A Survey of Text to Speech Audio Anti-Spoofing Detection Speech Synthesis Deepfakes Voice Cloning Synthetic Speech Detection and Audio Forensics

Abstract

This survey paper provides a comprehensive examination of advancements and challenges in text-to-speech (TTS), audio anti-spoofing detection, speech synthesis, deepfakes, voice cloning, synthetic speech detection, and audio forensics. These technologies play a pivotal role in enhancing human-computer interaction, offering natural and expressive interfaces. The integration of neural network-based approaches in TTS has improved speech fidelity, yet challenges such as data scarcity persist. Voice cloning and deepfakes present both opportunities and societal threats, necessitating robust anti-spoofing measures. The survey highlights innovative methodologies, including the use of synthetic data to enhance automatic speech recognition (ASR) systems, and underscores the importance of benchmarks for improving TTS system generalization. Despite advancements, issues like over-smoothing in predictions and noise robustness remain. The paper also explores the transformative potential of these technologies across various domains, including entertainment and security, and emphasizes future research directions to enhance system adaptability and performance. By synthesizing recent findings, this survey aims to guide ongoing research and innovation in the dynamic field of speech technology.

1 Introduction

1.1 Scope and Significance

This survey provides an extensive examination of text-to-speech (TTS), audio anti-spoofing detection, speech synthesis, deepfakes, voice cloning, synthetic speech detection, and audio forensics. These technologies are crucial for enhancing human-computer interaction through more natural and expressive interfaces. TTS systems, which convert text into spoken words, are integral to applications such as assistive technologies and virtual assistants. The adoption of neural network-based methods in TTS, particularly for high-fidelity speech generation, signifies notable progress, despite ongoing challenges like data scarcity in Automatic Speech Recognition (ASR) systems [1].

The survey investigates various TTS technologies, including neural and hybrid TTS, each presenting distinct advantages and drawbacks [2]. The development of personalized TTS systems, optimized for mobile deployment, meets the increasing demand for adaptive and expressive speech solutions [3]. Additionally, the control of speaker identity and style in TTS models, which traditionally depend on reference recordings, necessitates innovative approaches to enhance creative applications [4].

Voice cloning, which transforms a source speaker's voice into that of a target speaker while maintaining linguistic integrity, presents both opportunities and challenges. Deepfakes, capable of convincingly mimicking individuals, pose significant societal threats, including extortion and misinformation [5]. The synthesis of accented TTS, which aims to generate speech variants with specific accents, highlights the complexities involved in achieving naturalness across diverse linguistic contexts [6].

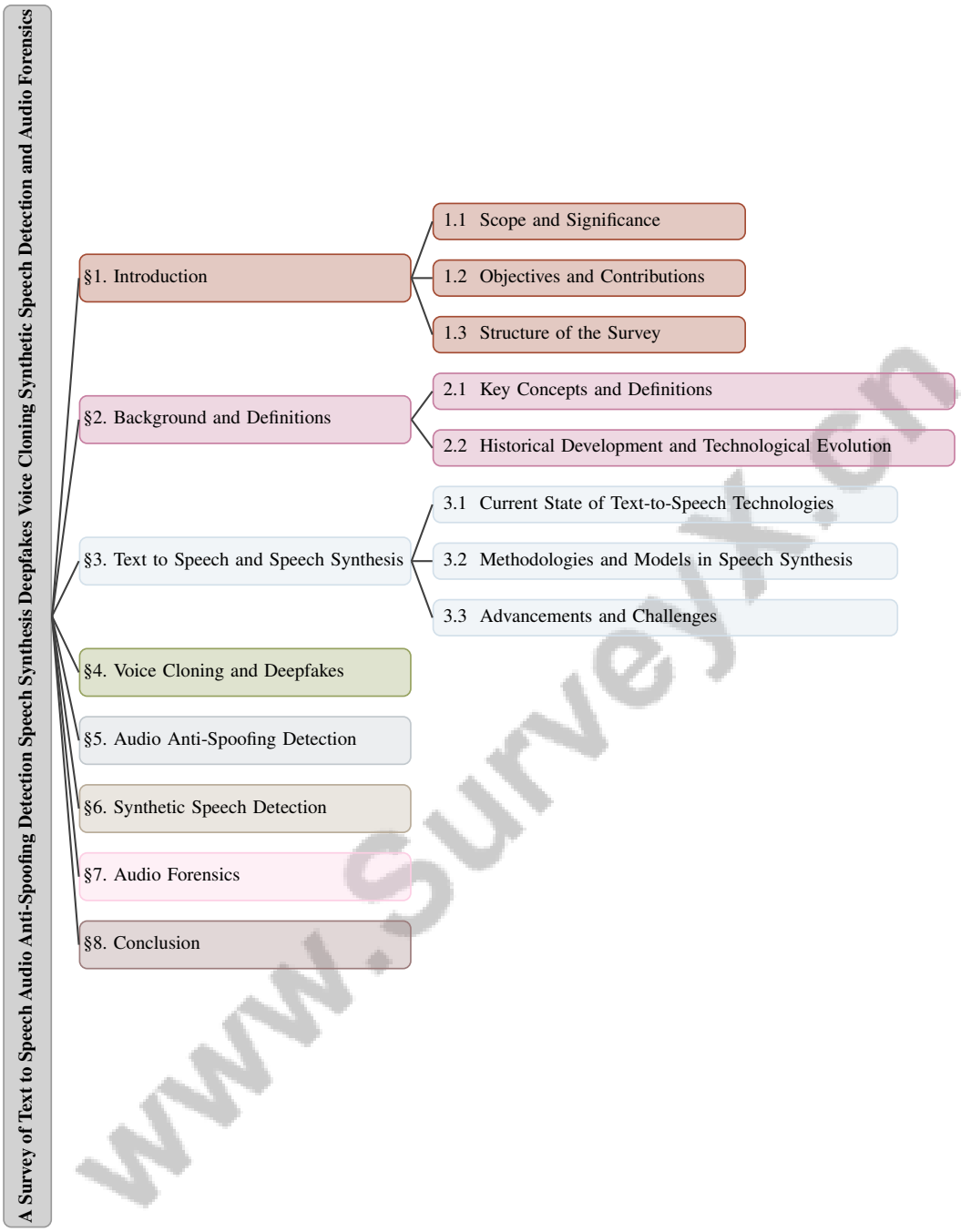


Figure 1: chapter structure

The generation of synthetic audio data from TTS systems trained on ASR corpora enhances ASR performance, particularly in low-resource settings, underscoring the importance of diverse speaker profiles in multi-speaker TTS systems [1]. This survey aims to provide a thorough overview of these technological advancements and challenges, guiding future research and innovation in the evolving field of speech technology.

1.2 Objectives and Contributions

The primary objective of this survey is to analyze recent advancements, methodologies, and applications in TTS, audio anti-spoofing detection, speech synthesis, deepfakes, voice cloning, synthetic

speech detection, and audio forensics. By synthesizing recent research findings, it clarifies the academic and practical significance of audio synthesis and audio-visual multimodal processing technologies. The survey examines various TTS systems—including concatenative, formant synthesis, and statistical parametric approaches—highlighting their advantages and limitations regarding voice naturalness and application suitability. It also provides insights into current advancements, such as neural and hybrid TTS, which offer valuable perspectives for researchers and industry practitioners [7, 8].

A significant contribution of this survey is the exploration of multi-speaker TTS systems capable of generating speech that mimics various target speakers without extensive retraining [9]. The survey emphasizes the role of synthetic data in enhancing ASR systems, particularly in low-resource environments, by comparing synthetic training data with real data [10]. Additionally, the integration of ASR and TTS within a closed-loop speech chain model facilitates joint training on labeled and unlabeled data, outperforming traditional separate systems [11].

The necessity for robust benchmarks to improve TTS systems' generalization to novel spoofing algorithms is also addressed, enhancing the reliability of voice authentication systems [12]. The vulnerabilities of automatic speaker verification systems to spoofing attacks using low-quality audio recordings from uncontrolled environments are examined, underscoring the need for resilient anti-spoofing measures [13].

Innovative methodologies, such as incorporating future contextual information into TTS synthesis through a pseudo lookahead generated by a pretrained language model, are discussed to enhance naturalness without increasing latency [14]. The survey highlights the importance of detecting synthetic spoken misinformation and provides a benchmark for evaluating detection models [15].

