IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Voice Diarization And Voice Cloning

Pulaganti Varshitha, Likitha DC, Kokkulwar Sruthi, Koushik N, Manjula BS

Student, Student, Student, Student, Asst. Professor
Department of Information Science and Engineering,
Global Academy of Technology, Bengaluru, India

Abstract: Automatic speaker diarization and voice cloning have become essential components in modern speech-driven applications, including intelligent transcription systems, conversational analytics, and personalized audio generation. This paper proposes a unified framework that integrates WhisperX, PyAnnote, and neural Text-to-Speech (TTS) models to achieve highly accurate voice diarization and realistic voice cloning. The system first applies WhisperX for ASR transcription with word-level alignment, ensuring fast and precise text extraction. Speaker segments are detected using PyAnnote's state-of-the-art speaker embedding and clustering techniques, which significantly improve diarization accuracy in multi-speaker environments. Each identified speaker embedding is then mapped to a corresponding voice profile used by a neural TTS model to generate synthetic speech that preserves pitch, timbre, and prosodic characteristics of the original speaker. The integration of these modules produces a seamless pipeline capable of identifying speakers, transcribing speech, and replicating voices with high naturalness. Experiments on dialogue datasets demonstrate notable improvements in diarization error rate, transcription accuracy, and perceptual voice similarity. The proposed system shows potential for applications in personalized virtual assistants, automated content creation, podcast enhancement, call center analytics, and accessibility technologies.

Introduction

In recent years, speech processing technologies have become central to human–computer interaction, multimedia analysis, and automated communication systems. Among these, voice cloning have emerged as crucial research areas due to their ability to understand "who spoke when" and to generate human-like synthetic voices. The combination of accurate diarization and natural voice cloning is essential for building advanced transcription systems, conversational AI, personalized audio applications, and assistive technologies. With the advancement of deep learning models such as WhisperX for high-accuracy transcription, PyAnnote for robust speaker diarization, and modern Text-to-Speech (TTS) architectures for voice cloning, this domain has gained significant importance in both research and industry.The applications of these technologies span across multiple fields. In media and broadcasting, diarization helps segment interviews, podcasts, and debates automatically. In education and accessibility, cloned voices can generate personalized audio lessons or support speech-impaired individuals through voice restoration. Customer service centers use diarization for call analytics, agent evaluation, and conversation summarization. Additionally, voice cloning enables personalized virtual assistants, audiobook generation, and multilingual voice translation. In forensic and security domains, speaker identification and diarization support investigation and audio verification tasks, making the technology widely useful across sectors.Despite these advances, several challenges remain. Speaker diarization still struggles with overlapping speech, background noise, accent variations, and short speaker turns. Similarly, voice cloning requires high-quality data and may fail to capture emotional expression, speaking style, or prosody consistency. Integrating diarization with TTS models introduces additional complexity, especially when mapping speaker embeddings accurately for cloning. Ethical issues such as privacy, misuse of cloned voices, and data security also pose significant concerns. These challenges highlight the need for continuous research, benchmarking, and comparative analysis, making a survey of existing methods essential for understanding current progress and identifying future research directions.

II. LITERATURE REVIEW

1.Speaker Diarization

Speaker diarization has evolved significantly with advanced neural architectures focusing on separation, embedding learning, voice activity detection (VAD), and multi-speaker interaction modeling.

1.1 Neural Separation–Diarization Models (Extended)

Maiti et al. (2024) introduced a breakthrough framework for Multi-Channel Conversational Speaker Separation via Neural Diarization [1]. Their model integrates beamforming, neural separation networks, and attention-driven diarization heads, enabling the system to jointly perform both tasks without treating them as independent modules. This integrated design is particularly effective in real conversational environments where challenges such as overlapping speech, environmental reverberation, and microphone array distortions limit the performance of classical pipelines. By leveraging multichannel spatial cues, the model improves the attribution of speech segments even in high-density speaker interactions. This direction signals the evolution toward holistic, end-to-end neural diarization systems optimized for multi-speaker, real-world scenarios.Building on these advancements, the earlier EEND-SS (End-To-End Neural Diarization with Speech Separation) model by Maiti et al. (2022) [6] established the foundation for handling
flexible numbers of speakers using transformer-based architectures. Unlike traditional clustering-based diarization, which fails under heavy overlap, EEND-SS incorporates speech separation modules inside the diarization model, enabling more accurate segmentation and speaker labeling. Its ability to process complex conversational patterns shows how transformer-based multi-speaker modeling is becoming central in modern diarization research.

