradually and extends it to PVAE-TTS, a novel multi-speaker adaptive TTS that progressively learns normalized representation to adapt to new speakers with limited data. It also uses dynamic style layer normalization (DSLN) to improve style adaptation quality. Cory et al. [187] presented a novel speaker encoder that relies on an encoder coupled with a multi-scale approach to learn the local and global features of the target speaker. They used SC-GlowTTS transformer variation as their baseline, and they demonstrated that their method outperforms a modified GE2E speaker encoder with angular prototypical loss. WavThruVec [188] is a two-stage architecture consisting of an encoder (text2vec) that converts text into Wav2Vec 2.0 embeddings using a pre-trained Wav2Vec 2.0 model and a decoder (vec2wav) to obtain a generated waveform. These contributions improved the generalizability of unseen speakers. Kim & Jeong et al. [189] proposed a TTS transfer learning methodology, using VITS as a baseline, that employs Wav2Vec 2.0 representations to utilize a large amount of unlabeled data for TTS model pre-training. AdaSpeech 4 [190] builds on top of AdaSpeech by introducing three major contributions: factorizing speaker characteristics as basis vectors and combining them through attention, employing CLN on the extracted speaker representation, and proposing a loss based on basis vector distribution. These contributions led to an improvement in speech quality and speaker similarity in a ZS-TTS scenario. Eskimez et al. [191] introduced a novel duration model, called total-duration aware (TDA), utilizing both text input and the total target duration. The purpose of this module is to maintain speech quality across various speech rate configurations while providing a more precise speech duration, leading to a performance boost in intelligibility and preserving the speaker's voice characteristics. VALL-T [192] is a generative transducer model designed to address robustness issues, such as hallucinations and high word error rates, observed in decoder-only TTS algorithms

TABLE V

COMPARISON OF ZERO-SHOT TTS ALGORITHMS COVERING NATURALNESS (NAT) AND SPEECH QUALITY (QUAL). FOR MULTIPLE DATASETS, RESULTS ARE AVERAGED. PAPERS ARE ARRANGED IN ASCENDING ORDER BY PUBLICATION DATE.

Paper	Test set, additional details	TTS Evaluation
DiffVoice [201]	LibriTTS (5 utt/reference, 21 reference audios) — [*P=Prompt]	Nat (NMOS) $\uparrow$ : DiffVoice-P (4.66) > VITS (x-vector) (4.14) > YTTS (3.46) > FS2 (x-vector) (3.42) > MSS (2.82)
GZS-TV [123]	LibriTTS test-set (5 utt/spk, 37 spks); VCTK (2 utt/spk, 108 spks)	Nat (NMOS) $\uparrow$ : GZS-TV (3.76) > YTTS (3.65) > SS (3.22) > MSS (3.20) Qual (WER) $\downarrow$ : GZS-TV (8.55) > YTTS (8.95) > SS (9.65) > MSS (9.85)
STEN-TTS [238]	VCTK (10 spks)	Nat (NMOS) $\uparrow$ : STEN-TTS (3.01) > DiffSinger [217] (2.55) > SS (2.50)
HierSpeech++ (HS++) [133]	LibriTTS test-clean (4837 samples) LibriTTS test-other (5120 samples)	Nat (NMOS) $\uparrow$ : HS++ (4.55) > HS (4.05) > VALL-E X (3.52) > YTTS (3.51) > XTTS (3.17) Qual (WER) $\downarrow$ : HS++ (5.20) > HS (7.88) > YTTS (11.97) > XTTS (28.89) > VALL-E X (32.88)
Latent Filling [233]	VCTK (11 spks)	Nat (NMOS) $\uparrow$ : Proposed (3.82) > YTTS (3.16) > SC-GlowTTS (2.02) Qual (WER) $\downarrow$ : Proposed (1.10) > YTTS (4.78) > SC-GlowTTS (28.56)
Mega-TTS 2 (MTTS2) [135]	LibriSpeech test-clean (100 sec/spk, 20 spks)	Qual (QMOS) $\uparrow$ : MTTS2-300s (4.12) > MTTS2-3s (3.99) > VALL-E-3s (3.89) > VALL-E-20s (3.41) Qual (WER) $\downarrow$ : MTTS2-300s (2.23) > MTTS2-3s (2.46) > VALL-E-3s (5.83) > VALL-E-20s (8.77)
USAT [92]	ESLTTS, VCTK, LibriTTS test set (10 utt/spk, 30 spks) each	Nat (NMOS) $\uparrow$ : USAT (3.89) > YTTS (3.73) > SS (3.32) > MSS (3.30) Qual (WER) $\downarrow$ : USAT (10.40) > YTTS (10.70) > SS (11.20) > MSS (11.70)
MRMI-TTS [126]	VCTK (9 spks); LibriTTS test-clean	Nat (NMOS) $\uparrow$ : MRMI-TTS (3.97) > MSS (3.73) > SS (3.70) = FS2 (d-vector) (3.70) Qual (MCD) $\downarrow$ : MRMI-TTS (5.02) > MSS (5.17) > SS (5.28) = FS2 (d-vector) (5.28)
P-Flow [169]	LibriSpeech test-clean (4-10 sec/sample/spk, 40 spks) = 2.2h [similar to VALL-E]	Qual (WER) $\downarrow$ : P-Flow (2.60) > VALL-E (5.90) > YTTS (7.70)
StyleTTS 2 (STTS2) [147]	LibriTTS test-clean (500 samples)	Nat (NMOS) $\uparrow$ : STTS2 (4.15) > STTS+HiFi-GAN (3.91) > VITS (3.69) > YTTS (2.35)
VALL-E 2 [205]	LibriSpeech test-clean (4-10 sec/sample/spk, 40 spks) = 2.2h — *GroupSize = 1	Qual (WER) $\downarrow$ : VALL-E 2 [FTS] (0.65) > VALL-E [FTS] (0.75) > VALL-E 2 (1.50) > VALL-E (3.10)
CM-TTS [164]	VCTK (512 samples); LJSpeech (512 samples)	Qual (QMOS) $\uparrow$ : CM-TTS (T=4) (3.82) > DiffGAN-TTS [218] (T=4) (3.66) Qual (MCD) $\downarrow$ : DiffGAN-TTS [218] (T=4) (8.66) > CM-TTS (T=4) (8.90) Qual (WER) $\downarrow$ : DiffGAN-TTS [218] (T=4) (7.98) > CM-TTS (T=4) (8.97)
NaturalSpeech 3 (NS3) [141]	LibriSpeech test-clean (1 utt/spk, 40 spks) *All are trained using LibriLight dataset	Nat (CMOS) $\uparrow$ : NS3 (reference) > NS2 (-0.18) > MTTS2 (-0.20) > VB (-0.32) > VALL-E (-0.60) Qual (WER) $\downarrow$ : NS3 (1.81) > NS2 (1.94) > VB (2.14) > MTTS2 (2.32) > VALL-E (6.11)
SpeechX [176]	LibriSpeech test-clean (4-10 sec/sample/spk, 40 spks) = 2.2h [similar to VALL-E]	Qual (WER) $\downarrow$ : SpeechX (VALL-E init) (4.66) > SpeechX (Random init) (5.40) > VALL-E (5.90)
VoiceCraft [177]	LibriTTS (125 utt); YouTube-GigaSpeech test-set (125 utt) Objective—Subjective: 125—40 utt/dataset	Nat (NMOS) $\uparrow$ : VoiceCraft (4.17) > XTTS v2 [219] (3.96) > VALL-E (3.86) > YTTS (2.79) Qual (WER) $\downarrow$ : XTTS v2 [219] (3.60) > VoiceCraft (4.50) > YTTS (6.60) > VALL-E (7.10)
Xue et al. [138]	LibriSpeech test-clean (1249 sentences) iEMOCAP (1079 sentences)	Nat (NMOS) $\uparrow$ : Proposed (3.79) > XTTS (3.44) > VALL-E (3.24) Qual (MCD) $\downarrow$ : Proposed (7.72) > XTTS (7.95) > VALL-E (8.76)
Lee et al. [139]	LibriTTS test-clean (500 sentences)	Nat (NMOS) $\uparrow$ : Proposed (3.94) > VALL-E X (3.75) > VITS (ECAPA TDNN) (3.52) Qual (CER) $\downarrow$ : Proposed (2.34) > VITS (ECAPA TDNN) (4.81) > VALL-E X (8.72)
SimpleSpeech [167]	LibriTTS test-clean following [220] (1 utt/spk)	Nat (NMOS) $\uparrow$ : SimpleSpeech (4.45) > XTTS (4.19) > VALL-E X (4.07) Qual (MCD) $\downarrow$ : SimpleSpeech (7.51) > VALL-E X (8.16) > XTTS (9.93) Qual (WER) $\downarrow$ : SimpleSpeech (3.30) > XTTS (5.60) = VALL-E X (5.60)
DINO-VITS [150]	CHiME3 (15 reference-audio/spk, 8 spks)	Nat (NMOS) $\uparrow$ : [Clean] DINO-VITS (4.00) > YTTS (3.96) Nat (NMOS) $\uparrow$ : [Noisy] DINO-VITS (3.55) > YTTS+denoiser (3.28) > YTTS (3.11)
Zhou et al. [204]	LibriTTS (test-clean & test-other)— *WavLM and Hubert uses K-means (K=1024)	Nat (NMOS) $\uparrow$ : Proposed (w/ WavLM) (3.54) > Proposed (w/ Hubert) (3.24) > VALL-E (2.50) Qual (WER) $\downarrow$ : Proposed (w/ WavLM) (7.02) > Proposed (w/ Hubert) (9.90) > VALL-E (23.31)
TacoLM [163]	LibriSpeech test-clean (4-10 sec/sample/spk, 40 spks) = 2.2h [similar to VALL-E]	Nat (CMOS) $\uparrow$ : TacoLM (0.12) > VALL-E (reference) Qual (WER) $\downarrow$ : TacoLM (5.96) > VALL-E (6.25)
Small-E [168]	LibriTTS test-clean (15 samples)	Nat (NMOS) $\uparrow$ : MetaVoice (3.80) > Small-E (3.16) > YTTS (2.56)
MaskGCT [232]	LibriSpeech test-clean	Nat (CMOS) $\uparrow$ : NS3 (0.16) > MaskGCT (0.10) > GT (Ref) > VoiceCraft (-0.33) > VB (-0.41) > VALL-E (-0.52) > XTTS v2 (-0.98) Qual (WER) $\downarrow$ : NS3 (1.94) > VB (2.03) > MaskGCT (2.63) > XTTS v2 (4.20) > VoiceCraft (4.68) > VALL-E (5.90)

like VALL-E. These algorithms suffer from the absence of explicit duration modeling, which results in poor performance due to the lack of monotonic alignment constraints. VALL-T introduced shifting relative position embeddings for the input phonemes, enforcing monotonic alignment constraints and enhancing TTS robustness. On the other hand, other works focus on speaker conditioning techniques to enable and improve the performance of ZS-TTS. Choi et al. [193] introduced a speaker-normalized affine coupling (SNAC) layer to enable ZS-TTS through conditioning techniques. It normalizes the input through the predicted parameters from speaker embeddings. SC-CNN [194] is a speaker-conditioning technique that predicts convolutional kernels from speaker embeddings and applies 1D convolution with these kernels to phoneme sequences. Chen et al. [195] proposed a two-branch speaker control module (TSCM) designed to enhance speaker similarity in few-shot and zero-shot TTS. This module includes a gated convolutional network (GCN) for local feature modeling and a gated recurrent unit (GRU) to obtain utterance-level features. The authors demonstrate the module’s effectiveness through integration with FastSpeech 2 and VITS.

More recently, ZS-TTS systems started integrating diffusion models for high-quality speech synthesis. Jiang et al. [196] introduced a ZS-TTS model focusing on timbre generalization and prosody modeling. This includes a global vector to model speaker timbre while guiding prosody modeling. In addition, they adopted a DDPM-based pitch predictor

and a prosody adapter for hierarchical prosody modeling. Furthermore, Guided-TTS 2 [80], previously discussed for its few-shot TTS generation, is also capable of zero-shot TTS through a pre-trained speaker encoder. Bang et al. [197] proposed a ZS-TTS system by building on top of the diffusion-based Grad-TTS [198] and enabling speech synthesis for unseen speakers. They utilize a pre-trained speaker recognition model to obtain speaker information from reference speech and expand it through information perturbation. Similarly, Grad-StyleSpeech [199] is a ZS-TTS model that utilizes a hierarchical transformer encoder and score-based diffusion model, proposed in Grad-TTS, for generating speech while replicating the target speaker voice. U-Style [200] is a Grad-TTS-based ZS-TTS algorithm, enabling control over speech synthesis by utilizing independent speaker and style audio references from unseen speakers. It effectively uses a cascaded U-Net structure and signal perturbation to disentangle speaker-style information. Additionally, it employs normalization techniques and a diffusion-based decoder to enhance the quality and naturalness of synthesized speech. DiffVoice [201] is a novel TTS algorithm that starts with encoding speech into phoneme-level representation using VAE and is enhanced using adversarial training. It then jointly models these representations and phoneme durations with a diffusion model. More recently, NaturalSpeech 2 [202] tackles the challenge of general large-scale TTS systems’ robustness, including unstable prosody, skipping/repeating, and voice quality. It uses

TABLE VI  
COMPARISON OF ZERO-SHOT TTS ALGORITHMS COVERING SPEAKER SIMILARITY (SIM). FOR MULTIPLE DATASETS, RESULTS ARE AVERAGED UNLESS STATED OTHERWISE. PAPERS ARE ARRANGED IN ASCENDING ORDER BY PUBLICATION DATE.