Moreover, the survey streamlines the process of adding new languages and voices to TTS platforms using build automation technologies [16]. It addresses the issue of unnatural computerized voices during interactions, which can negatively impact user experience, and discusses the mitigation of unauthorized exploitation of personal audio samples through high-quality deepfake speech synthesis [17].

The examination of adversarial training methods reveals significant improvements in keyword spotting (KWS) model accuracy on real speech data, achieving up to 12

1.3 Structure of the Survey

The survey is structured to deliver an in-depth analysis of technologies and methodologies essential to TTS, audio anti-spoofing detection, speech synthesis, deepfake detection, voice cloning, synthetic speech detection, and audio forensics. It covers various TTS technologies, including concatenative, formant synthesis, and neural TTS, while assessing their effectiveness and applications. The advancements in deep learning that enhance synthesized speech quality and the emerging challenges in deepfake audio detection and multimodal media authenticity are also discussed. By reviewing current research, public datasets, and deep-learning techniques, the survey aims to bridge gaps in the literature and propose future research directions that can lead to improved detection systems and methodologies in these critical areas [18, 8, 19, 20, 21]. The paper begins with an **Introduction**, outlining the survey's scope, significance, objectives, and contributions, establishing a foundation for the subsequent detailed exploration of these technologies and their implications.

The following section, **Background and Definitions**, provides foundational insights into key concepts and terms, offering precise definitions and discussing the historical development and technological evolution of these fields. This section draws parallels to frameworks such as the Ai-TTS model, which incorporates explicit accent intensity control into TTS processes [22].

In the **Text to Speech and Speech Synthesis** section, the survey assesses the current state of TTS technologies and explores various methodologies and models, including the VQTTS method, which addresses mel-spectrogram prediction and elucidates its advantages and limitations [14]. This section also highlights recent advancements and challenges within the field.

The **Voice Cloning and Deepfakes** section examines the techniques and technologies involved in voice cloning and deepfake creation, discussing their privacy and security implications. Innovations leading to more realistic synthetic voices and their potential impacts are further analyzed.

Subsequently, the **Audio Anti-Spoofing Detection** section explores strategies and technologies for detecting and preventing audio spoofing, addressing vulnerabilities in automatic speaker verification systems. The significance of benchmarks and datasets in evaluating anti-spoofing technologies is emphasized.

The **Synthetic Speech Detection** section analyzes methods for detecting synthetic speech, discussing the effectiveness of various detection techniques and the challenges in accurately identifying synthetic audio content.

In the **Audio Forensics** section, the role of audio forensics in verifying the authenticity and integrity of audio recordings is discussed. Methodologies used in audio analysis and challenges faced in forensic investigations are explored.

In the **Conclusion**, the essential findings and insights derived from this comprehensive survey on audio synthesis and audio-visual multimodal processing are synthesized, emphasizing advancements in TTS technologies, including neural TTS and hybrid systems. Future research directions, such as enhancing voice naturalness, multilingual support, and emotional expression in synthesized speech, are discussed, alongside potential applications across various sectors, providing valuable guidance for researchers and practitioners in the field [7, 8, 19, 17, 23]. The importance of ongoing research and collaboration to advance these technologies and address their challenges is underscored.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Key Concepts and Definitions

In the realm of speech technology, text-to-speech (TTS) and speech synthesis are pivotal for research and application. TTS systems convert text into spoken words, often using intermediate representations like mel spectrograms synthesized into audio via neural vocoders. Traditional TTS systems, with their multiple independent stages, face inefficiencies that complicate training and degrade performance [24]. They also risk losing visual features inherent in text due to reliance on discrete symbols [2]. Recent advancements address these issues with end-to-end models that improve architectural efficiency and speech synthesis quality [25]. For instance, synthetic speech generation combined with text augmentation enhances Automatic Speech Recognition (ASR) training in low-resource settings [1]. However, managing a wide range of voice characteristics remains challenging, affecting listener perception [26].

Speech synthesis extends beyond TTS to generate artificial speech closely mimicking human qualities, addressing unique acoustic properties of different speaker groups, such as children [6]. Disentangling content and style factors is crucial for expressiveness while minimizing content leakage [27]. Voice conversion (VC) transforms a source speaker's voice into a target speaker's while preserving linguistic content, enhancing TTS versatility in multilingual and multispeaker contexts [1]. The need for sophisticated models is underscored by existing TTS systems' inability to manage diverse voice characteristics effectively [26].

Generative adversarial networks (GANs) have significantly influenced speech synthesis, enhancing naturalness and expressiveness. GAN-based architectures effectively capture text-based prosody, essential for generating natural and expressive speech [27]. Benchmarks addressing audio generation challenges provide standardized metrics for system performance evaluation across methodologies [25].

A comprehensive understanding of TTS technologies, including concatenative, formant synthesis, and statistical parametric methods, alongside recent advancements like neural and hybrid TTS, equips researchers to navigate the complex landscape of speech technology. This knowledge enables leveraging enhanced user experiences in accessibility tools and virtual assistants while addressing challenges related to naturalness, multilingual support, and high-quality data generation for TTS applications [17, 28, 29, 8].

2.2 Historical Development and Technological Evolution

The evolution of text-to-speech (TTS) systems and audio forensics is marked by technological advancements enhancing synthesized speech's naturalness and robustness. Early TTS systems used

rule-based, concatenative synthesis with unit selection, limiting their ability to capture temporal dependencies, resulting in less natural outputs [30]. The shift to statistical parametric synthesis introduced adaptable, data-driven methodologies, improving natural-sounding speech through advanced algorithms and linguistic models, enhancing user experiences in accessibility tools and virtual assistants [28, 31, 17]. This evolution accelerated with neural network architectures, significantly improving naturalness and expressiveness. The neural source-filter (NSF) model enhanced waveform modeling by addressing complex temporal dependencies.

Despite advancements, challenges like 'content leakage,' where content information inadvertently influences style embeddings, persist, compromising style transfer integrity [27]. Text normalization and phonemization remain significant obstacles to achieving fully data-driven approaches [24]. Innovations like flow matching generative models, such as F5-TTS, aim to tackle these limitations using diffusion models for robust and flexible TTS alignment [32].

Simultaneously, audio forensics has evolved to address challenges from synthetic speech and deepfakes. Comprehensive datasets, including those from the Blizzard and Voice Conversion Challenges, have been crucial for progress [30]. However, the field still faces challenges, notably the need for datasets that generalize across diverse languages and speech types for effective synthetic speech detection [20].

Ongoing TTS and audio forensics evolution necessitates continuous innovation. As demand for personalized voice assistants and open vocabulary keyword spotting grows, the field must develop robust evaluation frameworks and refine methodologies to address emerging challenges [33]. Future research should explore fully data-driven approaches that eliminate text normalization and phonemization while enhancing model architectures to improve audio quality [24]. Developing sophisticated models capable of managing diverse voice characteristics and preventing content leakage in style embeddings will be crucial for the future of speech synthesis technologies [27].

3 Text to Speech and Speech Synthesis

3.1 Current State of Text-to-Speech Technologies

Method Name	Model Architecture	Adaptability Challenges	Expressiveness Enhancement
EATS[24]	End-to-end	Unseen Speakers	Prosody Manipulation
D2[13]	End-to-end	Multi-speaker Scenarios	Adversarial Training
VQTTS[14]	Classification-based Model	Unseen Speakers	Diverse Prosodies
TA-NN[1]	Linguistic Frontend	Unseen Speakers	Voice Diversity
Ai-TTS[22]	Fastspeech2 Architecture	Unseen Speakers	More Natural
PTTS[4]	Style Content Encoder	Unseen Style Factors	Natural Varied Speech
Coco-Nut[26]	Retrieval Model Training	Low-resource Languages	Diverse Corpus Construction
TTRS[34]	Tacotron-based Multispeaker	Unseen Speakers Adaptation	Phoneme-level Prosody
SPVC[3]	Structured Pruning	Unseen Speakers	Natural And Human-like

Table 1: Overview of contemporary text-to-speech (TTS) models, highlighting their architectural frameworks, adaptability challenges, and enhancements in expressiveness. This table categorizes various models, such as EATS, D2, and VQTTS, according to their approach to handling unseen speakers, multi-speaker scenarios, and diverse prosodic features, thereby illustrating the breadth of strategies employed in modern TTS systems.