1.2 Target Speaker Voice Activity Detection (TSVAD) (Extended)

Chen et al. (2024) introduced Flow TSVAD, a target-oriented diarization system using latent flow matching to model temporal activation of a particular speaker in mixed audio settings [2]. Instead of diarizing all speakers, Flow TSVAD focuses on detecting whether a specific speaker is active, making it extremely useful in call centers, session monitoring, teleconferencing, and personalized voice assistants. This paradigm marks a shift from "who speaks when?" to "is the target speaker speaking now?"—a direction important for applications requiring speaker-specific tracking. By integrating flow-based generative modeling, the system effectively captures speaker dynamics even under high acoustic variability, making it more robust than classical TSVAD systems. This category of research represents a growing interest in speaker-personalized diarization, where the system adapts to individual speaker embeddings and dynamically refines detection based on previously observed speech patterns. Target-driven approaches significantly reduce computation when the entire speaker set is unnecessary, making them valuable for resource-efficient and privacy-preserving voice monitoring systems.

1.3 Metric Learning and Graph-Based Clustering (Extended)

Singh and Ganapathy (2021) introduced a self-supervised metric learning framework for extracting speaker discriminative embeddings without relying on extensive labeled data [3]. Their use of graph clustering improves the cohesion of speaker groups by modeling relationships between embeddings in a non-linear, unsupervised manner. This is particularly relevant for low-resource languages, where labeled diarization datasets are limited. The use of self-supervised training also highlights an increasing movement
toward reducing annotation dependency, which is one of the main constraints in scaling diarization systems globally.Expanding on this direction, Li et al. (2025) used Graph Attention Networks (GAT) combined with Label Propagation Algorithms to identify overlapping speaker communities more efficiently [5]. Traditional diarization struggles with speaker overlapping because embeddings from overlapping segments often lie between multiple speaker clusters. Graph based models mitigate this issue by capturing global structural relationships between embeddings, enabling more robust clustering in multi-speaker conversations. This demonstrates the power of graph neural networks as an emerging paradigm in diarization—one that complements neural embedding extraction with topology-aware speaker modeling.

1.4 Weakly Supervised and Real-Time Systems (Extended)

Thienpondt and Demuynck (2024) proposed a weakly supervised diarization framework that relies on lightweight speaker embeddings derived from VAD-guided training [4]. Their approach significantly reduces inference computation, enabling diarization on edge devices and mobile platforms where memory, latency, and power are constrained. Weakly supervised learning also decreases reliance on fully annotated datasets, using coarse labels to train speaker embedding networks efficiently. This trend highlights a broader shift toward scalable diarization systems designed to run in real-world deployments such as smart home devices, wearable assistants, and low-latency speech interfaces.The move toward real-time diarization also reveals industry demand for on-device privacy, low-latency responses, and efficient resource management. Weak supervision broadens the feasibility of deploying diarization systems in environments where computational overhead must be minimized without sacrificing accuracy.

1.5 End-to-End Speaker-Attributed ASR (Extended)

Kanda et al. (2021) introduced the Transcribe-to-Diarizeframework, a paradigm shift integrating ASR and diarization inside the decoding process [7]. By embedding diarization within automatic speech recognition, the system eliminates the multi-step pipeline of VAD → Embedding → Clustering, which often suffers from error propagation. Instead, diarization becomes a by-product of transcription, making it more consistent and more accurate for use cases such as meeting transcription, lecture note generation, and multi-speaker documentation.This work represents a significant trend toward unified ASR–Diarization models driven by transformer-based architectures. These systems are capable of handling unlimited numbers of speakers, adapt to dynamic conversational structures, and deliver higher accuracy in real-time transcription scenarios. Such approaches indicate the future direction for diarization—where tasks are not isolated but fused into a single, end-to-end optimized framework that mirrors real conversational speech dynamics.