Paper	Test set, additional details	TTS Evaluation
DiffVoice [201]	LibriTTS (5 utt/reference, 21 reference audios) — *P=Prompt	Sim (SMOS) $\dagger$ : DiffVoice-P (4.61) > VITS (x-vector) (3.80) > YTTS (3.51) > FS2 (x-vector) (3.44) > MSS (3.15) Sim (SECS) $\ddagger$ : DiffVoice-P (0.85) > YTTS (0.81) > VITS (x-vector) (0.80) > MSS (0.76) > FS2 (x-vector) (0.75)
GZS-TV [123]	LibriTTS test-set (5 utt/spk, 37 spks); VCTK (2 utt/spk, 108 spks)	Sim (SMOS) $\dagger$ : GZS-TV (3.79) > YTTS (3.59) > MSS (3.31) > SS (3.30) Sim (SECS) $\ddagger$ : GZS-TV (0.81) > YTTS (0.79) > MSS (0.75) = SS (0.75)
STEN-TTS [238]	VCTK (110 spks)	Sim (SMOS) $\dagger$ : STEN-TTS (3.72) > SS (2.55) > DiffSinger [217] (2.50)
HierSpeech++ (HS++) [133]	LibriTTS test-clean (4837 samples) LibriTTS test-other (5120 samples)	Sim (SMOS) $\dagger$ : HS++ (3.68) > XTTS (3.49) > HS (3.31) > VALL-E X (3.21) > YTTS (3.06) Sim (SECS) $\ddagger$ : HS++ (0.89) > VALL-E X (0.85) > HS (0.80) = YTTS (0.80) > XTTS (0.77)
Latent Filling [233]	VCTK (11 spks)	Sim (SMOS) $\dagger$ : Proposed (3.34) > YTTS (2.79) > SC-GlowTTS (2.08) Qual (SECS) $\ddagger$ : Proposed (0.84) > YTTS (0.81) > SC-GlowTTS (0.60)
Mega-TTS 2 (MTTS2) [135]	LibriSpeech test-clean (100 sec/spk, 20 spks)	Sim (SMOS) $\dagger$ : MTTS2-300s (4.01) > MTTS2-3s (3.75) > VALL-E-3s (3.70) > VALL-E-20s (3.25) Sim (SECS) $\ddagger$ : MTTS2-300s (0.93) > MTTS2-3s (0.90) > VALL-E-3s (0.89) > VALL-E-20s (0.81)
USAT [92]	ESLTTS, VCTK, LibriTTS test set (10 utt/spk, 30 spks) each	Sim (SECS) $\dagger$ : USAT (0.73) > YTTS (0.71) > SS (0.65) = MSS (0.65) Sim (SMOS) $\ddagger$ : USAT (3.69) > YTTS (3.42) > MSS (3.10) > SS (3.06)
MRMI-TTS [126]	VCTK (9 spks); LibriTTS test-clean	Sim (SMOS) $\dagger$ : MRMI-TTS (4.02) > MSS (3.82) > SS (3.78) = FS2 (d-vector) (3.78) Sim (SECS) $\ddagger$ : MRMI-TTS (0.82) > MSS (0.81) > SS (0.80) = FS2 (d-vector) (0.80)
P-Flow [169]	LibriSpeech test-clean (4-10 sec/sample/spk, 40 spks) = 2.2h [similar to VALL-E]	Sim (SECS) $\dagger$ : VALL-E (0.58) > P-Flow (0.54) > YTTS (0.34)
StyleTTS 2 (STTS2) [147]	LibriTTS test-clean (500 samples)	Sim (SMOS) $\dagger$ : STTS2 (4.03) > STTS+HiFi-GAN (4.01) > VITS (3.54) > YTTS (2.42)
VALL-E 2 [205]	LibriSpeech test-clean (4-10 sec/sample/spk, 40 spks) = 2.2h — *GroupSize = 1	Sim (SECS) $\dagger$ : VALL-E 2 [FTS] (0.69) > VALL-E [FTS] (0.68) > VALL-E 2 (0.64) > VALL-E (0.63) Sim (SMOS) $\ddagger$ : VALL-E 2 (4.61) > VALL-E (4.45)
CM-TTS [164]	VCTK (512 samples); LJSpeech (512 samples)	Sim (SECS) $\dagger$ : CM-TTS (T=4) (0.72) > DiffGAN-TTS [218] (T=4) (0.71)
NaturalSpeech 3 (NS3) [141]	LibriSpeech test-clean (1 utt/spk, 40 spks) *All are trained using LibriLight dataset	Sim (SMOS) $\dagger$ : NS3 (4.01) > NS2 (3.65) > MTTS2 (3.63) > VB (3.52) > VALL-E (3.46) Sim (SECS) $\ddagger$ : NS3 (0.67) > NS2 (0.55) > MTTS2 (0.55) > VB (0.48) > VALL-E (0.47)
SpeechX [176]	LibriSpeech test-clean (4-10 sec/sample/spk, 40 spks) = 2.2h [similar to VALL-E]	Sim (SECS) $\dagger$ : SpeechX (VALL-E init) (0.58) > SpeechX (Random init) (0.57) = VALL-E (0.57)
VoiceCraft [177]	LibriTTS (125 utt); YouTube-GigaSpeech test-set (125 utt) Objective—Subjective: 125—40 utt/dataset	Sim (SMOS) $\dagger$ : VoiceCraft (4.34) > VALL-E (4.07) > XTTS v2 [219] (3.44) > YTTS (2.79) Sim (SECS) $\ddagger$ : VoiceCraft (0.55) > VALL-E (0.50) > XTTS v2 [219] (0.47) > YTTS (0.41)
Xue et al. [138]	LibriSpeech test-clean (1249 sentences) IEMOCAP (1079 sentences)	Sim (SMOS) $\dagger$ : Proposed (3.86) > XTTS (3.44) > VALL-E (3.29) Sim (SECS) $\ddagger$ : Proposed (0.76) > XTTS (0.71) > VALL-E (0.68)
Lee et al. [139]	LibriTTS test-clean (500 sentences)	Sim (SMOS) $\dagger$ : Proposed (3.64) > VITS (ECAPA TDNN) (3.23) > VALL-E X (3.08) Sim (SECS) $\ddagger$ : Proposed (0.51) > VALL-E X (0.43) > VITS (ECAPA TDNN) (0.39)
SimpleSpeech [167]	LibriTTS test-clean following [220] (1 utt/spk)	Sim (SMOS) $\dagger$ : VALL-E X (4.15) > SimpleSpeech (4.14) > XTTS (4.00) Sim (SECS) $\ddagger$ : SimpleSpeech (0.96) > XTTS (0.95) = VALL-E X (0.95)
DINO-VITS [150]	ChiME3 (15 reference-audio/spk, 8 spks)	Sim (SMOS) $\dagger$ : [Clean] DINO-VITS (3.85) > YTTS (3.33) Sim (SECS) $\ddagger$ : [Noisy] DINO-VITS (3.52) > YTTS+denoiser (3.35) > YTTS (3.20)
TacoLM [163]	LibriSpeech test-clean (4-10 sec/sample/spk, 40 spks) = 2.2h [similar to VALL-E]	Sim (SMOS) $\dagger$ : TacoLM (3.75) > VALL-E (3.50) Sim (SECS) $\ddagger$ : TacoLM (0.87) > VALL-E (0.86)
Zhou et al. [204]	LibriTTS (test-clean & test-other) — *WavLM and Hubert uses K-means (K=1024)	Sim (SMOS) $\dagger$ : Proposed (w/ WavLM) (3.75) > Proposed (w/ Hubert) (3.46) > VALL-E (3.11)
Small-E [168]	LibriTTS test-clean (15 samples)	Sim (SMOS) $\dagger$ : MetaVoice (3.91) > Small-E (3.08) > YTTS (2.54)
MaskGCT [232]	LibriSpeech test-clean	Sim (SMOS) $\dagger$ : MaskGCT (4.27) > NS3 (4.26) > VB (3.80) > VoiceCraft (3.52) > VALL-E (3.47) > XTTS v2 (3.02) Sim (SECS) $\ddagger$ : MaskGCT (0.69) > NS3 (0.67) > VB (0.64) > XTTS v2 (0.51) > VALL-E (0.50) > VoiceCraft (0.50)

a neural audio codec with a residual vector quantizer and a diffusion model to generate latent vectors conditioned on text input, and the decoder uses these vectors to generate the target speech. RALL-E [203] introduced chain-of-thought (CoT) prompting within an NCLM to improve robustness in LLM-based TTS. This means breaking a complex task into a series of more straightforward ones to improve performance. This model starts by predicting prosody tokens (pitch and duration). Then, it predicts speech tokens conditioned on those and the input sequence. In addition, it utilizes the predicted duration to mask irrelevant phonemes and prosody tokens to ensure the model concentrates only on the relevant parts. Zhou et al. [204] proposed a phonetic-enhanced language modeling method by leveraging SSL to train the AR language model. Subsequently, the mapping from phonetic variations to fine-grained acoustic details occurs in an NAR manner. This approach reduces error propagation during training, improving robustness and naturalness. VALL-E 2 [205] introduced two VALL-E enhancements achieving human parity. The first is repetition-aware sampling, which enhances the stability of the decoding process compared to random sampling. Additionally, five-time sampling (FTS) generates five outputs and selects the best one based on specific metrics, improving robustness. The second modification, grouped code modeling, increases inference speed by organizing codec codes into groups.

As TTS algorithms develop, researchers focus on exploring

algorithms and their components. Cooper et al. [72] explored the effect of various neural speaker embeddings on speaker similarity performance of unseen speakers. They showed that LDE with angular softmax loss performs better than x-vectors in speaker verification. SATTS [206] explored speaker attractors, a speech separation technique, for ZS-TTS algorithms. Speaker attractors are embedding vectors that pull time-frequency bins of the same speaker together while pushing those of other speakers, which proved useful with worse recording conditions. Recently, Azizah et al. [207] explored 24 Tacotron2-based zero-shot TTS variants and observed their performance on speech synthesis for dysphonia speakers. These variants include changing the type of input sequence, speaker embedding and its position, and loss function. They concluded that the best-performing models use a grapheme-phoneme input sequence, a speaker model to generate speaker embeddings placed at the TTS encoder only, and a combined loss using speaker consistency and frame-level speech loss.

**Non-English ZS-TTS** Zero-shot TTS algorithms have shown remarkable advancements in synthesizing human-like speech. However, it is important to focus on other languages to address the world's linguistic diversity. Papers [208]–[210] focus on Mandarin, paper [211] focus on Japanese, while [212], [213] focus on the Vietnamese language. Zhou et al. [208] proposed a content-dependent fine-grained speaker embedding (CDFSE) approach tackling Mandarin ZS-TTS.

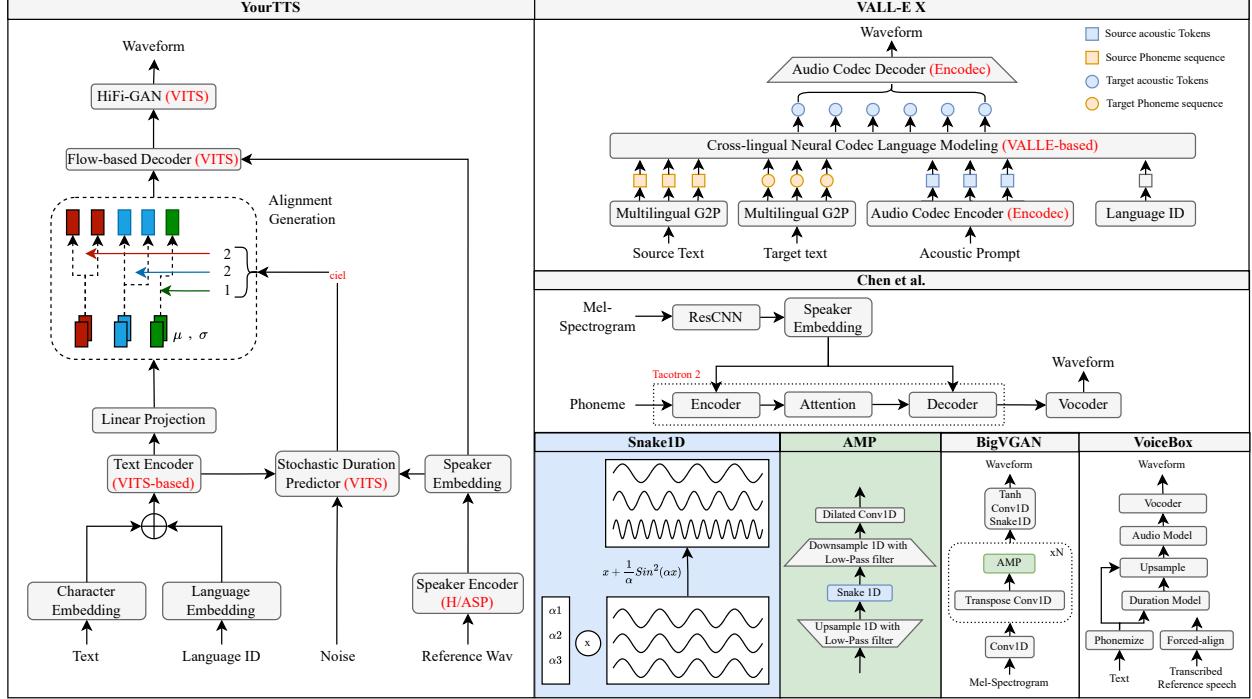


Fig. 7. Multilingual Voice Cloning algorithms, adapted from YourTTS [234], VALL-E X [229], BigVGAN [250], VoiceBox [179], and Chen et al. [222].