The current landscape of text-to-speech (TTS) technologies showcases significant advancements that enhance the naturalness and expressiveness of synthesized speech. Central to these developments is the End-to-End Adversarial Text-to-Speech (EATS) model, which streamlines the TTS training process through minimal supervision while achieving high-fidelity outputs comparable to state-of-the-art systems [24].

The evolution of TTS has seen a shift towards end-to-end solutions that integrate various components of the TTS pipeline. For example, combining a codec network with an acoustic model allows for joint optimization of speech representation learning and waveform reconstruction, thereby improving output quality and intelligibility [13]. The VQTTS approach exemplifies this trend by utilizing a classification-based acoustic model and a vocoder to generate waveforms from quantized acoustic features and prosody, further enhancing synthesized speech [14].

Despite these advancements, challenges persist, particularly in synthesizing speech for unseen speakers and low-resource languages where training data is limited [1]. The Ai-TTS model addresses

this issue by synthesizing speech with accent variations, improving TTS adaptability across diverse linguistic contexts [22]. Additionally, models like PromptTTS enhance user-friendliness by extracting style and content representations from input prompts [4].

Moreover, recent research has focused on making computer-generated speech sound more natural and human-like. The Coco-Nut approach constructs a diverse corpus of paired speech and voice characteristic descriptions from various internet sources to improve the expressiveness of synthesized speech [26]. Phoneme-level prosody manipulation has also enabled the generation of expressive voices, such as rapping and singing, showcasing the versatility of TTS systems [34].

Current TTS technologies reflect ongoing efforts to mitigate the impracticality of extensive training data and large models for mobile deployment [3]. The BASE TTS model demonstrates this by generating speech from text through discrete speech representations, known as speechcodes, and converting them into waveforms using efficient methodologies [3]. This approach significantly advances the field by enhancing the naturalness and expressiveness of synthesized speech, addressing the challenge of achieving human-like quality [17].

Table 1 provides a comprehensive summary of current text-to-speech (TTS) technologies, detailing their model architectures, adaptability challenges, and expressiveness enhancements, which are critical for understanding the advancements and ongoing challenges in the field. Figure 2 illustrates the key advancements in text-to-speech (TTS) technologies, focusing on end-to-end models, adaptability challenges, and improvements in expressiveness and naturalness. It categorizes significant models and approaches that have contributed to enhancing TTS systems' performance and adaptability across diverse linguistic contexts.

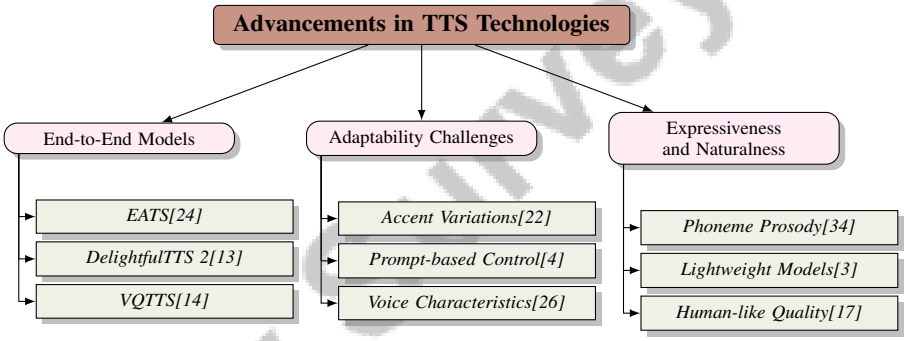


Figure 2: This figure illustrates the key advancements in text-to-speech (TTS) technologies, focusing on end-to-end models, adaptability challenges, and improvements in expressiveness and naturalness. It categorizes significant models and approaches that have contributed to enhancing TTS systems' performance and adaptability across diverse linguistic contexts.

3.2 Methodologies and Models in Speech Synthesis

The evolution of speech synthesis methodologies has been significantly influenced by neural network-based approaches, which have markedly improved the naturalness and expressiveness of synthesized speech. Traditional TTS systems relied on rule-based and concatenative synthesis techniques, which, while foundational, lacked flexibility and naturalness [17]. These early methods have largely been replaced by advanced models that leverage neural networks to overcome previous limitations.

One notable advancement is the Structured Pruning for Voice Cloning (SPVC) technique, which utilizes structured pruning to create lightweight TTS models without sacrificing audio quality [3]. This method addresses the challenge of deploying TTS systems on resource-constrained devices while maintaining high-quality output. Another innovative approach is the visual-text to speech (vTTS) model, which synthesizes speech directly from visual text images, enhancing the naturalness and expressiveness of generated speech [2].

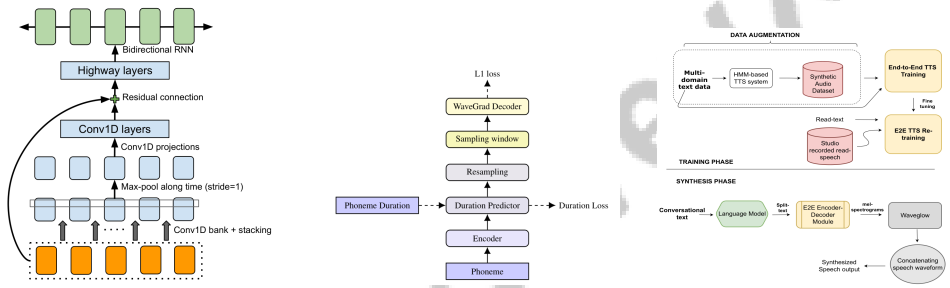
The integration of large pre-trained neural networks for text generation has been explored to augment training data for automatic speech recognition (ASR) models. The Text Generation for Speech Synthesis (TGSS) method employs these networks to generate synthetic texts for conversion to speech, significantly enhancing ASR performance in low-resource settings [1]. Similarly, the PromptTTS

framework uses a style encoder, content encoder, and speech decoder to generate speech, offering a user-friendly approach to TTS [4].

The Fastpitch TTS model, pretrained on a diverse adult speech dataset and fine-tuned on a cleaned child speech dataset, exemplifies efforts to synthesize realistic child voices by addressing the unique acoustic properties of different speaker groups [6]. This highlights the importance of diverse training data for achieving high-quality speech synthesis across various speaker profiles.

End-to-end models have revolutionized the field by simplifying the TTS process and improving synthesized speech quality. The EATS model employs a neural network generator that maps input sequences of characters or phonemes directly to raw audio, using a feed-forward architecture with an aligner and decoder for high-fidelity speech generation [24]. Additionally, the DelightfulTTS framework leverages vector-quantized auto-encoders and adversarial training to learn intermediate speech representations, facilitating high-quality waveform reconstruction [13].

Despite these advancements, challenges remain, particularly in achieving expressive and natural-sounding speech across diverse linguistic contexts. Methods focusing on explicit accent intensity control, such as the Ai-TTS model, aim to address these challenges by incorporating accent intensity into the TTS process [22]. Furthermore, disentangling content and style factors is a critical research area, with collaborative and adversarial learning strategies employed to enhance expressive capabilities while minimizing content leakage [27].



(a) Convolutional Neural Network with Highway and Convolutional Projections[35]

(b) WaveGrad Decoder for Phoneme Duration Prediction[36]

(c) End-to-End Text-to-Speech (TTS) System with Data Augmentation and Fine-Tuning[37]

Figure 3: Examples of Methodologies and Models in Speech Synthesis

As illustrated in Figure 3, three notable methodologies and models emerge in the exploration of advancements in TTS and speech synthesis. The first example, a convolutional neural network (CNN) with highway and convolutional projections, enhances the network's ability to learn complex patterns through residual connections, leading to more accurate speech synthesis. The second example, the WaveGrad Decoder for phoneme duration prediction, employs components like a Duration Predictor to refine the temporal aspects of synthesized speech. Lastly, the end-to-end TTS system with data augmentation and fine-tuning utilizes multi-domain text data and an HMM-based system to generate synthetic audio, emphasizing the importance of data augmentation and fine-tuning for high-quality speech synthesis. Collectively, these examples underscore the diverse methodologies and innovative models driving progress in the field of speech synthesis [35, 36, 37].