2. Voice Cloning and TTS

Voice cloning research has rapidly progressed due to zero shot learning, multi-style modeling, adversarial training, and multi-reference architectures.

2.1 Zero-Shot Voice Cloning Models

Gao et al. (2023) developed U-Style, a cascading U-net architecture that integrates multi-level speaker and style modeling [8]. The system enables zero-shot speaker cloning with expressive style transfer using only a few seconds of reference audio. This model shows how multi-scale feature extraction improves both timbre replication and prosodic variation.Li et al. (2024) presented a multi-modal adversarial training framework for zero-shot voice cloning, combining text and audio conditioning [9]. By training adversarially, the model prevents identity leakage, improves robustness to noise, and generates more consistent speaker embeddings. This addresses a common limitation in zero-shot cloning: generalization to unseen speakers.

2.2 Security and Anti-Clone Research

Liu et al. (2025) introduced CloneShield [10], a landmark contribution focusing on defense against unauthorized cloning. The authors propose a universal perturbation that can be applied to any audio file to prevent neural TTS models from accurately cloning the speaker's identity. This work signals the beginning of speech privacy protection research, which is essential as voice cloning systems become more accessible.

2.3 Style Adaptation and Prosody Modeling

Gupta and Singh (2025) presented DS-TTS, a model that introduces Dynamic Dual-Style Feature Modulation for zeroshot adaptation [11]. The dual-style mechanism allows the system to control both speaker identity and emotional style independently, creating more natural and expressive cloned voices. This contributes to improved control over prosody, emotion, speaking rate, and expressiveness.

2.4 Multi-Reference Learning for Stable Cloning

Kumar et al. (2024) proposed MRMI-TTS, a mutual information-driven TTS system that incorporates multiple reference audio samples for cloning [12]. Multiple references allow the model to capture variations in pitch, tone, and rhythm, significantly increasing stability and speaker similarity. This resolves limitations in single-reference zeroshot systems, which often struggle to maintain consistent identity across longWith the combination of deep learning and self-supervised techniques, voice diarization- the process of figuring out

"who spoke when" in an audio recording- has undergone significant change. Numerous researches have made contributions to increasing the precision, effectiveness, and adaptability of diarization in practical settings.
Comparative Table of Speaker Diarization & Voice Cloning:

Paper No.	Authors & Year	Method / Model Used	Limitations
[1]	Maiti et al., 2024	Multi-Channel Neural Diarization + Speech Separation	High computational cost; difficult for real-time use
[2]	Chen et al., 2024	Flow TSVAD using Latent Flow Matching	Works only for target-speaker scenarios. requires target embedding
[3]	Singh & Ganapathy, 2021	Self-Supervised Metric Learning + Graph Clustering	Accuracy depends on embedding quality; weak for overlapping speech
[4]	Thienpondt & Demuynck, 2024	Weakly Supervised VAD + Lightweight Speaker Embeddings	Less robust during overlapping or noisy conditions
[5]	Li et al., 2025	Graph Attention Networks + Label Propagation.	Higher training and graph Construction cost.
[6]	Maiti et al., 2022	EEND-SS (Joint End-to-End Diarization + Separation)	Requires high GPU resources; struggles on very long audio
[7]	Kanda et al., 2021	Transcribe-to-Diarize (ASR + Diarization Integration)	Strongly depends on ASR accuracy. errors propagate
[8]	Gao et al., 2023	U-Style (Cascading U-Nets + multi-level Style Modeling)	Struggles with strong emotions or heavy accents
[9]	Li et al., 2024	Multi-Modal Adversarial TTS Training	Complex training; high resource requirement
[10]	Liu et al., 2025	CloneShield (Universal Perturbation Defense)	Causes slight audio distortion; not generative model
[11]	Gupta & Singh, 2025	DS-TTS (Dynamic Dual-Style Feature Modulation)	Requires careful tuning; higher inference time
[12]	Kumar et al., 2024	MRMI-TTS (MultiReference + Mutual Information)	Needs multiple high-quality reference audios; slower training