They extract local content and speaker embeddings from a reference speech, introducing pronunciation information. This approach improves speaker similarity for unseen speakers by leveraging a content-dependent reference attention module. iEmoTTS [209] is a novel TTS system for cross-speaker emotion transfer through prosody-timbre disentanglement. This paper proposed a system consisting of an emotion encoder, a prosody predictor, a timbre encoder, and a probability-based emotion intensity control mechanism. These contributions led to speech synthesis while controlling emotion intensity for new speakers. SponTTS [210] is a two-stage approach, based on neural bottleneck (BN) features, for spontaneous TTS generation used in casual conversations. The first stage, text2BN, based on FastSpeech, maps text to speaker-independent BN features conditioned on spontaneous style representation using a conditional VAE (CVAE). The second stage, BN2Wave, following a VITS-based model, transforms BN intermediate values conditioned on speaker embeddings to generate speech similar to the target speaker. For Japanese, Fujita et al. [211] introduced a zero-shot TTS method, incorporating adapters into an SSL model fine-tuned using noisy reference speech. They also included a speech enhancement module for further noise reduction. For Vietnamese, Calib-StyleSpeech [212] is a FastSpeech2-based ZS-TTS algorithm that tackles accent imbalance between datasets. Their model includes a style extraction block to condition the acoustic features. Simultaneously, an independent block extracts a content vector to disentangle content-style information through a mutual information constraint. Adapt-TTS [213] is a zero-shot TTS model that uses an extracting mel-vector (EMV) to extract speaker and style information and a diffusion-based mel spectrogram

denoiser to improve the quality of the generated voice.

**Summary:** Zero-Shot TTS algorithms witnessed the most significant increase in papers within voice cloning in the past few years. After the release of SPEAR-TTS and VALL-E, many works started employing codec-based architecture for zero-shot TTS. Tables V and VI show the performance of many zero-shot TTS architectures in naturalness, speech quality, and speaker similarity. Similar to FS-TTS, calculating WER and CER require using a pre-trained ASR to evaluate the speech quality of the generated waveform. These systems include Whisper ASR (base or large), CTC-based HuBERT [28], Conformer Transducer [117], and Wav2Vec 2.0 [27] to quantify the speech quality performance of TTS systems. Speaker Embedding Cosine similarity scores are obtained through the Resemblyzer package, WavLM [148], XLSR-53 [214], TitaNet-L, and SpeechBrain toolkit [215], [216].

#### D. Multilingual Voice Cloning

Building multilingual TTS while maintaining speaker characteristics is crucial for advancing personalized chatbots. We will explore speaker adaptation, FS-TTS, and ZS-TTS across multiple languages. Fig. 7 shows some of these algorithms, while Table VIII lists work focusing on such systems.

**2 Languages:** This paragraph will cover multilingual TTS systems with only two languages in the following order: speaker adaptation, FS-TTS, and ZS-TTS systems. Xin et al. [221] proposed a speaker adaptation method for mono/cross-lingual speaker adaptation for English and Japanese. They train a language-independent speaker encoder and use the resultant speaker embeddings to train a monolingual multi-speaker TTS. Finally, they adapt to new speakers by fine-tuning the model

with a speaker consistency loss. Other works focus on multilingual FS-TTS algorithms, Chen et al. [222] introduced a novel speaker embedding network that enables bilingual learning between English and Chinese, even for unseen speakers. This consists of a ResCNN [223] speaker encoder, a Tacotron2-based mel spectrogram generation network, and a Griffin-Lim vocoder. AdaDurIAN [224] improved the DurIAN-based average model and adapted it into a few-shot mono/cross-lingual speaker adaptation through a few minutes of monolingual data. The authors' contributions include building a speaker-independent content encoder, fine-tuning speaker embedding and decoder instead of the entire model for better pronunciation, and adopting an LSTM post-net module to improve speech quality. FastGraphTTS [225] is an end-to-end TTS framework that incorporates syntactic information to improve prosody consistency between the input text and the generated speech using a graph neural network (GNN). Other works focus on ZS-TTS algorithms, Zheng et al. [226] presented a multi-stream text encoder and efficient speaker representation to improve the performance for zero-shot cross-lingual TTS on unseen speakers. They utilize a speaker encoder with random sampling and a language adversarial loss for speaker-content-language disentanglement. NNSpeech [227] is a speaker-guided CVAE that enables unseen speaker voice cloning with only one utterance. Cai et al. [228] proposed a bilingual Mandarin-English TTS system operating in two distinct ways depending on the data available per speaker. In the case of limited data, they used cascading multi-speech modules, including a VC module for monolingual and cross-lingual speaker adaptation. VALL-E X [229] is an extension of the VALL-E network to enable cross-lingual speech synthesis. It introduced training a multilingual conditional codec language modeling and synthesizing the unseen speaker's characteristics into the target language. More recently, Seed-TTS [230] is a family of large-scale TTS models capable of performing ZS-TTS and emotion control through instruction fine-tuning. They use a self-distillation approach for timbre disentanglement, enabling high-quality VC and a reinforcement learning-based approach to enhance the model's robustness. Additionally, they introduced a diffusion-based NAR variant that performs speech generation in an end-to-end manner without pre-estimated phoneme durations. MobileSpeech [231] is a light-weight FastSpeech2-based ZS-TTS model that leverages a parallel speech mask codec decoder that uses hierarchical information from the speech codec and employs weight mechanisms across its layers during generation, achieving competitive performance in both English and Mandarin while reducing latency. Wang et al. [232] proposed Masked Generative Codec Transformer (MaskGCT), a two-stage fully NAR TTS model that does not require an explicit phoneme duration predictor and text-speech alignment supervision. The first stage predicts semantic tokens using a speech SSL model, while the second stage, based on SoundStorm [143], uses a masked generative codec transformer conditioned on the semantic tokens. Unlike previous methods, Latent Filling [233] is a speaker embedding latent space augmentation through a consistency loss to improve the performance of intra-lingual (English to English) and cross-lingual (Korean to English) ZS-TTS. This ensures

TABLE VII  
TABLE OF LANGUAGE ACRONYMS.

Acronym	Full Form		Acronym	Full Form		Acronym	Full Form
EN	English		PT	Portuguese		JV	Javanese
ZH	Chinese		SP	Spanish		VI	Vietnamese
KO	Korean		JA	Japanese		DE	German
SU	Sundanese		ID	Indonesian		FR	French
HI	Hindi		MR	Marathi		-	-

TABLE VIII  
MULTILINGUAL VOICE CLONING PAPERS, FOLLOWING ABBREVIATIONS FROM TABLE VII.

Ref	Languages		Ref	Languages
[221]	EN-JA		[222], [224]–[232]	EN-ZH
[233]	EN-KO		[234]	EN-FR-PT
[236]	HI-DE-SP-MR		[237]	EN-ID-JV-SU
[238]	EN-ZH-ID-JA-VI		[239]	EN-DE-SP-FR-KO
[179]	6 languages		[244]	6 languages
[246]	7 languages		[243]	8 languages
[250]	9 languages		[245]	11 languages
[241]	12 languages		[247]	16 languages
[240]	50 language locales		-	-

that the generated acoustic features of the augmented speaker embedding maintain the same speaker representation.

**3 to 5 Languages:** Unlike TTS models with only two languages, more languages require a general algorithm as languages differ in their alphabets and pronunciation. We will start with limited multilingual algorithms. YourTTS [234] is an extension of VITS for zero-shot TTS and multilingual training. This includes using text instead of phonemes, as not all languages have suitable open-source grapheme-to-phoneme converters. It also integrates a trainable language embedding with input characters and a H/ASP speaker encoder [235] to enable multilingual zero-shot TTS. Parrot-TTS [236] is a modularized TTS model that leverages SSL and performs well on adapting to unseen languages while preserving the speaker's voice characteristics, and it performs well for unseen speaker ZS-TTS in four languages. This architecture trains a seq2seq module as its text encoder and a pre-trained HuBERT as its speech encoder. The embeddings obtained from both generate the output waveform using a HiFi-GANv2 vocoder. Azizah et al. [237] proposed a novel partial network-based transfer learning to handle low-resource language. They utilized text and language encoders to improve multilingual performance. In addition, they used an explicit style control for TTS conditioning and an utterance-level reconstruction loss to improve the performance of zero-shot speaker adaptation. STEN-TTS [238] is a diffusion-based multilingual ZS-TTS capable of cross-lingual synthesis with a 3-second reference audio. This method also introduces style-enhanced normalization (STEN) into the diffusion framework to maintain the speaker's style, a challenge to previous diffusion-based TTS systems. Hemati et al. [239] introduced a Tacotron2-based model suitable for multilingual speaker adaptation. Unlike previous methods, they

use the International Phonetic Alphabet (IPA) as a direct input that goes through a trainable LUT for better generalization across languages. They explore transfer learning to enhance inter-lingual and cross-lingual speaker adaptation performance.

**5+ Languages:** This section discusses the broader realm of multilingual algorithms. Yang et al. [240] explored building a universal seq2seq TTS model capable of synthesizing any speaker’s voice in any language. This framework includes a transformer-based acoustic model, a WaveNet vocoder, and global conditioning from speaker and language networks. The global conditioning uses speaker embeddings obtained from a LUT followed by a mapping network to generate features with similar dynamic range. Similarly, the language embeddings are obtained from a language network with the same architecture. To enable a large multilingual universal synthesis, the authors proposed concatenating these features with the output of the acoustic model. Lux et al. [241] proposed a modified encoder and the usage of language-agnostic meta-learning (LAML) for zero-shot cross-lingual TTS that performs well for unseen speakers. This work uses articulatory feature vectors [242] to share knowledge across languages, making it a language-agnostic approach. It also introduced the usage of word boundaries to improve multilingual TTS performance. Jeon et al. [243] explored language-agnostic voice cloning through a multi-level attention aggregation approach, evaluating its effectiveness on unseen speakers within seen languages and unseen speakers and languages. Voicebox [179] is a versatile text-guided speech generative model capable of noise removal, content editing, diverse sample generation, style conversion, and mono/cross-lingual zero-shot TTS. It is a flow-matching model trained through infilling given a text and audio context. ZMM-TTS [244] is a multilingual zero-shot TTS model that leverages discrete speech representation from an SSL model and is capable of ZS-TTS for unseen speakers and languages, including low-resource languages. The authors explored the effect of various input representations on multilingual TTS performance, such as characters, IPA, or pre-trained phoneme representation, in which the latter demonstrated better overall performance. CLAM-TTS [245] is a codec-based TTS model that employs a probabilistic residual vector quantization to improve token length compression while performing with comparable or better performance compared to SOTA neural codec-based TTS models. Jeong et al. [246] introduced a transfer learning scheme consisting of two stages: unsupervised pre-training using speech-only data by leveraging self-supervised speech representations obtained through Wav2Vec 2.0. Then, supervised fine-tuning to generalize to new languages and speakers efficiently. To handle multilingual diversity, they utilized XLSR-53 [79], based on Wav2Vec 2.0, to obtain language-independent linguistic features useful for various downstream tasks. XTTS [247] builds on top of Tortoise TTS [248] to enable multilingual training. This model consists of VQ-VAE, an encoder comprising various components, GPT-2, and a HiFi-GAN vocoder. They use a BPE tokenizer [249] to obtain text tokens as input to GPT-2. They also romanized Korean, Japanese, and Chinese before tokenization. Researchers mainly focus on acoustic modeling, leading to less research on other components, such as the vocoder to handle

speaker diversity. BigVGAN [250] introduced a universal GAN-based vocoder that achieves SOTA performance on zero-shot out-of-distribution conditions, including unseen speakers and languages. They enhanced the GAN-based generator by introducing a periodic activation function that provides inductive bias into audio synthesis. It also introduced an anti-aliased multi-periodicity composition (AMP) module that composes multiple signals with learnable periodicities and a low-pass filter to reduce high-frequency artifacts.

## V. DATASETS & EVALUATION METRICS

Open-source speech datasets, summarized in Table IX, are vital for advancing monolingual and multilingual voice cloning algorithms. For example, speaker verification can improve speaker similarity, while ASR pre-trained models are used to test the intelligibility of the synthesized speech.