3.3 Advancements and Challenges

Recent advancements in TTS and speech synthesis technologies have led to notable improvements in speech intelligibility, audio quality, and speaker similarity. The development of two-stage TTS systems has demonstrated superior performance compared to state-of-the-art methods, highlighting the potential for enhanced speech synthesis quality [38]. DelightfulTTS 2 has further optimized representation learning, effectively addressing the limitations of traditional TTS models [13].

The integration of advanced methodologies, such as generative adversarial networks (GANs) and diffusion models, has been pivotal in achieving these improvements. The VQTTS approach has significantly narrowed the quality gap between ground-truth and predicted acoustic features, achieving state-of-the-art performance in TTS synthesis [14]. Additionally, the Unet-TTS model has

shown significant enhancements in synthesizing expressive speech with accurate speaker and style characteristics, outperforming baseline methods in subjective and objective evaluations [39].

Innovations have also focused on enhancing the diversity and expressiveness of synthesized speech. Watanabe et al.'s approach utilizes a broad range of voice characteristics from real-world data, significantly improving synthesized speech diversity and expressiveness [26]. The Ai-TTS model further advances this area by effectively controlling accent intensity, achieving over 80

Despite these advancements, challenges persist, particularly in synthesizing speech for unseen speakers and low-resource languages. Data scarcity remains a significant barrier, necessitating innovative approaches like model pretraining to enhance performance on domain-mismatched text [1]. Additionally, issues such as over-smoothing in TTS predictions hinder the ability of models to capture variations in speaking styles, which is critical for achieving natural-sounding speech.

Efforts to enhance noise robustness in TTS systems have also been explored, with models like ReFlow-TTS demonstrating improvements in intelligibility and audio quality under noisy conditions. However, challenges such as stuttering and pitch instability in adverse environments highlight the need for further research and innovation in this area. Moreover, the ability to generate intelligible speech from out-of-vocabulary (OOV) and rare characters, as demonstrated by the vTTS model, remains an ongoing challenge [2].

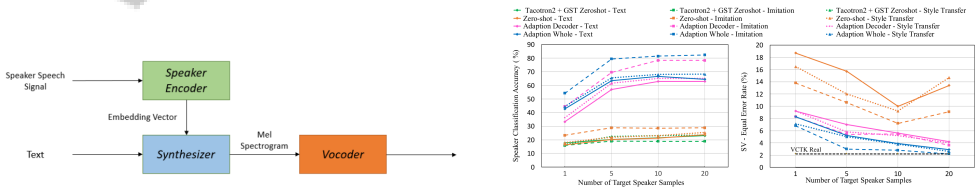
4 Voice Cloning and Deepfakes

4.1 Techniques and Technologies in Voice Cloning

The evolution of voice cloning technologies is largely attributed to advancements in deep learning and neural network architectures that enhance speech synthesis. Techniques like x-vector embeddings extract speaker-specific features, enabling the generation of speech that closely resembles a target speaker's voice even with limited data, thus reducing training requirements [40, 41]. Generative adversarial networks (GANs) have been pivotal, with models such as GAN-TTS using a conditional feed-forward generator and a Transformer-based discriminator to improve the naturalness and diversity of synthetic voices [42, 43].

Transfer learning further optimizes data usage and training time, exemplified by the NAUTILUS system, which facilitates voice cloning from both transcribed and untranscribed speech [44]. This system's use of a speaker encoder, multispeaker Tacotron 2 synthesizer, and universal SC-WaveRNN vocoder enhances attention mechanisms, producing natural and expressive speech across different speakers. Prosody cloning has also advanced TTS applications, notably for anonymization and audiobook customization [45]. Models like ComSpeech, which integrate audio-visual information, exemplify high-quality synthetic voice generation [11].

Augmenting training datasets with diverse TTS data improves customizable keyword spotting systems [15]. Techniques leveraging automatic labeling and natural language prompts highlight voice cloning's potential in creating high-fidelity speech synthesis systems [9]. The VQTTS method, transforming acoustic models into classification models, and PromptTTS, enabling style control through text descriptions, represent significant advancements in this field [14, 4].



(a) Speaker Synthesis System[46]

(b) Speaker Classification and Equal Error Rate Analysis for Different Adaptation Strategies[47]

Figure 4: Examples of Techniques and Technologies in Voice Cloning

Voice cloning and deepfake technologies have rapidly progressed, utilizing sophisticated techniques to produce accurate synthetic voices. As shown in Figure 4, these technologies rely on systems like the Speaker Synthesis System, which includes a Speaker Encoder, Synthesizer, and Vocoder. The

effectiveness of adaptation strategies in voice synthesis is evaluated using speaker classification and equal error rate metrics, providing insights into the accuracy and adaptability of these technologies [46, 47].

4.2 Privacy and Security Implications

The rapid development of voice cloning and deepfake technologies raises significant privacy and security concerns. The ability to create highly realistic synthetic voices poses ethical challenges, particularly regarding privacy and misuse for purposes like identity theft and misinformation [48]. Fine-tuning models for specific speakers enhances performance but also increases the risk of unauthorized voice replication [49]. Research into generating deceptive audio highlights the dual-use risks of these technologies, emphasizing the need for ethical considerations in their development and deployment [50, 48].

Advanced Deepfake Speech Detection (DSD) models, using ensemble techniques and diverse datasets, have been developed to enhance detection capabilities [20]. However, challenges persist, particularly due to insufficient training data for certain linguistic groups, which affects the performance of detection systems [51]. Accurate prosody modeling without speaker mismatch is critical for generating natural speech, potentially reducing privacy concerns [52]. The dual-use nature of these technologies necessitates a balanced approach that considers both benefits and risks to ensure advancements in speech synthesis protect individual privacy and security.

4.3 Innovations in Synthetic Voice Realism

Recent advances in TTS and voice cloning technologies have significantly enhanced the realism of synthetic voices. Deep learning techniques have enabled the development of models that closely mimic human speech [17]. The visual-text to speech (vTTS) model, using convolutional neural networks to extract visual features from text images, exemplifies these advancements [2]. GANs further enhance realism by capturing text-based prosody, essential for generating natural and expressive speech [17].

Despite progress, challenges remain in achieving zero-shot synthesis and mobile deployment due to computational constraints. While recent TTS systems produce human-like speech, modeling highly expressive voices remains challenging due to a trade-off between expressiveness and signal quality. Techniques like data augmentation and GANs show promise in enhancing expressive speech quality, yet the pursuit of perfect naturalness continues to pose obstacles [15, 1]. Developing multilingual datasets is crucial to overcoming these challenges, ensuring that TTS systems effectively capture diverse linguistic characteristics and speaker profiles. Future research should focus on accessible and scalable solutions that enhance the naturalness and expressiveness of synthetic speech, broadening their applicability across diverse contexts.

In recent years, the field of audio anti-spoofing detection has seen significant advancements, yet it continues to face various challenges that necessitate ongoing research. To better understand these dynamics, Figure 5 illustrates the hierarchical structure of audio anti-spoofing detection. This figure outlines key strategies and vulnerabilities within the domain, while also emphasizing the critical role of benchmarks and datasets. By highlighting technological advancements and identifying the challenges that lie ahead, the figure serves as a comprehensive overview of future research directions essential for combating sophisticated spoofing techniques.