III. RESEARCH METHODOLOGY

Despite significant progress in voice diarization and cloning, current systems still face limitations in accuracy, robustness, and adaptability. Modern diarization models like pyannote.audio and WhisperX achieve strong performance on clean audio, but their accuracy drops in noisy, real-world environments such as classrooms, meetings, or outdoor settings. Existing models struggle with overlapping speech, spontaneous conversation, emotional speech, and Multilingual dialogues—scenarios common in practical applications. Additionally, most available datasets do not fully capture these real-world complexities, limiting the generalization ability of current diarization approaches.Another major research gap lies in the integration of diarization with high-quality voice cloning pipelines. Although TTS systems can generate natural-sounding speech, maintaining speaker identity when the cloned voice

is generated from diarized segments remains a challenge. The diarization step often introduces segmentation errors, which negatively affect embedding extraction and voice similarity in the cloning phase. Current frameworks lack end-to-end optimization, meaning diarization, speaker embedding extraction, and synthesis are treated as separate tasks, leading to cumulative errors and inconsistent results. There is also limited research on achieving high fidelity cloning from very short or noisy audio samples.Finally, there is a lack of systematic evaluation frameworks that jointly measure diarization quality, speaker identity preservation, naturalness, and real-time efficiency. Research rarely addresses computational challenges, such as enabling lightweight models suitable for mobile or edge devices. Ethical and security concerns—like voice spoofing, misuse, consent, and identity protection—are also under-explored in current literature. Overall, these gaps highlight the need for improved multi-speaker datasets, robust end-to-end pipelines, cross-lingual support, real-time optimized models, and stronger ethical guidelines to support the next generation of diarization and voice cloning systems.

IV. FUTURE SCOPE

The field of speech processing is evolving rapidly, and the proposed system offers several promising avenues for future development. One major direction is improving diarization accuracy in highly complex acoustic environments. Current models face difficulty in scenarios involving overlapping speech, background noise, multilingual conversations, and spontaneous dialogues. Future work can integrate advanced transformer-based architectures, self-supervised models like WavLM or MMSpeech, and large audio-language models to significantly enhance segment separation and minimize speaker confusion. Additionally, real-time diarization with near-zero latency remains an important research direction for applications such as live broadcasting, meeting transcription, and intelligent assistants. Another key area of enhancement involves the voice cloning pipeline. While current TTS models can generate high-quality synthetic voices, they often struggle with emotional variability, long-term consistency, and maintaining natural prosody. Future research can focus on expressive voice cloning, where models learn emotional cues, conversational tone, and personality traits. Extending cloning capabilities to low-resource languages and code-mixed speech will also increase system usability, especially in multilingual countries like India. Incorporating reinforcement learning and prosody conditioned generative models can help achieve more human-like speech output.Finally, the project can evolve into a scalable end-to-end platform integrating diarization, transcription, and cloning for intelligent human–machine interaction. Future systems may support personalized AI agents that adapt to user voices, help in accessibility tools (e.g., voice for people with speech impairment), and enable more secure authentication through voice biometrics. Ethical considerations such as deepfakedetection, secure voice identity handling, and consent-based usage will also drive future advancements. Overall, the project serves as a foundation for developing highly intelligent, multilingual, ethical, and user-adaptive speech processing systems WhereESS$_0$iserror sum of squares of APT, ESS$_1$iserror sum of squares of CAPM, Nisnumber of observations, K$_0$is number of independent variables of the APT and K$_1$ isnumber of independent variables of the CAPM.As according to the ratio when:

R> 1 means CAPM is more strongly supported by data under consideration than APT.
R < 1 means APT is more strongly supported by data under consideration than CAPM.