Evaluating voice cloning algorithms is crucial for an in-depth comparison focusing on various TTS aspects and the evaluation metrics are summarized in Table X. Speech naturalness measures how natural the voice sounds, measured subjectively using the Mean Opinion Score (MOS). Speech quality focuses on the clarity and intelligibility of the speech, evaluated using Mel-Cepstral Distortion (MCD), Character Error Rate (CER), and Word Error Rate (WER), in which the latter two depend on the pre-trained ASR system used for evaluation. Speech similarity focuses on the resemblance between the generated voice and the target speaker, evaluated using Speaker Embedding Cosine Similarity (SECS), which calculates the cosine similarity between the embeddings of the generated speech and the reference audio. Although this metric is objective, there is still some subjectivity due to the speaker encoder choice. Speech expression evaluates the style and emotions of the generated speech using prosody-related metrics such as Gross Pitch Error (GPE), Voicing Decision Error (VDE), and F0 Frame Error (FFE). Finally, Generation speed can be quantified using the real-time factor (RTF), indicating the time required to generate one second of audio.

## VI. APPLICATIONS & MISUSE

**Entertainment & Content Creation:** Voice cloning is highly valuable in the entertainment and content creation industry. For instance, automating voice-over for animated movies saves time and allows more focus on film design [284]. Moreover, it can speed up the movie dubbing process, expanding its reach to international audiences while preserving voice characteristics and emotions [285]. Content creators can benefit by having the option to edit audio narration to insert words in case of misspellings during the initial video [286].

**Personal Assistants:** Personalized assistants can benefit from expressive and realistic sounds as they improve the user experience. For example, the ability to generate speech with an appropriate emotional tone can ensure the user is engaged and comforted with the expressiveness of the AI-generated speech. Cerence introduced “My Car, My Voice” to let users create custom voices for their in-car assistants [287].

**Advertisement:** Voice cloning offers various possibilities to transform the advertisement industry [2]. It allows the creation

TABLE IX  
SPEECH DATASETS ACROSS VARIOUS LANGUAGES WITH A FOCUS ON TTS DATASETS.

Dataset	Task	Amount [hours (h), utterances (utt)]	Speakers [Male, Female]	Sampling Rate	Languages
Single Language					
CHiME3 [251]	ASR	11698 utt	83 [6+, 6+]	16 kHz	English
LibriSpeech [252]	ASR	982h	2484 [1283, 1201]	16 kHz	English
GigaSpeech [253]	ASR	[10kh labeled, 30kh unlabeled/weakly labeled]	N/A	16 kHz	English
VCTK [254]	TTS	44h	109 [N/A, N/A]	48 kHz	English
LJSpeech [255]	TTS	24h	1 [0, 1]	22.05 kHz	English
ESLTTS [92]	TTS	37h	134 [N/A, N/A]	24.00 kHz	English
Blizzard 2011 [256]	TTS	16.6h	1 [0, 1]	16 kHz	English
Blizzard 2013 [257]	TTS	319h	1 [0, 1]	Varied	English
Hi-Fi TTS [55]	TTS	292h	10 [4, 6]	44.1 kHz	English
LibriLight [258]	ASR	[10h labeled, 60kh unlabeled]	7000+ [N/A, N/A]	16 kHz	English
LibriTTS [259]	TTS	586h	2456 [1271, 1185]	24 kHz	English
JSUT [260]	TTS	10h	1 [0, 1]	48 kHz	Japanese
JVS [261]	VC, TTS	30h	100 [49, 51]	24 kHz	Japanese
CSMSC [262]	TTS	12h	1 [0, 1]	48 kHz	Mandarin
AISHELL-1 [263]	ASR	170h	400 [186, 214]	16 kHz	Mandarin
AISHELL-2 [264]	ASR	1000h	1991 [845, 1146]	16 kHz, 44.1 kHz	Mandarin
AISHELL-3 [265]	TTS	85h	218 [42, 176]	44.1 kHz	Mandarin
WENETSPEECH [266]	ASR	[10kh labeled, 2.4kh weakly labeled, 10kh unlabeled]	N/A	16 kHz	Mandarin
DidiSpeech-1 [267]	VC, TTS, ASR	572h	4500 [2094, 2406]	48 kHz	Mandarin
DidiSpeech-2 [267]	VC, TTS, ASR	227h	1500 [797, 703]	48 kHz	Mandarin
TITML-IDN [268]	ASR	14.5h	20 [11, 9]	16 kHz	Indonesian
INDspeech News LVCSR [269]	ASR	43.35h	400 [200, 200]	16 kHz	Indonesian
KhazakTTS 2 [270]	TTS	271h	5 [2, 3]	22.05 kHz	Kazakh
Multilingual datasets					
CommonVoice [271]	ASR, LID	2508h	58250 [N/A, N/A]	48 kHz	Multilingual
MLS [272]	ASR, TTS	50.5kh	6332 [3124, 3208]	16 kHz	Multilingual
OpenSLR [273]	TTS	25.4h	139 [49, 90]	22.05 kHz	Multilingual
Emilia [274]	TTS	101kh	N/A	24 kHz	Multilingual
VocCeleb1 [275]	SI, SV	352h	1251 [690, 561]	16 kHz	N/A
VocCeleb2 [276]	SI, SV	2442h	6112 [3761, 2351]	N/A	N/A
Emotional datasets					
IEEMOCAP [277]	SER	12h	10 [5, 5]	48kHz	English
ESD [278]	EVC	29h	20 [N/A, N/A]	N/A	English, Mandarin

of a brand voice tailored to different target audiences. This also includes multilingual voice cloning, where you can maintain the speaker's characteristics to reach a higher audience.

**Accessibility:** Few-shot and zero-shot voice cloning can provide value toward restoring the communication ability of people who lost their voice [45]. Dysphonia is a speech disorder leading to unclear and difficult-to-understand speech, affecting their social life and mental well-being. Voice cloning can assist in such cases as it can improve their speech intelligibility while maintaining voice characteristics [48].

**Misuse:** Voice cloning algorithm misuse has been occurring lately at an alarming rate as less data is required to replicate someone's voice. It is essential to understand that anyone can be a victim of an impersonation or financial scam. Hutiri et al. [288] presented a speech generation harm taxonomy, discussing various aspects of the non-ethical usage of voice cloning algorithms. For example, using people's voices by training and releasing their AI models can lead to identity theft, data laundering, impersonation, and coercion. Further-

more, voice cloning can lead to consumer fraud and financial losses. In August of 2019, a company's CEO was tricked by an audio deepfake to perform a fraudulent transfer of \$243,000 [289], proving the importance of understanding its potential misuse and developing detection algorithms.

## VII. FUTURE RESEARCH DIRECTIONS

**Lack of TTS Benchmark:** Researchers are focused on enhancing monolingual and multilingual voice cloning. However, objective comparisons are tricky due to testing set variation across papers, making it difficult to grasp the status of voice cloning. Establishing standardized methods for evaluation is crucial for further voice cloning development.

**Conversation-Style Speech Synthesis:** TTS research focuses mainly on synthesizing reading-style speech, disregarding casual tones, which can be helpful in conversational style generation. There are very few datasets aimed at addressing this type of task, so exploring conversational TTS could be an exciting area of research. AdaSpeech 3 [54] focuses on

TABLE X  
EVALUATION METRICS, WHERE EACH VARIABLE WILL CORRESPOND TO EITHER THE REFERENCE X OR THE PREDICTION  $X'$ .

Metric	Formula/Scale	Focus	Task
Objective			
MCD [279] ↓	$\frac{1}{T} \sum_{t=0}^{T-1} \sqrt{\sum_{k=1}^K (c_{t,k} - c'_{t,k})^2}$	Speech Quality	Measures spectrum similarity, where $c_{t,k}$ is the k-th MFCC of t-th time frame from audio waveform.
CER/WER ↓	Insertions+Deletions+Substitutions All characters/Words	Speech Quality	Measures CER/WER of transcribed speech using a pre-trained ASR and the reference text.
SECS ↑	$\frac{\text{Embeddings} \cdot \text{Embeddings}'}{\ \text{Embeddings}\  \times \ \text{Embeddings}'\ }$	Speaker Similarity	Measures the cosine similarity between the speaker's generated embedding and the ground truth.
SV-EER [280] ↓	$\frac{\text{False Accept (FA) Errors}}{\text{Imposter Attempts}} = \frac{\text{False Reject (FR) Errors}}{\text{Legitimate Attempts}}$	Speaker Similarity	Measures the performance of speaker verification systems, where False-Accuracy Rate (FAR) = False-Rejection Rate (FRR).
GPE [281] ↓	$\frac{\sum_t [  p_t - p'_t  > 0.2p_t ] [v_t] [v'_t]}{\sum_t [v_t] [v'_t]}$	Speech Expression	Voice Frame percentage with at least 20% from reference. $p_t$ is the pitch signal from audio, and $v_t$ is voicing decision.
VDE [281] ↓	$\frac{\sum_{t=0}^{T-1} [v_t \neq v'_t]}{T}$	Speech Expression	Frame percentage with incorrect voicing decisions. $v_t$ is voicing decisions audio, where T is the total frames.
FFE [282] ↓	$\frac{\sum_{t=0}^{T-1} [ p_t - p'_t  > 0.2p_t] [v_t] [v'_t] + [v_t \neq v'_t]}{T}$	Speech Expression	Percentage of frames where an error either by VDE or GPE occurs.
RTF ↓	NA	Inference Speed	The time in seconds that the device requires to synthesize one second of the waveform.
Subjective			
MOS ↑	5-point scale	NA	Average rating across evaluators following a point-scale. This is used to evaluate naturalness, similarity, quality, and intelligibility.
CMOS ↑	7-point scale	NA	The evaluator rates preference on a 7-point scale (-3, 3), where positive values indicate improvement while negative indicates a decline.
MUSHRA [283] ↑	100-point scale	NA	Evaluator rates speech, including reference, generated, hidden, and anchor signals with access to all signals for consistent comparison.
AB preference test ↑	Model A %: Model B %: No Preference %	NA	Evaluator compares synthesized speech of two methods. If more than two methods, this method can be called ABX preference test.

spontaneous speech synthesis, mimicking how people interact in conversations or a podcast setting.

**Fine-grained Disentanglement:** Representation disentanglement is a crucial technique for achieving fine-grained control over TTS systems. Even though most work concentrates on content-timbre disentanglement, there is a critical need for enhanced control over prosodic features for manipulating voice expression effectively. The ability to replicate emotions and expressions can be a valuable asset for many HCI applications.

### VIII. CONCLUSION

Voice cloning research has increased rapidly, driven by the industrial interest in digitization and digital twins. This survey systematically examines the advancements made in synthetic voice generation, specifically voice cloning. We categorize the methods into four main areas: speaker adaptation, few-shot voice cloning, zero-shot voice cloning, and multilingual voice cloning. We then classify those works based on their contributions and the challenges they tackle to enhance voice cloning performance. We also explore speech-related datasets and the evaluation metrics used to compare voice cloning algorithms. Understanding these methods is essential for enhancing their performance and developing algorithms for their detection, thereby mitigating the risks of audio deepfake generation.

### REFERENCES

- [1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, ‘A survey on neural speech synthesis’, arXiv preprint arXiv:2106.15561, 2021.
- [2] A. Kadam, S. Rane, A. K. Mishra, S. K. Sahu, S. Singh, and S. K. Pathak, ‘A Survey of Audio Synthesis and Lip-syncing for Synthetic Video Generation’, EAI Endorsed Transactions on Creative Technologies, vol. 8, no. 28, pp. e2–e2, 2021.
- [3] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, ‘Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward’, Applied intelligence, vol. 53, no. 4, pp. 3974–4026, 2023.
- [4] Z. Khanjani, G. Watson, and V. P. Janeja, ‘Audio deepfakes: A survey’, Frontiers in Big Data, vol. 5, p. 1001063, 2023.
- [5] S. C. Ramu, D. Saxena, and V. Mali, ‘A Survey on Voice Cloning and Automated Video Dubbing Systems’, in 2024 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), 2024, pp. 1–5.
- [6] L. Zhao and F. Chen, ‘Research on voice cloning with a few samples’, in 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA), 2020, pp. 323–328.
- [7] H.-T. Luong and J. Yamagishi, ‘Nautilus: a versatile voice cloning system’, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2967–2981, 2020.
- [8] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, ‘Neural voice cloning with a few samples’, Advances in neural information processing systems, vol. 31, 2018.
- [9] J. Cong, S. Yang, L. Xie, G. Yu, and G. Wan, ‘Data Efficient Voice Cloning from Noisy Samples with Domain Adversarial Training’, in Proc. Interspeech 2020, 2020, pp. 811–815.
- [10] Z. Shaheen, T. Sadekova, Y. Matveeva, A. Shirshova, and M. Kudinov, ‘Exploiting Emotion Information in Speaker Embeddings for Expressive Text-to-Speech’, in INTERSPEECH 2023, 2023, pp. 2038–2042.
- [11] A. R. Mandeel, M. S. Al-Radhi, and T. G. Csapó, ‘Enhancing End-to-End Speech Synthesis by Modeling Interrogative Sentences with Speaker Adaptation’, in 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2023, pp. 158–163.
- [12] S.-F. Huang, C.-J. Lin, D.-R. Liu, Y.-C. Chen, and H.-Y. Lee, ‘Meta-tts: Meta-learning for few-shot speaker adaptive text-to-speech’, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1558–1571, 2022.
- [13] S.-F. Huang, C.-P. Chen, Z.-S. Chen, Y.-P. Tsai, and H.-Y. Lee, ‘Personalized Lightweight Text-to-Speech: Voice Cloning with Adaptive Structured Pruning’, in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.
- [14] Y. Nakai, Y. Saito, K. Udagawa, and H. Saruwatari, ‘Multi-Task Adversarial Training Algorithm for Multi-Speaker Neural Text-to-Speech’, in 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2022, pp. 743–748.
- [15] Y. Zhang et al., ‘Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning’, in Interspeech 2019, 2019, pp. 2080–2084.
- [16] C.-M. Chien and H.-Y. Lee, ‘Hierarchical prosody modeling for non-autoregressive speech synthesis’, in 2021 IEEE Spoken Language Technology Workshop (SLT), 2021, pp. 446–453.
- [17] J. Shen et al., ‘Natural tts synthesis by conditioning wavenet on mel spectrogram predictions’, in 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2018, pp. 4779–4783.