5 Audio Anti-Spoofing Detection

5.1 Strategies and Technologies for Anti-Spoofing

The advancement of audio technologies necessitates robust anti-spoofing detection mechanisms to combat sophisticated voice synthesis techniques like deepfakes and voice cloning, which increasingly threaten personal voice data security [48]. Machine learning, especially deep learning, enhances the accuracy of these detection systems. The DelightfulTTS framework, for instance, employs a two-stage training strategy with vector-quantized auto-encoders and adversarial training, improving TTS systems' resilience against spoofing attacks while maintaining speech naturalness [13]. Similarly,

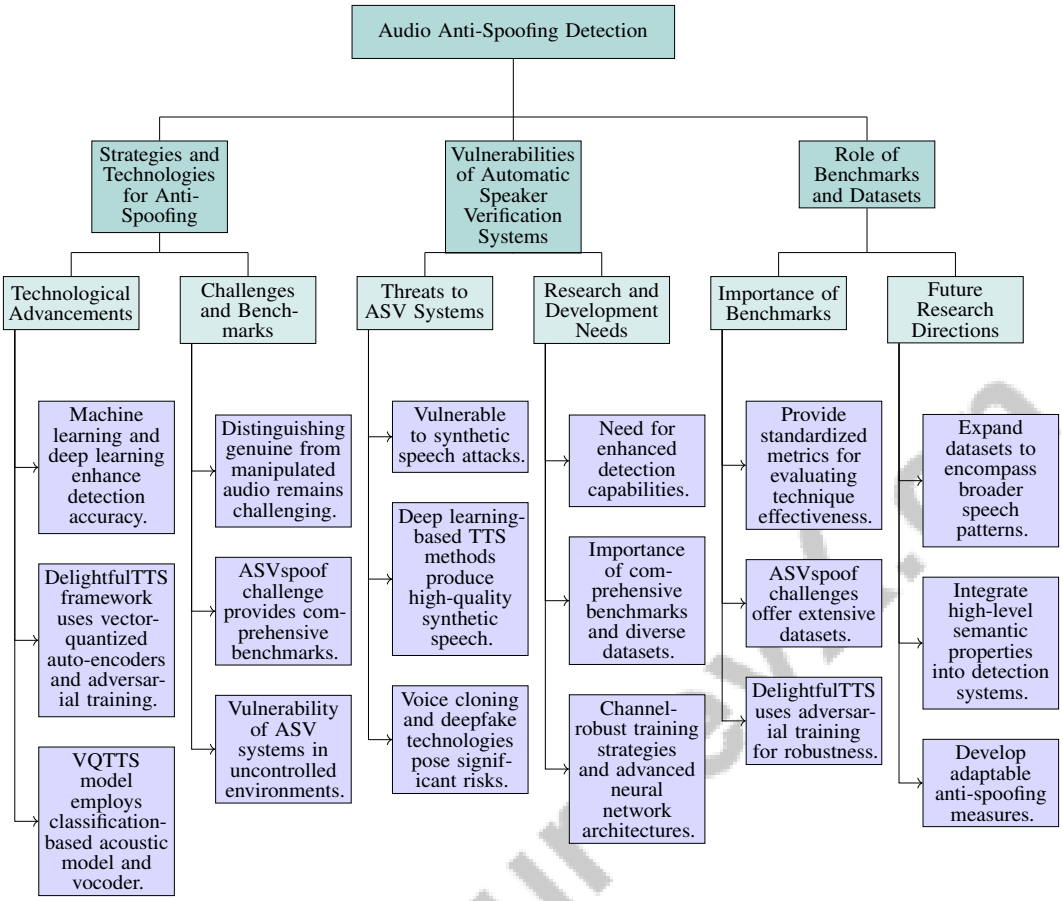


Figure 5: This figure illustrates the hierarchical structure of audio anti-spoofing detection, outlining key strategies, vulnerabilities, and the role of benchmarks and datasets. It highlights technological advancements, challenges, and future research directions in combating sophisticated spoofing techniques.

the VQTTS model advances speech synthesis by using a classification-based acoustic model and vocoder, complicating adversarial exploitation [14].

Addressing automatic speaker verification (ASV) vulnerabilities, the DelightfulTTS framework demonstrates adversarial training's potential to enhance TTS robustness [13]. Synthetic data also plays a crucial role in improving automatic speech recognition (ASR) systems in low-resource settings by mimicking real-world scenarios to accommodate diverse speaker profiles [1, 10].

Despite these advancements, distinguishing genuine from manipulated audio remains challenging. Comprehensive benchmarks, such as the ASVspoof challenge, are essential for evaluating anti-spoofing technologies [12]. The vulnerability of ASV systems to low-quality audio recordings in uncontrolled environments underscores the need for ongoing innovation in audio anti-spoofing detection [13].

5.2 Vulnerabilities of Automatic Speaker Verification Systems

Automatic Speaker Verification (ASV) systems, while effective for voice-based authentication, are vulnerable to advanced spoofing techniques, notably synthetic speech attacks. These vulnerabilities are exacerbated by deep learning-based TTS methods that produce high-quality synthetic speech closely mimicking human voices [53]. Unseen conditions, particularly channel effects, further complicate detection, as current methods struggle to maintain accuracy across diverse acoustic environments [54].

Voice cloning and deepfake technologies, utilizing generative models like GANs, pose significant risks to ASV system integrity, necessitating enhanced detection capabilities to prevent misuse, such as identity theft [53]. Research is crucial to develop resilient anti-spoofing measures, including comprehensive benchmarks and diverse datasets reflecting potential spoofing attack complexities. Recent studies emphasize the need for innovative approaches, such as channel-robust training strategies and advanced neural network architectures, to enhance ASV systems' effectiveness in modern communication contexts [55, 20, 54, 56, 57].

5.3 Role of Benchmarks and Datasets

Benchmark	Size	Domain	Task Format	Metric
ASR-Personalization[58]	1,000	Speech Recognition	Model Personalization	Word Error Rate
SCC-Benchmark[59]	4,360,000	Speech Command Classification	Keyword Spotting	Accuracy, F1-score
TTS-Bench[60]	10,000	Speech Synthesis	Speech Distribution Evaluation	WER, WER-Ratio
MLAAD[61]	28,345	Audio Deepfake Detection	Attribute Classification	Accuracy, F1-score
SpMis[31]	360,611	Spoken Misinformation Detection	Misinformation Detection	Accuracy, F1-score
FMFCC-A[62]	50,000	Synthetic Speech Detection	Detection	Log-loss, EER
TTSDS[63]	3,500	Speech Synthesis	Quality Evaluation	TTSDS, WER
ASVspoof5[64]	1,000	Voice Anti-Spoofing	Deepfake Detection	EER

Table 2: This table presents a comprehensive overview of various benchmarks utilized in the evaluation of audio processing technologies, specifically focusing on speech recognition, synthesis, and detection tasks. It details the size, domain, task format, and evaluation metrics for each benchmark, providing a foundational reference for researchers aiming to enhance audio anti-spoofing and detection systems.

Robust benchmarks and datasets are crucial for evaluating audio anti-spoofing technologies, providing standardized metrics for assessing technique effectiveness across scenarios. Table 2 provides a detailed overview of representative benchmarks in the field of audio processing, highlighting their relevance and application in advancing anti-spoofing technologies and synthetic speech detection. The ASVspoof challenges offer extensive datasets of genuine and spoofed speech samples, serving as benchmarks for developing advanced anti-spoofing models like DelightfulTTS, which uses adversarial training to enhance TTS robustness [65, 13].

Innovative methodologies, such as the Unsupervised Text-to-Speech Synthesis (UTSS) model, leverage unlabelled data to improve synthetic speech detection systems, particularly valuable in scenarios with scarce labeled data [66]. Despite advancements, challenges persist in distinguishing genuine from manipulated audio in real-world scenarios. Expanding benchmark datasets to encompass broader speech patterns and contextual factors is essential. Recent research emphasizes integrating high-level semantic properties, such as speaker identity cues and voice prosody, into detection systems. Exploring deep learning techniques to identify deepfake speech is crucial, as these technologies generate highly realistic audio posing significant risks if misused. A comprehensive approach combining innovative detection strategies with diverse and representative data is vital for addressing synthetic speech detection challenges and ensuring robust performance across applications [67, 58, 68, 20, 21]. Future research should focus on developing adaptable anti-spoofing measures to ensure audio communication systems' integrity and security in a digital world.

6 Synthetic Speech Detection

6.1 Detection Techniques and Methodologies

Synthetic speech detection is increasingly vital due to advances in text-to-speech (TTS) and voice cloning technologies. Techniques such as the Textual Echo Cancellation (TEC) framework enhance detection by integrating audio and text inputs, improving the distinction between genuine and synthetic speech in intelligent devices [69]. Comprehensive datasets like the FMFCC-A, with 50,000 utterances, and the SpMis, with 360,611 samples, are crucial for developing robust detection models in Mandarin speech and misinformation contexts, respectively [62, 31]. The IPAD dataset further supports the evaluation of methodologies through diverse impersonation samples [70].