V. RESULTS AND DISCUSSION

This project successfully demonstrates an integrated system for voice diarization, speaker identification, and voice cloning, leveraging state-of-the-art frameworks such as WhisperX for accurate ASR alignment, PyAnnote for robust diarization, and modern TTS models for high-quality speech synthesis. The combination of these technologies allows precise segmentation of multi-speaker audio, reliable attribution of speech segments to individual voices, and the ability to generate natural-sounding synthetic speech that
closely resembles the target speaker. Through experimental evaluation, the system shows strong performance in noisy, real-world audio conditions, validating its practical applicability.
Overall, the project addresses key challenges in multispeaker audio processing, including overlap detection, alignment precision, and speaker embedding consistency. By integrating powerful deep learning models and optimizing workflow pipelines, the system significantly enhances the accuracy and reliability of conversational audio analysis. The developed framework serves as a useful foundation for building advanced applications such as automated meeting summarizers, personalized virtual assistants, and voice-based authentication systems.In summary, the project contributes meaningfully to the growing field of speech

technologies by demonstrating an efficient pipeline for diarization and voice cloning. It highlights both the potential and limitations of current deep
learning approaches, laying the groundwork for future improvements using larger datasets, multimodal learning, and real-time deployment techniques. The work establishes a strong baseline that can be extended for academic, industrial, and real-world communication applications**Acknowledgment**

Thepreferredspellingoftheword "acknowledgment" inAmericaiswithoutan "e" afterthe "g".Avoidthestiltedexpression,
"Oneofus(R.B.G.)thanks..."Instead,try"R.B.G.thanks".Putapplicablesponsoracknowledgmentshere;DONOTplacethemonthefirstpageofyourpaperorasafootnote.

REFERENCES

[1] S. Maiti, Y. Ueda, S. Watanabe, C. Zhang, and M. Yu, "Multi-Channel Conversational Speaker Separation via Neural Diarization," arXiv, Apr. 2024.

[2] Z. Chen, B. Han, S. Wang, Y. Jiang, and Y. Qian, "Flow-TSVAD: Target-Speaker Voice Activity Detection via Latent Flow Matching," arXiv, Sept. 2024.

[3] P. Singh and S. Ganapathy, "Self-Supervised Metric Learning With Graph Clustering For Speaker Diarization," arXiv, Sept. 2021.

[4] J. Thienpondt and K. Demuynck, "Speaker Embeddings With Weakly Supervised Voice Activity Detection For Efficient Speaker Diarization," Odyssey, May 2024.

[5] Z. Li, J. Wang, X. Li, W. Li, L. Luo, L. Li, and Q. Hong, "Speaker Diarization with Overlapping Community Detection Using Graph Attention Networks and Label Propagation Algorithm," arXiv, Jun. 2025.

[6] S. Maiti, Y. Ueda, S. Watanabe, C. Zhang, and M. Yu, "EEND-SS: Joint End-to-End Neural Speaker Diarization and Speech Separation for Flexible Number of Speakers," IEEE SLT, 2022.

[7] N. Kanda, X. Xiao, Y. Gaur, X. Wang, Z. Meng, Z. Chen, T. Yoshioka, "Transcribe-to-Diarize: Neural Speaker Diarization for Unlimited Number of Speakers using End-to-End Speaker-Attributed ASR," arXiv, Oct. 2021.

[8] S. Gao, J. Huang, J. Wu, X. Xiao, and S. Watanabe, "U-Style: Cascading U-nets with Multi-level Speaker and Style Modeling for Zero-Shot Voice Cloning," arXiv, Oct. 2023.

[9] H. Li, J. Wang, Y. Zhang, L. He, and Y. Qian, "Multi-modal Adversarial Training for Zero-Shot Voice Cloning," arXiv, Aug. 2024.

[10] Z. Liu, Y. Shi, M. Hong, H. Wang, and C. Zhao, "CloneShield: Universal Perturbation Against Zero-Shot Voice Cloning," arXiv, May 2025.

[11] A. Gupta and M. Singh, "DS-TTS: Zero-Shot Speaker Style Adaptation via Dynamic Dual-Style Feature Modulation," arXiv, June 2025.

[12] R. Kumar, J. Park, L. Wang, and T. Hayashi, "MRMI-TTS: MultiReference Audios and Mutual Information Driven Zero-Shot Voice Cloning," arXiv, Jun. 2024.