- [18] Y. Wang et al., ‘Tacotron: Towards End-to-End Speech Synthesis’, in Interspeech 2017, 2017, pp. 4006–4010.
- [19] D. Griffin and J. Lim, ‘Signal estimation from modified short-time Fourier transform’, IEEE Transactions on acoustics, speech, and signal processing, vol. 32, no. 2, pp. 236–243, 1984.
- [20] S. Dieleman et al., ‘Wavenet: A generative model for raw audio’, arXiv preprint arXiv:1609.03499, vol. 12, 2016.
- [21] Y. Ren et al., ‘FastSpeech 2: Fast and High-Quality End-to-End Text to Speech’, in International Conference on Learning Representations, 2021.
- [22] Y. Ren et al., ‘FastSpeech: fast, robust and controllable text to speech’, in Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA: Curran Associates Inc., 2019.
- [23] J. Kim, J. Kong, and J. Son, ‘Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech’, in International Conference on Machine Learning, 2021, pp. 5530–5540.
- [24] C. Wang et al., ‘Neural codec language models are zero-shot text to speech synthesizers’, arXiv preprint arXiv:2301.02111, 2023.
- [25] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, ‘High Fidelity Neural Audio Compression’, Transactions on Machine Learning Research, 2023.
- [26] T. B. Brown et al., ‘Language models are few-shot learners’, in Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 2020.
- [27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, ‘wav2vec 2.0: A framework for self-supervised learning of speech representations’, Advances in neural information processing systems, vol. 33, pp. 12449–12460, 2020.
- [28] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, ‘Hubert: Self-supervised speech representation learning by masked prediction of hidden units’, IEEE/ACM transactions on audio, speech, and language processing, vol. 29, pp. 3451–3460, 2021.
- [29] A. Bapna et al., ‘SLAM: A unified encoder for speech and language modeling via speech-text joint pre-training’, arXiv preprint arXiv:2110.10329, 2021.
- [30] J. Chen, M. Ma, R. Zheng, and L. Huang, ‘Mam: Masked acoustic modeling for end-to-end speech-to-text translation’, arXiv preprint arXiv:2010.11445, 2020.
- [31] R. Zheng, J. Chen, M. Ma, and L. Huang, ‘Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation’, in International Conference on Machine Learning, 2021, pp. 12736–12746.
- [32] H. Bai, R. Zheng, J. Chen, M. Ma, X. Li, and L. Huang, ‘A3T: Alignment-Aware Acoustic and Text Pretraining for Speech Synthesis and Editing’, in International Conference on Machine Learning, 2022, pp. 1399–1411.
- [33] I. Goodfellow et al., ‘Generative adversarial nets’, Advances in neural information processing systems, vol. 27, 2014.
- [34] C. Finn, P. Abbeel, and S. Levine, ‘Model-agnostic meta-learning for fast adaptation of deep networks’, in International conference on machine learning, 2017, pp. 1126–1135.
- [35] J. Zaïdi, H. Seuté, B. van Niekerk, and M.-A. Carboneau, ‘Daft-Exprt: Cross-Speaker Prosody Transfer on Any Text for Expressive Speech Synthesis’, in Interspeech 2022, 2022, pp. 4591–4595.
- [36] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, ‘Film: Visual reasoning with a general conditioning layer’, in Proceedings of the AAAI conference on artificial intelligence, 2018, vol. 32.
- [37] N. Ellinas et al., ‘Controllable speech synthesis by learning discrete phoneme-level prosodic representations’, Speech Communication, vol. 146, pp. 22–31, 2023.
- [38] C. Du, Y. Guo, X. Chen, and K. Yu, ‘Speaker adaptive text-to-speech with timbre-normalized vector-quantized feature’, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.
- [39] Y. Fan, Y. Qian, F. K. Soong, and L. He, ‘Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis’, in 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2015, pp. 4475–4479.
- [40] Y. Zhao, D. Saito, and N. Minematsu, ‘Speaker Representations for Speaker Adaptation in Multiple Speakers’ BLSTM-RNN-Based Speech Synthesis’, in Interspeech 2016, 2016, pp. 2268–2272.
- [41] R. Doddipatla, N. Braunschweiler, and R. Maia, ‘Speaker Adaptation in DNN-Based Speech Synthesis Using d-Vectors’, in Interspeech 2017, 2017, pp. 3404–3408.
- [42] Z. Cai, C. Zhang, and M. Li, ‘From Speaker Verification to Multi-speaker Speech Synthesis, Deep Transfer with Feedback Constraint’, in Interspeech 2020, 2020, pp. 3974–3978.
- [43] W. Cai, J. Chen, J. Zhang, and M. Li, ‘On-the-fly data loader and utterance-level aggregation for speaker and language recognition’, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1038–1051, 2020.
- [44] G. Ruggiero, E. Zovato, L. Di Caro, and V. Pollet, ‘Voice cloning: a multi-speaker text-to-speech synthesis approach based on transfer learning’, arXiv preprint arXiv:2102.05630, 2021.
- [45] Y. Jia et al., ‘Transfer learning from speaker verification to multispeaker text-to-speech synthesis’, Advances in neural information processing systems, vol. 31, 2018.
- [46] J. Yang, J.-S. Bae, T. Bak, Y.-I. Kim, and H.-Y. Cho, ‘GANSpeech: Adversarial Training for High-Fidelity Multi-Speaker Speech Synthesis’, in Interspeech 2021, 2021, pp. 2202–2206.
- [47] B. Chen, C. Du, and K. Yu, ‘Neural fusion for voice cloning’, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1993–2001, 2022.
- [48] L. Wadoux, N. Barbot, J. Chevrel, and D. Lalive, ‘Voice Cloning for Voice Disorders: Impact of Phonetic Content’, in International Conference on Text, Speech, and Dialogue, 2023, pp. 293–303.
- [49] W.-N. Hsu et al., ‘Hierarchical Generative Modeling for Controllable Speech Synthesis’, in International Conference on Learning Representations, 2019.
- [50] Y. Wang et al., ‘Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis’, in International conference on machine learning, 2018, pp. 5180–5189.
- [51] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, ‘Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder’, in Interspeech 2018, 2018, pp. 3067–3071.
- [52] P. Neekhara, J. Li, and B. Ginsburg, ‘Adapting TTS models for new speakers using transfer learning’, arXiv preprint arXiv:2110.05798, 2021.
- [53] M. Chen et al., ‘AdaSpeech: Adaptive Text to Speech for Custom Voice’, in International Conference on Learning Representations, 2021.
- [54] Y. Yan et al., ‘Adaptive Text to Speech for Spontaneous Style’, in Interspeech 2021, 2021, pp. 4668–4672.
- [55] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Y. Zhang, ‘Hi-Fi Multi-Speaker English TTS Dataset’, in Interspeech 2021, 2021, pp. 2776–2780.
- [56] K. Inoue, S. Hara, M. Abe, T. Hayashi, R. Yamamoto, and S. Watanabe, ‘Semi-supervised speaker adaptation for end-to-end speech synthesis with pretrained models’, in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7634–7638.
- [57] H. Zhang and Y. Lin, ‘Unsupervised Learning for Sequence-to-Sequence Text-to-Speech for Low-Resource Languages’, in Interspeech 2020, 2020, pp. 3161–3165.
- [58] A. Van Den Oord, O. Vinyals, and Others, ‘Neural discrete representation learning’, Advances in neural information processing systems, vol. 30, 2017.
- [59] Y. Yan et al., ‘Adaspeech 2: Adaptive text to speech with untranscribed data’, in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6613–6617.
- [60] K. Klapsas et al., ‘Self supervised learning for robust voice cloning’, in Interspeech 2022, 2022, pp. 4935–4939.
- [61] J.-B. Grill et al., ‘Bootstrap your own latent-a new approach to self-supervised learning’, Advances in neural information processing systems, vol. 33, pp. 21271–21284, 2020.
- [62] J. Shen et al., ‘Non-attentive tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling’, arXiv preprint arXiv:2010.04301, 2020.
- [63] K. Inoue, S. Hara, and M. Abe, ‘Module Comparison Of Transformer-Tts For Speaker Adaptation Based On Fine-Tuning’, in 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2020, pp. 826–830.
- [64] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, ‘Neural speech synthesis with transformer network’, in Proceedings of the AAAI conference on artificial intelligence, 2019, vol. 33, pp. 6706–6713.
- [65] K. Song et al., ‘AdaVITS: Tiny VITS for low computing resource speaker adaptation’, in 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2022, pp. 319–323.
- [66] Y. Li, R. Bhardwaj, A. Mehrish, B. Cheng, and S. Poria, ‘HYPERTTS: Parameter Efficient Adaptation in Text to Speech Using Hypernetworks’, in Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 8516–8527.