Evaluation metrics such as Equal Error Rate (EER) and Area Under the Curve (AUC) are standard measures for assessing model performance in differentiating genuine from synthetic speech [15, 71]. Metrics like Mel-Cepstral Distortion (MSD) and F0 Frame Error (FFE) are used alongside human evaluations to assess the naturalness and similarity of synthesized speech [45].

Adversarial techniques, including the AASIST3 method, enhance detection through advanced neural networks and data pre-processing [56]. Experiments on datasets like TIMIT-TTS, with nearly 80,000 synthetic tracks, demonstrate adversarial training's potential to improve detection accuracy [18]. The classification of acoustic models and vocoders, utilizing diverse TTS system datasets, enhances synthetic speech identification [61]. Subjective evaluations, such as Mean Opinion Scores (MOS), provide insights into the perceptual quality of synthesized speech [52].

The rapid evolution of synthetic speech technologies necessitates sophisticated detection techniques to address challenges posed by deepfake speech and misinformation. Combining automatic speaker verification with prosody analysis is promising for enhancing detection capabilities. Specialized datasets like SpMis address the urgent need for effective synthetic spoken misinformation detection [31, 21, 67, 20]. Leveraging diverse datasets, robust evaluation metrics, and innovative adversarial strategies will enhance synthetic speech detection systems across various applications and linguistic contexts.

6.2 Evaluation Metrics and Performance

Synthetic speech detection methods are evaluated using various metrics that collectively assess model performance. The Equal Error Rate (EER) is a key metric indicating the balance between false acceptances and rejections, essential for security-sensitive applications [54]. The minimum tandem detection cost function (min-tDCF) provides insights into trade-offs between error types and their associated costs, crucial for assessing the economic viability of detection systems [54]. Word error rates (WER), including cpWER and MIMO-WER, are vital for analyzing transcription accuracy in speech recognition systems dealing with synthetic speech [10].

Despite these metrics, challenges remain in capturing emotional expressiveness, complex linguistic variations, and real-time synthesis without latency [17]. This underscores the need for advanced evaluation frameworks to capture synthetic speech nuances and their impacts on detection systems.

7 Audio Forensics

7.1 Role of Audio Forensics in Authenticity Verification

Audio forensics is essential in verifying audio recording authenticity, playing a critical role in legal and security sectors. Its primary function is to meticulously analyze audio evidence to ensure its integrity, confirming recordings are unaltered [68]. This involves detecting anomalies such as edits or splices that may undermine the evidence's credibility [72].

The field continually evolves, adapting to advances in audio synthesis technologies. Techniques like spectral, waveform, and phase analysis are pivotal in identifying inconsistencies indicative of tampering [73]. These methods enable forensic experts to scrutinize audio signals thoroughly, distinguishing between authentic and manipulated recordings.

The rise of deepfake technologies and advanced voice cloning poses significant challenges to audio forensics, as these technologies facilitate the creation of highly realistic synthetic voices. Consequently, the field increasingly relies on machine learning and artificial intelligence to enhance its ability to detect deepfakes and other synthetic audio manipulations [74, 54].

Beyond legal applications, audio forensics is crucial in media verification, ensuring the authenticity of audio content on news and online platforms. This is vital in combating misinformation, where fabricated audio can mislead audiences and manipulate public opinion [75]. By providing reliable verification means, audio forensics enhances information dissemination integrity and protects privacy rights.

7.2 Methodologies in Audio Analysis

Audio forensic analysis employs diverse methodologies to examine recordings for authenticity, integrity, and manipulation, which are crucial in legal and security contexts. The sophistication of audio deepfakes and speech synthesis technologies heightens the importance of accurate audio evidence assessment [18, 76, 77, 78, 21].

Spectral analysis, a foundational technique, inspects audio signal frequency spectra to uncover anomalies suggesting tampering, such as abrupt spectral content changes [68]. Waveform analysis examines audio waveform shape and amplitude to identify discontinuities indicative of manipulation [72]. Phase analysis complements these methods by examining phase information within audio signals, effectively identifying phase discontinuities from splicing or manipulation [73].

The emergence of deepfake technologies necessitates advanced detection methodologies in audio forensics. Machine learning and artificial intelligence have become integral, providing powerful tools for analyzing audio signals and detecting deepfakes [54]. These technologies facilitate model creation capable of learning complex audio patterns, thereby improving forensic analysis accuracy.

Comprehensive datasets, such as those from ASVspoof challenges, are crucial for advancing audio forensics. They offer diverse genuine and spoofed audio samples, aiding the development and evaluation of detection models [65]. Utilizing these resources enables researchers to create more robust methodologies that can effectively distinguish between authentic and manipulated audio content.

7.3 Challenges in Forensic Investigations

Forensic investigations of audio recordings face numerous challenges, particularly due to the rapid evolution of audio synthesis and manipulation technologies. Differentiating genuine from manipulated audio content is a primary challenge. The sophistication of deepfake technologies and voice cloning complicates identifying subtle manipulations, such as splicing and prosody modification, which can undermine recording integrity and facilitate malicious exploitation [67, 20, 57]. Recent innovations, such as the Pivotal Objective Perturbation (POP) method, introduce imperceptible noise to original speech samples, enhancing resilience against unauthorized synthesis and improving detection capabilities.

Another significant challenge is the scarcity of comprehensive datasets representing the diversity of languages, accents, and speaker characteristics encountered in real-world scenarios. This limitation restricts the generalizability of forensic methodologies, particularly in detecting synthetic speech across various linguistic and cultural contexts [20]. Furthermore, the need for datasets encompassing a wide range of spoofing attacks, including those generated by advanced text-to-speech and voice conversion technologies, exacerbates this challenge [65].

The complexity of audio signals and the intricacies of genuine and manipulated speech necessitate sophisticated methodologies for robust detection models. While traditional techniques like spectral, waveform, and phase analysis are useful, they may not suffice against advanced synthetic speech technologies. Thus, integrating machine learning and artificial intelligence into forensic investigations is essential for enhancing detection capabilities and improving the accuracy of audio analysis [54].

Moreover, the dynamic nature of audio synthesis technologies demands continuous adaptation and innovation in forensic methodologies. As new techniques for generating synthetic speech emerge, forensic experts must refine detection strategies to maintain the integrity and credibility of audio evidence [32].

8 Conclusion

8.1 Applications and Future Directions

The advancements in TTS, audio anti-spoofing detection, speech synthesis, deepfakes, voice cloning, synthetic speech detection, and audio forensics have significantly impacted various domains, notably in entertainment, where the creation of realistic synthetic voices has transformed content production. These technologies enhance user engagement and interaction, offering more immersive experiences. Future research should focus on evaluating the naturalness and intelligibility of generated speech,

especially in ASR systems. In audio forensics, the development of robust detection models and comprehensive datasets is crucial for ensuring the authenticity of audio recordings. Establishing standardized benchmarks will aid in distinguishing genuine from manipulated audio, enhancing the reliability of forensic analyses. Further efforts are needed to expand datasets and improve the resilience of audio watermarking techniques against diverse perturbations.

The integration of synthetic speech data into ASR training has proven beneficial, particularly in low-resource settings. Future research should explore the subjective evaluation of generated child speech and its implications for ASR applications. Promising research directions include optimizing inference performance and integrating model components into a unified, end-to-end trainable framework. Exploring fully parallel TTS generation and alternative modeling techniques, such as flow matching generative models, could yield more robust and flexible TTS systems. The impact of synthetic data on ASR system personalization presents another promising avenue, with neural text augmentation showing significant improvements in ASR performance. This suggests potential applications in enhancing voice cloning and synthetic speech detection technologies.

Future research should also focus on multi-speaker style transfer, refining system capabilities and disentanglement techniques. Developing publicly available paired corpora of speech and voice characteristic descriptions will further advance the field. Exploring new methodologies and models, such as direct speech-to-speech translation, represents another promising direction. Enhancing TTS systems' adaptability to various linguistic contexts and speaker profiles remains crucial, with ongoing efforts to create accessible and scalable solutions that improve the naturalness and expressiveness of synthesized speech. In audio forensics, expanding datasets to cover a broader range of spoofing attacks is essential for safeguarding audio communications in an increasingly digital world. Addressing these challenges will empower researchers and practitioners to enhance the integrity of audio communications.