- [67] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, ‘Generalized end-to-end loss for speaker verification’, in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4879–4883.
- [68] Y. Wu, H. Zhao, X. Liang, and Y. Sun, ‘VStyclone: Real-time Chinese voice style clone’, Computers and Electrical Engineering, vol. 105, p. 108534, 2023.
- [69] P. Cheng and Z. Ling, ‘Speaker Adaption with Intuitive Prosodic Features for Statistical Parametric Speech Synthesis’, in Proceedings of the 6th International Conference on Digital Signal Processing, 2022, pp. 187–193.
- [70] T. Wang, J. Tao, R. Fu, J. Yi, Z. Wen, and R. Zhong, ‘Spoken Content and Voice Factorization for Few-Shot Speaker Adaptation’, in Proc. Interspeech 2020, 2020, pp. 796–800.
- [71] S. Choi, S. Han, D. Kim, and S. Ha, ‘Attentron: Few-Shot Text-to-Speech Utilizing Attention-Based Variable-Length Embedding’, in Proc. Interspeech 2020, 2020, pp. 2007–2011.
- [72] E. Cooper et al., ‘Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings’, in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6184–6188.
- [73] H. Lu, X. Wu, H. Guo, S. Liu, Z. Wu, and H. Meng, ‘Unifying One-Shot Voice Conversion and Cloning with Disentangled Speech Representations’, in ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 11141–11145.
- [74] H. Lu, D. Wang, X. Wu, Z. Wu, X. Liu, and H. Meng, ‘Disentangled speech representation learning for one-shot cross-lingual voice conversion using  $\beta$ -vae’, in 2022 IEEE Spoken Language Technology Workshop (SLT), 2023, pp. 814–821.
- [75] J. Lian, C. Zhang, and D. Yu, ‘Robust disentangled variational speech representation learning for zero-shot voice conversion’, in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6572–6576.
- [76] T. Sadekova, V. Gogoryan, I. Vovk, V. Popov, M. Kudinov, and J. Wei, ‘A Unified System for Voice Cloning and Voice Conversion through Diffusion Probabilistic Modeling’, in Proc. Interspeech 2022, 2022, pp. 3003–3007.
- [77] H. Zhang and Y. Lin, ‘Improve few-shot voice cloning using multi-modal learning’, in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8317–8321.
- [78] S.-H. Lee, S.-B. Kim, J.-H. Lee, E. Song, M.-J. Hwang, and S.-W. Lee, ‘HierSpeech: Bridging the gap between text and speech by hierarchical variational inference using self-supervised representations for speech synthesis’, Advances in Neural Information Processing Systems, vol. 35, pp. 16624–16636, 2022.
- [79] A. Babu et al., ‘XLS-R: Self-supervised cross-lingual speech representation learning at scale’, arXiv preprint arXiv:2111. 09296, 2021.
- [80] S. Kim, H. Kim, and S. Yoon, ‘Guided-tts 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data’, arXiv preprint arXiv:2205. 15370, 2022.
- [81] H. Kim, S. Kim, and S. Yoon, ‘Guided-tts: A diffusion model for text-to-speech via classifier guidance’, in International Conference on Machine Learning, 2022, pp. 11119–11133.
- [82] J. Ho, A. Jain, and P. Abbeel, ‘Denoising diffusion probabilistic models’, Advances in neural information processing systems, vol. 33, pp. 6840–6851, 2020.
- [83] H. Kim, S. Kim, J. Yeom, and S. Yoon, ‘UnitSpeech: Speaker-adaptive Speech Synthesis with Untranscribed Data’, in Proc. INTERSPEECH 2023, 2023, pp. 3038–3042.
- [84] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, ‘VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop’, in International Conference on Learning Representations, 2018.
- [85] Z. Kons, S. Shechtman, A. Sorin, C. Rabinovitz, and R. Hoory, ‘High Quality, Lightweight and Adaptable TTS Using LPCNet’, in Proc. Interspeech 2019, 2019, pp. 176–180.
- [86] N. Morioka, H. Zen, N. Chen, Y. Zhang, and Y. Ding, ‘Residual adapters for few-shot text-to-speech speaker adaptation’, arXiv preprint arXiv:2210. 15868, 2022.
- [87] Y. Jia, H. Zen, J. Shen, Y. Zhang, and Y. Wu, ‘PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS’, in Interspeech 2021, 2021, pp. 151–155.
- [88] C.-P. Hsieh, S. Ghosh, and B. Ginsburg, ‘Adapter-Based Extension of Multi-Speaker Text-To-Speech Model for New Speakers’, in Proc. INTERSPEECH 2023, 2023, pp. 3028–3032.
- [89] A. Łaćucki, ‘Fastpitch: Parallel text-to-speech with pitch prediction’, in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6588–6592.
- [90] H. Chen and P. N. Garner, ‘Diffusion transformer for adaptive text-to-speech’, in 12th Speech Synthesis Workshop (SSW) 2023, 2023.
- [91] W. Peebles and S. Xie, ‘Scalable diffusion models with transformers’, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4195–4205.
- [92] 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, 2024.
- [93] Y. Deng, L. He, and F. Soong, ‘Modeling multi-speaker latent space to improve neural tts: Quick enrolling new speaker and enhancing premium voice’, arXiv preprint arXiv:1812. 05253, 2018.
- [94] J.-H. Kim, S.-H. Lee, J.-H. Lee, H.-G. Jung, and S.-W. Lee, ‘GC-TTS: Few-shot speaker adaptation with geometric constraints’, in 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2021, pp. 1172–1177.
- [95] H.-G. Jung and S.-W. Lee, ‘Few-shot learning with geometric constraints’, IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 11, pp. 4660–4672, 2020.
- [96] J. Lee and J.-H. Chang, ‘One-Shot Speaker Adaptation Based on Initialization by Generative Adversarial Networks for TTS’, in Proc. Interspeech 2022, 2022, pp. 2978–2982.
- [97] A. Gabryś et al., ‘Voice filter: Few-shot text-to-speech speaker adaptation using voice conversion as a post-processing module’, in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 7902–7906.
- [98] J. Xue, Y. Deng, Y. Han, Y. Li, J. Sun, and J. Liang, ‘ECAPA-TDNN for Multi-speaker Text-to-speech Synthesis’, in 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2022, pp. 230–234.
- [99] B. Desplanques, J. Thienpondt, and K. Demuynck, ‘ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification’, in Interspeech 2020, 2020, pp. 3830–3834.
- [100] H. Zhao, H. Liu, M. Cao, and Y. Liu, ‘Few-Shot Custom Speech Synthesis with Multi-Angle Fusion’, in 2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP), 2023, pp. 1609–1613.
- [101] Y. Chen et al., ‘Sample Efficient Adaptive Text-to-Speech’, in International Conference on Learning Representations, 2018.
- [102] H. B. Moss, V. Aggarwal, N. Prateek, J. González, and R. Barra-Chicote, ‘Boffin tts: Few-shot speaker adaptation by bayesian optimization’, in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7639–7643.
- [103] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, ‘Meta-stylespeech: Multi-speaker adaptive text-to-speech generation’, in International Conference on Machine Learning, 2021, pp. 7748–7759.
- [104] S. Liu, D. Su, and D. Yu, ‘Meta-voice: Fast few-shot style transfer for expressive voice cloning using meta learning’, arXiv preprint arXiv:2111. 07218, 2021.
- [105] Y. Zheng, X. Li, and L. Lu, ‘Investigation of fast and efficient methods for multi-speaker modeling and speaker adaptation’, in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6618–6622.
- [106] T. Wang et al., ‘Prosody and voice factorization for few-shot speaker adaptation in the challenge m2voc 2021’, in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 8603–8607.
- [107] R. Fu, J. Tao, Z. Wen, J. Yi, T. Wang, and C. Qiang, ‘Bi-level style and prosody decoupling modeling for personalized end-to-end speech synthesis’, in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6568–6572.
- [108] W. Song et al., ‘Dian: Duration informed auto-regressive network for voice cloning’, in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 8598–8602.
- [109] C.-M. Chien, J.-H. Lin, C.-Y. Huang, P.-C. Hsu, and H.-Y. Lee, ‘Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech’, in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 8588–8592.
- [110] D. Ke et al., ‘AdaptiveFormer: A Few-shot Speaker Adaptive Speech Synthesis Model based on FastSpeech2’, in 2022 13th International

- Symposium on Chinese Spoken Language Processing (ISCSLP), 2022, pp. 225–229.
- [111] R. Li, D. Pu, M. Huang, and B. Huang, ‘Unet-tts: Improving unseen speaker and style transfer in one-shot voice cloning’, in ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8327–8331.
- [112] O. Ronneberger, P. Fischer, and T. Brox, ‘U-net: Convolutional networks for biomedical image segmentation’, in Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18, 2015, pp. 234–241.
- [113] Q. Wu, Q. Shen, J. Luan, and Y. Wang, ‘Msdtron: a high-capability multi-speaker speech synthesis system for diverse data using characteristic information’, in ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6327–6331.
- [114] J. Wu, J. Luan, and Y. Wang, ‘LightClone: Speaker-guided Parallel Subnet Selection for Few-shot Voice Cloning’, in Proc. INTERSPEECH 2023, 2023, pp. 4868–4872.
- [115] X. Yuan, R. Feng, M. Ye, C. Tuo, and M. Zhang, ‘AdaVocoder: Adaptive Vocoder for Custom Voice’, in Proc. Interspeech 2022, 2022, pp. 1586–1590.
- [116] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, ‘Robust speech recognition via large-scale weak supervision’, in International conference on machine learning, 2023, pp. 28492–28518.
- [117] A. Gulati et al., ‘Conformer: Convolution-augmented Transformer for Speech Recognition’, in Interspeech 2020, 2020, pp. 5036–5040.
- [118] O. Kuchaiev et al., ‘Nemo: a toolkit for building ai applications using neural modules’, arXiv preprint arXiv:1909.09577, 2019.
- [119] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, ‘X-vectors: Robust dnn embeddings for speaker recognition’, in 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2018, pp. 5329–5333.
- [120] N. R. Koluguri, T. Park, and B. Ginsburg, ‘Titant: Neural model for speaker representation with 1d depth-wise separable convolutions and global context’, in ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8102–8106.
- [121] V. Popov et al., ‘Fast and Lightweight On-Device TTS with Tacotron2 and LPCNet’, in Interspeech 2020, 2020, pp. 220–224.
- [122] R. Huang, Y. Ren, J. Liu, C. Cui, and Z. Zhao, ‘Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech’, Advances in Neural Information Processing Systems, vol. 35, pp. 10970–10983, 2022.
- [123] W. Wang, Y. Song, and S. Jha, ‘Generalizable Zero-Shot Speaker Adaptive Speech Synthesis with Disentangled Representations’, in INTERSPEECH 2023, 2023, pp. 4454–4458.
- [124] J. Lian, C. Zhang, G. K. Anumanchipalli, and D. Yu, ‘Unsupervised tts acoustic modeling for tts with conditional disentangled sequential vae’, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.
- [125] Y. Jeon, Y. Kim, and G. G. Lee, ‘Enhancing Zero-Shot Multi-Speaker TTS with Negated Speaker Representations’, in Proceedings of the AAAI Conference on Artificial Intelligence, 2024, vol. 38, pp. 18336–18344.
- [126] Y. T. Chen, W. Li, and B. Tang, ‘MRMI-TTS: Multi-Reference Audios and Mutual Information Driven Zero-Shot Voice Cloning’, ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 23, no. 5, pp. 1–14, 2024.
- [127] B. J. Choi, M. Jeong, M. Kim, and N. S. Kim, ‘Variable-Length Speaker Conditioning in Flow-Based Text-to-Speech’, IEEE Signal Processing Letters, 2024.
- [128] Y. Ren et al., ‘Fewer-token neural speech codec with time-invariant codes’, in ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 12737–12741.
- [129] D. Paul, S. Mukherjee, Y. Pantazis, and Y. Stylianou, ‘A Universal Multi-Speaker Multi-Style Text-to-Speech via Disentangled Representation Learning Based on Rényi Divergence Minimization’, in Interspeech, 2021, pp. 3625–3629.
- [130] 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.
- [131] K. Fujita, T. Ashihara, H. Kanagawa, T. Moriya, and Y. Ijima, ‘Zero-shot text-to-speech synthesis conditioned using self-supervised speech representation model’, in 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), 2023, pp. 1–5.
- [132] J. Y. Lee et al., ‘Hierarchical timbre-cadence speaker encoder for zero-shot speech synthesis’, in Proc. Interspeech, 2023.
- [133] S.-H. Lee, H.-Y. Choi, S.-B. Kim, and S.-W. Lee, ‘Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis’, arXiv preprint arXiv:2311.12454, 2023.
- [134] Z. Jiang et al., ‘Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias’, arXiv preprint arXiv:2306.03509, 2023.
- [135] Z. Jiang et al., ‘Mega-TTS 2: Boosting Prompting Mechanisms for Zero-Shot Speech Synthesis’, in The Twelfth International Conference on Learning Representations, 2024.
- [136] E. Kharitonov et al., ‘Speak, read and prompt: High-fidelity text-to-speech with minimal supervision’, Transactions of the Association for Computational Linguistics, vol. 11, pp. 1703–1718, 2023.
- [137] X. Zhu et al., ‘Vec-tok speech: Speech vectorization and tokenization for neural speech generation’, arXiv preprint arXiv:2310.07246, 2023.
- [138] J. Xue, Y. Deng, Y. Han, Y. Gao, and Y. Li, ‘Improving Audio Codec-based Zero-Shot Text-to-Speech Synthesis with Multi-Modal Context and Large Language Model’, in Interspeech 2024, 2024, pp. 682–686.
- [139] J. Y. Lee, M. Jeong, M. Kim, J.-H. Lee, H.-Y. Cho, and N. S. Kim, ‘High Fidelity Text-to-Speech Via Discrete Tokens Using Token Transducer and Group Masked Language Model’, in Interspeech 2024, 2024, pp. 3445–3449.
- [140] S. Lei et al., ‘Improving Language Model-Based Zero-Shot Text-to-Speech Synthesis with Multi-Scale Acoustic Prompts’, in ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 12662–12666.
- [141] Z. Ju et al., ‘NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models’, in Forty-first International Conference on Machine Learning, 2024.
- [142] J. Li, D. Li, S. Savarese, and S. Hoi, ‘Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models’, in International conference on machine learning, 2023, pp. 19730–19742.
- [143] Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi, ‘Soundstorm: Efficient parallel audio generation’, arXiv preprint arXiv:2305.09636, 2023.
- [144] M. Kim, M. Jeong, B. J. Choi, S. Kim, J. Y. Lee, and N. S. Kim, ‘Utilizing Neural Transducers for Two-Stage Text-to-Speech via Semantic Token Prediction’, arXiv preprint arXiv:2401.01498, 2024.
- [145] M. Jeong, M. Kim, J. Y. Lee, and N. S. Kim, ‘Efficient Parallel Audio Generation Using Group Masked Language Modeling’, IEEE Signal Processing Letters, 2024.
- [146] Y. A. Li, C. Han, and N. Mesgarani, ‘Styletts: A style-based generative model for natural and diverse text-to-speech synthesis’, IEEE Journal of Selected Topics in Signal Processing, 2025.
- [147] 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’, Advances in Neural Information Processing Systems, vol. 36, 2024.
- [148] S. Chen et al., ‘Wavlm: Large-scale self-supervised pre-training for full stack speech processing’, IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1505–1518, 2022.
- [149] Y. A. Li, C. Han, X. Jiang, and N. Mesgarani, ‘Phoneme-level bert for enhanced prosody of text-to-speech with grapheme predictions’, in ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.
- [150] V. Pankov et al., ‘DINO-VITS: Data-Efficient Zero-Shot TTS with Self-Supervised Speaker Verification Loss for Noise Robustness’, in Interspeech 2024, 2024, pp. 697–701.
- [151] Y. Liu et al., ‘Multilingual denoising pre-training for neural machine translation’, Transactions of the Association for Computational Linguistics, vol. 8, pp. 726–742, 2020.
- [152] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, ‘CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking’, in Interspeech 2023, 2023, pp. 5301–5305.
- [153] M. Kang, W. Han, S. J. Hwang, and E. Yang, ‘ZET-Speech: Zero-shot adaptive Emotion-controllable Text-to-Speech Synthesis with Diffusion and Style-based Models’, in Proc. INTERSPEECH 2023, 2023, pp. 4339–4343.
- [154] D. Paul, Y. Pantazis, and Y. Stylianou, ‘Speaker Conditional WaveRNN: Towards Universal Neural Vocoder for Unseen Speaker and Recording Conditions’, in Interspeech 2020, 2020, pp. 235–239.

- [155] E. Casanova et al., ‘SC-GlowTTS: An Efficient Zero-Shot Multi-Speaker Text-To-Speech Model’, in *Interspeech 2021*, 2021, pp. 3645–3649.
- [156] J. Kim, S. Kim, J. Kong, and S. Yoon, ‘Glow-tts: A generative flow for text-to-speech via monotonic alignment search’, *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.
- [157] J. S. Chung et al., ‘In Defence of Metric Learning for Speaker Recognition’, in *Interspeech 2020*, 2020, pp. 2977–2981.
- [158] J. Vainer and O. Dušek, ‘SpeedySpeech: Efficient Neural Speech Synthesis’, in *Interspeech 2020*, 2020, pp. 3575–3579.
- [159] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, ‘Language modeling with gated convolutional networks’, in *International conference on machine learning*, 2017, pp. 933–941.
- [160] S. Ogun, V. Colotte, and E. Vincent, ‘Stochastic Pitch Prediction Improves the Diversity and Naturalness of Speech in Glow-TTS’, in *INTERSPEECH 2023*, 2023, pp. 4878–4882.
- [161] H. Yoon, C. Kim, E. Song, H.-W. Yoon, and H.-G. Kang, ‘Pruning Self-Attention for Zero-Shot Multi-Speaker Text-to-Speech’, in *INTERSPEECH 2023*, 2023, pp. 4299–4303.
- [162] F. Lux, J. Koch, and N. T. Vu, ‘Exact prosody cloning in zero-shot multispeaker text-to-speech’, in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 962–969.
- [163] Y. Song, Z. Chen, X. Wang, Z. Ma, G. Yang, and X. Chen, ‘TacoLM: GaTed Attention Equipped Codec Language Model are Efficient Zero-Shot Text to Speech Synthesizers’, in *Interspeech 2024*, 2024, pp. 4433–4437.
- [164] X. Li et al., ‘CM-TTS: Enhancing Real Time Text-to-Speech Synthesis Efficiency through Weighted Samplers and Consistency Models’, in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 3777–3794.
- [165] Z. Ye et al., ‘Flashspeech: Efficient zero-shot speech synthesis’, in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 6998–7007.
- [166] K. Fujita, T. Asihara, M. Delcroix, and Y. Ijima, ‘Lightweight Zero-shot Text-to-Speech with Mixture of Adapters’, in *Interspeech 2024*, 2024, pp. 692–696.
- [167] D. Yang, D. Wang, H. Guo, X. Chen, X. Wu, and H. Meng, ‘SimpleSpeech: Towards Simple and Efficient Text-to-Speech with Scalar Latent Transformer Diffusion Models’, in *Interspeech 2024*, 2024, pp. 4398–4402.
- [168] T. Lemerle, N. Obin, and A. Roebel, ‘Small-E: Small Language Model with Linear Attention for Efficient Speech Synthesis’, in *Interspeech 2024*, 2024, pp. 3420–3424.
- [169] S. Kim et al., ‘P-flow: a fast and data-efficient zero-shot TTS through speech prompting’, *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [170] Y. Lei, S. Yang, J. Cong, L. Xie, and D. Su, ‘Glow-WaveGAN 2: High-quality Zero-shot Text-to-speech Synthesis and Any-to-any Voice Conversion’, in *Interspeech 2022*, 2022, pp. 2563–2567.
- [171] J. Cong, S. Yang, L. Xie, and D. Su, ‘Glow-WaveGAN: Learning Speech Representations from GAN-Based Variational Auto-Encoder for High Fidelity Flow-Based Speech Synthesis’, in *Interspeech 2021*, 2021, pp. 2182–2186.
- [172] H. Liu, T. Wang, R. Fu, J. Yi, Z. Wen, and J. Tao, ‘Unifyspeech: A unified framework for zero-shot text-to-speech and voice conversion’, *arXiv preprint arXiv:2301.03801*, 2023.
- [173] H.-S. Choi, J. Yang, J. Lee, and H. Kim, ‘NANSY++: Unified Voice Synthesis with Neural Analysis and Synthesis’, in *The Eleventh International Conference on Learning Representations*, 2023.
- [174] D. Ke et al., ‘Voicifier-LN: An Novel Approach to Elevate the Speaker Similarity for General Zero-shot Multi-Speaker TTS’, in *Proceedings of the 2022 5th International Conference on Artificial Intelligence and Pattern Recognition*, 2022, pp. 950–955.
- [175] M. Zhang, X. Zhou, Z. Wu, and H. Li, ‘Towards zero-shot multi-speaker multi-accent text-to-speech synthesis’, *IEEE Signal Processing Letters*, 2023.
- [176] X. Wang et al., ‘Speechx: Neural codec language model as a versatile speech transformer’, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [177] P. Peng, P.-Y. Huang, S.-W. Li, A. Mohamed, and D. Harwath, ‘VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild’, in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 12442–12462.
- [178] N. Kanda et al., ‘Making Flow-Matching-Based Zero-Shot Text-to-Speech Laugh as You Like’, *arXiv preprint arXiv:2402.07383*, 2024.
- [179] M. Le et al., ‘Voicebox: Text-guided multilingual universal speech generation at scale’, *Advances in neural information processing systems*, vol. 36, 2024.
- [180] J. Lee, J. Kim, and J.-H. Chang, ‘Zero-Shot Voice Cloning Using Variational Embedding with Attention Mechanism’, in *2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*, 2021, pp. 344–348.
- [181] S.-H. Lee, H.-W. Yoon, H.-R. Noh, J.-H. Kim, and S.-W. Lee, ‘Multispectrogram: High-diversity and high-fidelity spectrogram generation with adversarial style combination for speech synthesis’, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 13198–13206.
- [182] B. J. Choi, M. Jeong, M. Kim, S. H. Mun, and N. S. Kim, ‘Adversarial speaker-consistency learning using untranscribed speech data for zero-shot multi-speaker text-to-speech’, in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 1708–1712.
- [183] B. Schnell and P. N. Garner, ‘Investigating a neural all pass warp in modern TTS applications’, *Speech Communication*, vol. 138, pp. 26–37, 2022.
- [184] Z. Lyu and J. Zhu, ‘Enriching Style Transfer in multi-scale control based personalized end-to-end speech synthesis’, in *2022 12th International Conference on Information Science and Technology (ICIST)*, 2022, pp. 114–119.
- [185] A. Gorodetskii and I. Ozhiganov, ‘Zero-shot long-form voice cloning with dynamic convolution attention’, *arXiv preprint arXiv:2201.10375*, 2022.
- [186] J.-H. Lee, S.-H. Lee, J.-H. Kim, and S.-W. Lee, ‘PVAE-TTS: Adaptive text-to-speech via progressive style adaptation’, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6312–6316.
- [187] T. Cory and R. Iqbal, ‘Multi-scale speaker vectors for zero-shot speech synthesis’, in *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2022, pp. 496–501.
- [188] H. Siuzdak, P. Dura, P. van Rijn, and N. Jacoby, ‘WavThruVec: Latent speech representation as intermediate features for neural speech synthesis’, in *Interspeech 2022*, 2022, pp. 833–837.
- [189] M. Kim, M. Jeong, B. J. Choi, S. Ahn, J. Y. Lee, and N. S. Kim, ‘Transfer Learning Framework for Low-Resource Text-to-Speech using a Large-Scale Unlabeled Speech Corpus’, in *Interspeech 2022*, 2022, pp. 788–792.
- [190] Y. Wu et al., ‘AdaSpeech 4: Adaptive Text to Speech in Zero-Shot Scenarios’, in *Interspeech 2022*, 2022, pp. 2568–2572.
- [191] S. E. Eskimez et al., ‘Total-Duration-Aware Duration Modeling for Text-to-Speech Systems’, in *Interspeech 2024*, 2024, pp. 2290–2294.
- [192] C. Du et al., ‘VALL-T: Decoder-Only Generative Transducer for Robust and Decoding-Controllable Text-to-Speech’, *arXiv preprint arXiv:2401.14321*, 2024.
- [193] B. J. Choi, M. Jeong, J. Y. Lee, and N. S. Kim, ‘SNAC: Speaker-normalized affine coupling layer in flow-based architecture for zero-shot multi-speaker text-to-speech’, *IEEE Signal Processing Letters*, vol. 29, pp. 2502–2506, 2022.
- [194] H. Yoon, C. Kim, S. Um, H.-W. Yoon, and H.-G. Kang, ‘SC-CNN: Effective speaker conditioning method for zero-shot multi-speaker text-to-speech systems’, *IEEE Signal Processing Letters*, vol. 30, pp. 593–597, 2023.
- [195] 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, p. 28, 2024.
- [196] Y. Jiang, T. Li, F. Yang, L. Xie, M. Meng, and Y. Wang, ‘Towards Expressive Zero-Shot Speech Synthesis with Hierarchical Prosody Modeling’, in *Interspeech 2024*, 2024, pp. 2300–2304.
- [197] C.-W. Bang and C. Chun, ‘Effective Zero-Shot Multi-Speaker Text-to-Speech Technique Using Information Perturbation and a Speaker Encoder’, *Sensors*, vol. 23, no. 23, p. 9591, 2023.
- [198] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, ‘Grad-tts: A diffusion probabilistic model for text-to-speech’, in *International Conference on Machine Learning*, 2021, pp. 8599–8608.
- [199] M. Kang, D. Min, and S. J. Hwang, ‘Grad-stylespeech: Any-speaker adaptive text-to-speech synthesis with diffusion models’, in *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [200] T. Li et al., ‘U-Style: Cascading U-Nets With Multi-Level Speaker and Style Modeling for Zero-Shot Voice Cloning’, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