As the field evolves, future research should prioritize developing scalable solutions that enhance the naturalness and expressiveness of synthesized speech, ensuring applicability across diverse linguistic contexts and speaker profiles. Enhancing TTS systems' adaptability to various speaker characteristics and linguistic contexts is critical, with efforts directed toward improving control over voice characteristics and accent intensity. Additionally, future studies could investigate the meanings of extracted multi-level embeddings and aim to develop controllable multi-utterance style transfer capabilities.

8.2 Impact of Synthetic Data on ASR Personalization

The integration of synthetic data into ASR systems has become crucial for enhancing personalization and performance. Recent studies demonstrate that effectively utilizing synthetic data can significantly improve ASR performance, especially in data-scarce scenarios. This approach addresses challenges posed by limited real training data, enabling the development of more robust ASR models. A key aspect of ASR personalization is content alignment rather than stylistic matching, as effective model adaptation relies on content consistency. This underscores the need for synthetic data generation methodologies that prioritize content alignment to bolster ASR performance across diverse applications and speaker profiles. While the use of synthetic data has shown promise, its effectiveness can vary based on the underlying model architectures.

Despite advancements, challenges persist in expanding training data to encompass a wider array of contexts and enhancing algorithms for improved performance in diverse scenarios. Developing advanced algorithms that enhance the naturalness and expressiveness of TTS systems is crucial for elevating ASR performance, as these features are essential for achieving high-quality synthesized speech. By focusing on algorithms that effectively manage diverse linguistic contexts and speaker profiles, researchers can enhance the adaptability and personalization of ASR systems, ensuring their applicability across various fields.

References

[1] Zhuangqun Huang, Gil Keren, Ziran Jiang, Shashank Jain, David Goss-Grubbs, Nelson Cheng, Farnaz Abtahi, Duc Le, David Zhang, Antony D'Avirro, Ethan Campbell-Taylor, Jessie Salas, Irina-Elena Veliche, and Xi Chen. Text generation with speech synthesis for asr data augmentation, 2023.

[2] Yoshifumi Nakano, Takaaki Saeki, Shinnosuke Takamichi, Katsuhito Sudoh, and Hiroshi Saruwatari. vtts: visual-text to speech, 2022.

[3] Sung-Feng Huang, Chia ping Chen, Zhi-Sheng Chen, Yu-Pao Tsai, and Hung yi Lee. Personalized lightweight text-to-speech: Voice cloning with adaptive structured pruning, 2023.

[4] Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[5] Hongbin Liu, Moyang Guo, Zhengyuan Jiang, Lun Wang, and Neil Zhenqiang Gong. Audiomarkbench: Benchmarking robustness of audio watermarking, 2024.

[6] Rishabh Jain and Peter Corcoran. Improved child text-to-speech synthesis through fastpitch-based transfer learning, 2023.

[7] Zhaofeng Shi. A survey on audio synthesis and audio-visual multimodal processing, 2021.

[8] Md. Jalal Uddin Chowdhury and Ashab Hussan. A review-based study on different text-to-speech technologies, 2023.

[9] Dan Lyth and Simon King. Natural language guidance of high-fidelity text-to-speech with synthetic annotations, 2024.

[10] Samuele Cornell, Jordan Darefsky, Zhiyao Duan, and Shinji Watanabe. Generating data with text-to-speech and large-language models for conversational speech recognition, 2024.

[11] Qingkai Fang, Shaolei Zhang, Zhengrui Ma, Min Zhang, and Yang Feng. Can we achieve high-quality direct speech-to-speech translation without parallel speech data?, 2024.

[12] Yi-Chiao Wu, Patrick Lumban Tobing, Kazuki Yasuhara, Noriyuki Matsunaga, Yamato Ohtani, and Tomoki Toda. A cyclical approach to synthetic and natural speech mismatch refinement of neural post-filter for low-cost text-to-speech system, 2022.

[13] Yanqing Liu, Ruiqing Xue, Lei He, Xu Tan, and Sheng Zhao. Delightfultts 2: End-to-end speech synthesis with adversarial vector-quantized auto-encoders, 2022.

[14] Chenpeng Du, Yiwei Guo, Xie Chen, and Kai Yu. Vqtts: High-fidelity text-to-speech synthesis with self-supervised vq acoustic feature, 2024.

[15] Pai Zhu, Dhruuv Agarwal, Jacob W. Bartel, Kurt Partridge, Hyun Jin Park, and Quan Wang. Synth4kws: Synthesized speech for user defined keyword spotting in low resource environments, 2024.

[16] Prithwiraj Bhattacharjee, Rajan Saha Raju, Arif Ahmad, and M. Shahidur Rahman. End to end bangla speech synthesis, 2021.

[17] Harini s and Manoj G M. Text to speech synthesis, 2024.

[18] Davide Salvi, Brian Hosler, Paolo Bestagini, Matthew C. Stamm, and Stefano Tubaro. Timit-tts: a text-to-speech dataset for multimodal synthetic media detection, 2022.

[19] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis, 2021.

[20] Lam Pham, Phat Lam, Tin Nguyen, Hieu Tang, Dat Tran, Alexander Schindler, Taron Zakaryan, Alexander Polonsky, and Canh Vu. A comprehensive survey with critical analysis for deepfake speech detection, 2024.

[21] Julia Kaiwen Lau, Kelvin Kai Wen Kong, Julian Hao Yong, Per Hoong Tan, Zhou Yang, Zi Qian Yong, Joshua Chern Wey Low, Chun Yong Chong, Mei Kuan Lim, and David Lo. Synthesizing speech test cases with text-to-speech? an empirical study on the false alarms in automated speech recognition testing, 2023.

[22] Rui Liu, Haolin Zuo, De Hu, Guanglai Gao, and Haizhou Li. Explicit intensity control for accented text-to-speech, 2022.

[23] Ingmar Steiner and Sébastien Le Maguer. Creating new language and voice components for the updated marytts text-to-speech synthesis platform, 2018.

[24] Jeff Donahue, Sander Dieleman, Mikołaj Bińkowski, Erich Elsen, and Karen Simonyan. End-to-end adversarial text-to-speech, 2021.

[25] Cheng-I Jeff Lai, Erica Cooper, Yang Zhang, Shiyu Chang, Kaizhi Qian, Yi-Lun Liao, Yung-Sung Chuang, Alexander H. Liu, Junichi Yamagishi, David Cox, and James Glass. On the interplay between sparsity, naturalness, intelligibility, and prosody in speech synthesis, 2021.

[26] Aya Watanabe, Shinnosuke Takamichi, Yuki Saito, Wataru Nakata, Detai Xin, and Hiroshi Saruwatari. Building speech corpus with diverse voice characteristics for its prompt-based representation, 2024.

[27] Daxin Tan and Tan Lee. Fine-grained style modeling, transfer and prediction in text-to-speech synthesis via phone-level content-style disentanglement, 2021.

[28] Ahmet Gunduz, Kamer Ali Yuksel, Kareem Darwish, Golara Javadi, Fabio Minazzi, Nicola Sobieski, and Sebastien Bratieres. An automated end-to-end open-source software for high-quality text-to-speech dataset generation, 2024.

[29] David Ferris. Techniques and challenges in speech synthesis, 2017.

[30] Erica Cooper and Junichi Yamagishi. How do voices from past speech synthesis challenges compare today?, 2021.

[31] Peizhuo Liu, Li Wang, Renqiang He, Haorui He, Lei Wang, Huadi Zheng, Jie Shi, Tong Xiao, and Zhizheng Wu. Spmis: An investigation of synthetic spoken misinformation detection, 2024.

[32] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching, 2024.

[33] Mateusz Łajszczak, Guillermo Cámbara, Yang Li, Fatih Beyhan, Arent van Korlaar, Fan Yang, Arnaud Joly, Álvaro Martín-Cortinas, Ammar Abbas, Adam Michalski, Alexis Moinet, Sri Karlapati, Ewa Muszyńska, Haohan Guo, Bartosz Putrycz, Soledad López Gambino, Kayeon Yoo, Elena Sokolova, and Thomas Drugman. Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data, 2024.