- [201] Z. Liu, Y. Guo, and K. Yu, ‘Diffvoice: Text-to-speech with latent diffusion’, in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.
- [202] K. Shen et al., ‘NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers’, in The Twelfth International Conference on Learning Representations, 2024.
- [203] D. Xin et al., ‘Rall-e: Robust codec language modeling with chain-of-thought prompting for text-to-speech synthesis’, arXiv preprint arXiv:2404. 03204, 2024.
- [204] K. Zhou et al., ‘Phonetic Enhanced Language Modeling for Text-to-Speech Synthesis’, in Interspeech 2024, 2024, pp. 3440–3444.
- [205] S. Chen et al., ‘VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers’, arXiv preprint arXiv:2406. 05370, 2024.
- [206] N. Goswami and T. Harada, ‘SATTS: Speaker Attractor Text to Speech, Learning to Speak by Learning to Separate’, in Interspeech 2022, 2022, pp. 1203–1207.
- [207] K. Azizah, ‘Zero-Shot Voice Cloning Text-to-Speech for Dysphonia Disorder Speakers’, IEEE Access, 2024.
- [208] Y. Zhou et al., ‘Content-Dependent Fine-Grained Speaker Embedding for Zero-Shot Speaker Adaptation in Text-to-Speech Synthesis’, in Interspeech 2022, 2022, pp. 2573–2577.
- [209] G. Zhang et al., ‘iEmoTTS: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody and timbre’, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 1693–1705, 2023.
- [210] H. Li, X. Zhu, L. Xue, Y. Song, Y. Chen, and L. Xie, ‘SponTTS: modeling and transferring spontaneous style for TTS’, in ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 12171–12175.
- [211] K. Fujita et al., ‘Noise-robust zero-shot text-to-speech synthesis conditioned on self-supervised speech-representation model with adapters’, in ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 11471–11475.
- [212] N. T. Dat, L. Q. Tuong, and N. D. Dung, ‘Calib-StyleSpeech: A Zero-Shot Approach in Voice Cloning of High Adaptive Text to Speech System with Imbalanced Dataset’, in International Conference on Artificial Intelligence and Big Data in Digital Era, 2021, pp. 61–72.
- [213] P. P. Ngoc, C. T. Quang, and M. L. 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, 2023.
- [214] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, ‘Unsupervised Cross-Lingual Representation Learning for Speech Recognition’, in Interspeech 2021, 2021, pp. 2426–2430.
- [215] M. Ravanelli et al., ‘Open-Source Conversational AI with SpeechBrain 1.0’, arXiv [cs.LG]. 2024.
- [216] M. Ravanelli et al., ‘SpeechBrain: A General-Purpose Speech Toolkit’, arXiv [eess.AS]. 2021.
- [217] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, ‘Diffssinger: Singing voice synthesis via shallow diffusion mechanism’, in Proceedings of the AAAI conference on artificial intelligence, 2022, vol. 36, pp. 11020–11028.
- [218] S. Liu, D. Su, and D. Yu, ‘Diffgan-tts: High-fidelity and efficient text-to-speech with denoising diffusion gans’, arXiv preprint arXiv:2201. 11972, 2022.
- [219] G. Eren and The Coqui TTS Team, Coqui TTS. 2021.
- [220] Y. Gao, N. Morioka, Y. Zhang, and N. Chen, ‘E3 tts: Easy end-to-end diffusion-based text to speech’, in 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2023, pp. 1–8.
- [221] D. Xin, Y. Saito, S. Takamichi, T. Koriyama, and H. Saruwatari, ‘Cross-Lingual Speaker Adaptation Using Domain Adaptation and Speaker Consistency Loss for Text-To-Speech Synthesis’, in Interspeech, 2021, pp. 1614–1618.
- [222] M. Chen et al., ‘Cross-Lingual, Multi-Speaker Text-To-Speech Synthesis Using Neural Speaker Embedding’, in Interspeech, 2019, pp. 2105–2109.
- [223] C. Li et al., ‘Deep speaker: an end-to-end neural speaker embedding system’, arXiv preprint arXiv:1705. 02304, 2017.
- [224] Z. Zhang, Q. Tian, H. Lu, L.-H. Chen, and S. Liu, ‘Adadurian: Few-shot adaptation for neural text-to-speech with durian’, arXiv preprint arXiv:2005. 05642, 2020.
- [225] J. Wang, X. Zhang, A. Sun, N. Cheng, and J. Xiao, ‘FastGraphTTS: An Ultrafast Syntax-Aware Speech Synthesis Framework’, in 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), 2023, pp. 905–912.
- [226] Y. Zheng, Z. Zhang, X. Li, W. Su, and L. Lu, ‘Zero-shot cross-lingual transfer using multi-stream encoder and efficient speaker representation’, in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8027–8031.
- [227] B. Zhao, X. Zhang, J. Wang, N. Cheng, and J. Xiao, ‘nnspeech: Speaker-guided conditional variational autoencoder for zero-shot multi-speaker text-to-speech’, in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 4293–4297.
- [228] Z. Cai, Y. Yang, and M. Li, ‘Cross-lingual multi-speaker speech synthesis with limited bilingual training data’, Computer Speech & Language, vol. 77, p. 101427, 2023.
- [229] Z. Zhang et al., ‘Speak foreign languages with your own voice: Cross-lingual neural codec language modeling’, arXiv preprint arXiv:2303. 03926, 2023.
- [230] P. Anastassiou et al., ‘Seed-TTS: A Family of High-Quality Versatile Speech Generation Models’, arXiv preprint arXiv:2406. 02430, 2024.
- [231] S. Ji, Z. Jiang, H. Wang, J.-L. Zuo, and Z. Zhao, ‘MobileSpeech: A Fast and High-Fidelity Framework for Mobile Zero-Shot Text-to-Speech’, in Annual Meeting of the Association for Computational Linguistics, 2024.
- [232] Y. Wang et al., ‘MaskGCT: Zero-Shot Text-to-Speech with Masked Generative Codec Transformer’, in The Thirteenth International Conference on Learning Representations, 2025.
- [233] J.-S. Bae et al., ‘Latent Filling: Latent Space Data Augmentation for Zero-Shot Speech Synthesis’, in ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 11166–11170.
- [234] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, ‘Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone’, in International Conference on Machine Learning, 2022, pp. 2709–2720.
- [235] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, ‘Clova baseline system for the voxceleb speaker recognition challenge 2020’, arXiv preprint arXiv:2009. 14153, 2020.
- [236] N. Shah, S. Kosgi, V. Tambrahalli, S. Neha, A. Nelakanti, and V. Gandhi, ‘ParrotTTS: Text-to-speech synthesis exploiting disentangled self-supervised representations’, in Findings of the Association for Computational Linguistics: EACL 2024, 2024, pp. 79–91.
- [237] 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, vol. 10, pp. 5895–5911, 2022.
- [238] C. Tran, C. M. Luong, and S. Sakti, ‘STEN-TTS: Improving Zero-shot Cross-Lingual Transfer for Multi-Lingual TTS with Style-Enhanced Normalization Diffusion Framework’, in Proc. INTERSPEECH 2023, 2023, pp. 4464–4468.
- [239] H. Hemati and D. Borth, ‘Using ipa-based tacotron for data efficient cross-lingual speaker adaptation and pronunciation enhancement’, arXiv preprint arXiv:2011. 06392, 2020.
- [240] J. Yang and L. He, ‘Towards Universal Text-to-Speech’, in Interspeech 2020, 2020, pp. 3171–3175.
- [241] F. Lux, J. Koch, and N. T. Vu, ‘Low-Resource Multilingual and Zero-Shot Multispeaker TTS’, in Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2022, pp. 741–751.
- [242] F. Lux and T. Vu, ‘Language-Agnostic Meta-Learning for Low-Resource Text-to-Speech with Articulatory Features’, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 6858–6868.
- [243] Y. Jeon and G. Lee, ‘Multi-Level Attention Aggregation for Language-Agnostic Speaker Replication’, in Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), 2024, pp. 14–20.
- [244] C. Gong et al., ‘Zmm-tts: Zero-shot multilingual and multispeaker speech synthesis conditioned on self-supervised discrete speech representations’, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.
- [245] J. Kim, K. Lee, S. Chung, and J. Cho, ‘CLaM-TTS: Improving Neural Codec Language Model for Zero-Shot Text-to-Speech’, in The Twelfth International Conference on Learning Representations, 2024.
- [246] M. Jeong, M. Kim, B. J. Choi, J. Yoon, W. Jang, and N. S. Kim, ‘Transfer Learning for Low-Resource, Multi-Lingual, and Zero-Shot Multi-Speaker Text-to-Speech’, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.

- [247] E. Casanova et al., ‘XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model’, in Interspeech 2024, 2024, pp. 4978–4982.
- [248] J. Betker, ‘Better speech synthesis through scaling’, arXiv preprint arXiv:2305.07243, 2023.
- [249] ‘A new algorithm for data compression optimization’, arXiv preprint arXiv:1209.1045, 2012.
- [250] S.-G. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, ‘Big-VGAN: A Universal Neural Vocoder with Large-Scale Training’, in The Eleventh International Conference on Learning Representations, 2023.
- [251] J. Barker, R. Marixer, E. Vincent, and S. Watanabe, ‘The third ‘ChiME’ speech separation and recognition challenge: Analysis and outcomes’, Computer Speech & Language, vol. 46, pp. 605–626, 2017.
- [252] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, ‘LibriSpeech: an asr corpus based on public domain audio books’, in 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2015, pp. 5206–5210.
- [253] G. Chen et al., ‘GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio’, in Interspeech 2021, 2021, pp. 3670–3674.
- [254] J. Yamagishi, C. Veaux, and K. MacDonald, CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92). University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2019.
- [255] K. Ito and L. Johnson, ‘The LJ Speech Dataset’, 2017. [Online]. Available: <https://keithito.com/LJ-Speech-Dataset/>.
- [256] S. King and V. Karaikos. The blizzard challenge 2011. In Blizzard Challenge Workshop, 2011.
- [257] S. King and V. Karaikos. The blizzard challenge 2013. In Blizzard Challenge Workshop, 2013
- [258] J. Kahn et al., ‘Libri-Light: A Benchmark for ASR with Limited or No Supervision’, in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7669–7673.
- [259] H. Zen et al., ‘Libritts: A corpus derived from librispeech for text-to-speech’, arXiv preprint arXiv:1904.02882, 2019.
- [260] R. Sonobe, S. Takamichi, and H. Saruwatari, ‘JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis’, arXiv preprint arXiv:1711.00354, 2017.
- [261] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, ‘JVS corpus: free Japanese multi-speaker voice corpus’, arXiv preprint arXiv:1908.06248, 2019.
- [262] DataBaker. 2017, “Chinese standard mandarin speech corpus”. [Online]: <https://en.data-baker.com/datasets/freeDatasets/>
- [263] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, ‘Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline’, in 2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA), 2017, pp. 1–5.
- [264] J. Du, X. Na, X. Liu, and H. Bu, ‘Aishell-2: Transforming mandarin asr research into industrial scale’, arXiv preprint arXiv:1808.10583, 2018.
- [265] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, ‘Aishell-3: A multi-speaker mandarin tts corpus and the baselines’, arXiv preprint arXiv:2010.11567, 2020.
- [266] B. Zhang et al., ‘Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition’, in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6182–6186.
- [267] T. Guo et al., ‘Didispeech: A large scale mandarin speech corpus’, in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6968–6972.
- [268] D. P. Lestari, K. Iwano, and S. Furui, ‘A large vocabulary continuous speech recognition system for Indonesian language’, in 15th Indonesian Scientific Conference in Japan Proceedings, 2006, pp. 17–22.
- [269] S. Sakti, E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura, ‘Development of Indonesian large vocabulary continuous speech recognition system within A-STAR project’, in Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST), 2008.
- [270] S. Mussakhojayeva, Y. Khassanov, and H. A. Varol, ‘KazakhTTS2: Extending the Open-Source Kazakh TTS Corpus With More Data, Speakers, and Topics’, in Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 5404–5411.
- [271] R. Ardila et al., ‘Common voice: A massively-multilingual speech corpus’, arXiv preprint arXiv:1912.06670, 2019.
- [272] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, ‘Mls: A large-scale multilingual dataset for speech research’, arXiv preprint arXiv:2012.03411, 2020.
- [273] K. Sodimana et al., ‘A Step-by-Step Process for Building TTS Voices Using Open Source Data and Frameworks for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese’, in SLTU, 2018, pp. 66–70.
- [274] H. He et al., ‘Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation’, in 2024 IEEE Spoken Language Technology Workshop (SLT), 2024, pp. 885–890.
- [275] A. Nagrani, J. S. Chung, and A. Zisserman, ‘Voxceleb: a large-scale speaker identification dataset’, arXiv preprint arXiv:1706.08612, 2017.
- [276] J. S. Chung, A. Nagrani, and A. Zisserman, ‘Voxceleb2: Deep speaker recognition’, arXiv preprint arXiv:1806.05622, 2018.
- [277] C. Busso et al., ‘IEMOCAP: Interactive emotional dyadic motion capture database’, Language resources and evaluation, vol. 42, pp. 335–359, 2008.
- [278] K. Zhou, B. Sisman, R. Liu, and H. Li, ‘Emotional voice conversion: Theory, databases and ESD’, Speech Communication, vol. 137, pp. 1–18, 2022.
- [279] R. Kubichek, ‘Mel-cepstral distance measure for objective speech quality assessment’, in Proceedings of IEEE pacific rim conference on communications computers and signal processing, 1993, vol. 1, pp. 125–128.
- [280] J. H. L. Hansen and T. Hasan, ‘Speaker recognition by machines and humans: A tutorial review’, IEEE Signal processing magazine, vol. 32, no. 6, pp. 74–99, 2015.
- [281] T. Nakatani, S. Amano, T. Irino, K. Ishizuka, and T. Kondo, ‘A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments’, Speech communication, vol. 50, no. 3, pp. 203–214, 2008.
- [282] W. Chu and A. Alwan, ‘Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend’, in 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 3969–3972.
- [283] B. Series, ‘Method for the subjective assessment of intermediate quality level of audio systems’, International Telecommunication Union Radiocommunication Assembly, vol. 2, 2014.
- [284] P. Neekhara, S. Hussain, S. Dubnov, F. Koushanfar, and J. McAuley, ‘Expressive neural voice cloning’, in Asian Conference on Machine Learning, 2021, pp. 252–267.
- [285] J. Li and L. Zhang, ‘Zse-vits: A zero-shot expressive voice cloning method based on vits’, Electronics, vol. 12, no. 4, p. 820, 2023.
- [286] C. Tang, C. Luo, Z. Zhao, D. Yin, Y. Zhao, and W. Zeng, ‘Zero-Shot Text-to-Speech for Text-Based Insertion in Audio Narration’, in Proc. Interspeech 2021, 2021, pp. 3600–3604.
- [287] [Online]. Available: <https://www.cerence.com/news-releases/news-release-details/cerence-introduces-my-car-my-voice-new-voice-clone-solution>. [Accessed: 31-Aug-2024].
- [288] W. Hutiri, O. Papakyriakopoulos, and A. Xiang, ‘Not My Voice! A Taxonomy of Ethical and Safety Harms of Speech Generators’, in The 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 359–376.
- [289] J. Damiani, ‘A voice deepfake was used to scam A CEO out of \$243,000’, Forbes, 03-Sep-2019. [Online]. Available: <https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/>. [Accessed: 31-Aug-2024].

## IX. BIOGRAPHY SECTION

**Hussam Azzuni** received an MSc degree in Computer Vision from MBZUAI, UAE. He is passionate about developing Human-AI interaction algorithms and 3D reconstruction methodologies for AR/VR immersive experiences.

**Abdulmotaleb El Saddik** (Fellow, IEEE) is a Distinguished University Professor specializing in intelligent multimedia computing and digital twin technologies. His research integrates various technologies, including AI, IoT, AR/VR, haptics, and 5G, to enhance real-time interactions in the metaverse. He served as Editor-in-Chief, authored numerous publications and books, chaired conferences/workshops, and supervised over 150 researchers.