[34] Konstantinos Markopoulos, Nikolaos Ellinas, Alexandra Vioni, Myrsini Christidou, Panos Kakoulidis, Georgios Vamvoukakis, Georgia Maniati, June Sig Sung, Hyoungmin Park, Pirros Tsiakoulis, and Aimilios Chalamandaris. Rapping-singing voice synthesis based on phoneme-level prosody control, 2021.

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

[36] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. Wavegrad 2: Iterative refinement for text-to-speech synthesis, 2021.

[37] Ishika Gupta, Anusha Prakash, Jom Kuriakose, and Hema A. Murthy. Hmm-based data augmentation for e2e systems for building conversational speech synthesis systems, 2022.

[38] Joun Yeop Lee, Myeonghun Jeong, Minchan Kim, Ji-Hyun Lee, Hoon-Young Cho, and Nam Soo Kim. High fidelity text-to-speech via discrete tokens using token transducer and group masked language model, 2024.

[39] Rui Li, Dong Pu, Minnie Huang, and Bill Huang. Unet-tts: Improving unseen speaker and style transfer in one-shot voice cloning, 2022.

[40] Ambuj Mehrish, Navonil Majumder, Rishabh Bhardwaj, Rada Mihalcea, and Soujanya Poria. A review of deep learning techniques for speech processing, 2023.

[41] Paarth Neekhara, Jason Li, and Boris Ginsburg. Adapting tts models for new speakers using transfer learning, 2022.

[42] Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C. Cobo, and Karen Simonyan. High fidelity speech synthesis with adversarial networks, 2019.

[43] John Janiczek, Dading Chong, Dongyang Dai, Arlo Faria, Chao Wang, Tao Wang, and Yuzong Liu. Multi-modal adversarial training for zero-shot voice cloning, 2024.

[44] Hieu-Thi Luong and Junichi Yamagishi. Nautilus: a versatile voice cloning system, 2020.

[45] Florian Lux, Julia Koch, and Ngoc Thang Vu. Exact prosody cloning in zero-shot multispeaker text-to-speech, 2022.

[46] Giuseppe Ruggiero, Enrico Zovato, Luigi Di Caro, and Vincent Pollet. Voice cloning: a multi-speaker text-to-speech synthesis approach based on transfer learning, 2021.

[47] Paarth Neekhara, Shehzeen Hussain, Shlomo Dubnov, Farinaz Koushanfar, and Julian McAuley. Expressive neural voice cloning, 2021.

[48] Akshita Gupta, Tatiana Likhomanenko, Karren Dai Yang, Richard He Bai, Zakaria Aldeneh, and Navdeep Jaitly. Visatronic: A multimodal decoder-only model for speech synthesis. *arXiv preprint arXiv:2411.17690*, 2024.

[49] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-tts: A family of high-quality versatile speech generation models, 2024.

[50] Chunyong Yang, Pengfei Liu, Yanli Chen, Hongbin Wang, and Min Liu. The msxf tts system for icassp 2022 add challenge, 2022.

[51] Vinotha R, Hepsiba D, L. D. Vijay Anand, and Deepak John Reji. Empowering communication: Speech technology for indian and western accents through ai-powered speech synthesis, 2024.

[52] Wen-Chin Huang, Tomoki Hayashi, Xinjian Li, Shinji Watanabe, and Tomoki Toda. On prosody modeling for asr+tts based voice conversion, 2021.

[53] Mingrui Yuan and Zhiyao Duan. Spoofing speaker verification systems with deep multi-speaker text-to-speech synthesis, 2019.

[54] Xinhui Chen, You Zhang, Ge Zhu, and Zhiyao Duan. Ur channel-robust synthetic speech detection system for asvspoof 2021, 2021.

[55] Cemal Hanilci, Tomi Kinnunen, Md Sahidullah, and Aleksandr Sizov. Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise, 2016.

[56] Kirill Borodin, Vasiliy Kudryavtsev, Dmitrii Korzh, Alexey Efimenko, Grach Mkrtchian, Mikhail Gorodnichev, and Oleg Y. Rogov. Aasist3: Kan-enhanced aasist speech deepfake detection using ssl features and additional regularization for the asvspoof 2024 challenge, 2024.

[57] Zhisheng Zhang, Qianyi Yang, Derui Wang, Pengyang Huang, Yuxin Cao, Kai Ye, and Jie Hao. Mitigating unauthorized speech synthesis for voice protection, 2024.

[58] Karren Yang, Ting-Yao Hu, Jen-Hao Rick Chang, Hema Swetha Koppula, and Oncel Tuzel. Text is all you need: Personalizing asr models using controllable speech synthesis, 2023.

[59] Sebastião Quintas, Isabelle Ferrané, and Thomas Pellegrini. Enhancing synthetic training data for speech commands: From asr-based filtering to domain adaptation in ssl latent space, 2024.

[60] Christoph Minixhofer, Ondřej Klejch, and Peter Bell. Evaluating and reducing the distance between synthetic and real speech distributions, 2023.

[61] Nicholas Klein, Tianxiang Chen, Hemlata Tak, Ricardo Casal, and Elie Khoury. Source tracing of audio deepfake systems, 2024.

[62] Zhenyu Zhang, Yewei Gu, Xiaowei Yi, and Xianfeng Zhao. Fmfcc-a: A challenging mandarin dataset for synthetic speech detection, 2021.

[63] Christoph Minixhofer, Ondřej Klejch, and Peter Bell. Ttsds – text-to-speech distribution score, 2024.

[64] Octavian Pascu, Dan Oneata, Horia Cucu, and Nicolas M. Müller. Easy, interpretable, effective: opensmile for voice deepfake detection, 2024.

[65] Haibin Wu, Andy T. Liu, and Hung yi Lee. Defense for black-box attacks on anti-spoofing models by self-supervised learning, 2020.

[66] Junrui Ni, Liming Wang, Heting Gao, Kaizhi Qian, Yang Zhang, Shiyu Chang, and Mark Hasegawa-Johnson. Unsupervised text-to-speech synthesis by unsupervised automatic speech recognition, 2022.

[67] Luigi Attorresi, Davide Salvi, Clara Borrelli, Paolo Bestagini, and Stefano Tubaro. Combining automatic speaker verification and prosody analysis for synthetic speech detection, 2022.

[68] Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney. Generating synthetic audio data for attention-based speech recognition systems, 2020.

[69] Shaojin Ding, Ye Jia, Ke Hu, and Quan Wang. Textual echo cancellation, 2021.

[70] Hao Gu, JiangYan Yi, Chenglong Wang, Yong Ren, Jianhua Tao, Xinrui Yan, Yujie Chen, and Xiaohui Zhang. Utilizing speaker profiles for impersonation audio detection, 2024.

[71] Jee weon Jung, Yihan Wu, Xin Wang, Ji-Hoon Kim, Soumi Maiti, Yuta Matsunaga, Hye jin Shim, Jinchuan Tian, Nicholas Evans, Joon Son Chung, Wangyou Zhang, Seyun Um, Shinnosuke Takamichi, and Shinji Watanabe. Spoofceleb: Speech deepfake detection and sasv in the wild, 2024.

[72] Jie Pu, Yixiong Meng, and Oguz Elibol. Building synthetic speaker profiles in text-to-speech systems, 2022.

[73] Junzuo Zhou, Jiangyan Yi, Tao Wang, Jianhua Tao, Ye Bai, Chu Yuan Zhang, Yong Ren, and Zhengqi Wen. Traceablespeech: Towards proactively traceable text-to-speech with watermarking, 2024.

[74] Rui Liu, Jinhua Zhang, Guanglai Gao, and Haizhou Li. Betray oneself: A novel audio deepfake detection model via mono-to-stereo conversion, 2023.

[75] Yuxiang Zhang, Zhuo Li, Jingze Lu, Wenchao Wang, and Pengyuan Zhang. Synthetic speech detection based on temporal consistency and distribution of speaker features, 2023.

[76] Abdulazeez AlAli and George Theodorakopoulos. An rfp dataset for real, fake, and partially fake audio detection, 2024.

[77] Xin Wang and Junichi Yamagishi. A practical guide to logical access voice presentation attack detection, 2022.

[78] Xinfeng Li, Kai Li, Yifan Zheng, Chen Yan, Xiaoyu Ji, and Wenyuan Xu. Safeear: Content privacy-preserving audio deepfake detection, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.