Novelty Assessment Report

Paper: VibeVoice: Expressive Podcast Generation with Next-Token Diffusion

PDF URL: https://openreview.net/pdf?id=FihSkzyxdv

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Generating long-form, multi-speaker conversational audio like podcasts poses significant challenges for traditional Text-to-Speech (TTS) systems, particularly in scalability, speaker consistency, and natural turn-taking. We present VibeVoice , a novel model designed to synthesize expressive, long-form speech with multiple speakers in a zero-shot manner. A core component of our approach is the continuous speech tokenizers operating at an ultra-low frame rate of 7.5. This tokenizer effectively preserves audio fidelity while significantly boosting computational efficiency for processing long sequences. To facilitate training on authentic conversational dynamics, we have developed an annotation pipeline that generates pseudo transcriptions and turn-taking labels for extensive podcast data. Leveraging this data and our efficient tokenizer, VibeVoice employs the next-token diffusion framework. This enables VibeVoice to: (1) synthesize long-form speech (up to 30 minutes) with up to 4 speakers, surpassing the typical 1-2 speaker limits of many prior models; and (2) achieve a high degree of naturalness in turn-taking, pacing, and the rendition of subtle non-lexical cues (such as breaths and lip smacks), which are crucial for listener immersion and capturing the authentic vibe of expressive conversations.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Expressive Long-Form Multi-Speaker Conversational Speech Synthesis**

A total of **23 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Conversational Dialogue Speech Generation**
- **Emotion and Style Transfer in Multi-Speaker TTS**
- **Context-Aware Long-Form Speech Synthesis**
- **Conversational Speech Processing and Understanding**
- **Dialogue Response Generation with Emotion**
- **Personalized Speech and Dialog Modeling**

Complete Taxonomy Tree

- Expressive Long-Form Multi-Speaker Conversational Speech Synthesis Survey Taxonomy
- Conversational Dialogue Speech Generation
  - Podcast-Style Long-Form Dialogue Synthesis ★ (4 papers)
  - [0] VibeVoice: Expressive Podcast Generation with Next-Token Diffusion (Anon et al., 2026) View paper
  - [2] FireRedTTS-2: Towards Long Conversational Speech Generation for Podcast and Chatbot (Xie Kun, 2025) View paper
  - [5] SoulX-Podcast: Towards Realistic Long-form Podcasts with Dialectal and Paralinguistic Diversity (LIN Haopeng, 2025) View paper
  - [6] Podagent: A comprehensive framework for podcast generation (Xiao Yujia, 2025) View paper
  - Interactive Dual-Speaker Dialogue Synthesis (2 papers)
  - [1] Interactive Conversational Head Generation (Mohan Zhou, 2023) View paper
  - [14] DialoSpeech: Dual-Speaker Dialogue Generation with LLM and Flow Matching (Hanke Xie, 2025) View paper
  - Agent-Based Multi-Party Dialogue Generation (1 papers)
  - [3] DialogueAgents: A Hybrid Agent-Based Speech Synthesis Framework for Multi-Party Dialogue (Xiang Li, 2025) View paper
- Emotion and Style Transfer in Multi-Speaker TTS
  - Contrastive Learning for Emotion and Style Disentanglement (2 papers)
  - [7] Boosting Multi-Speaker Expressive Speech Synthesis with Semi-Supervised Contrastive Learning (Xinfa Zhu, 2023) View paper
  - [18] Multi-Speaker Expressive Speech Synthesis via Semi-supervised Contrastive Learning (Zhu Xin-fa, 2023) View paper
  - Multi-Factor Decoupling for Cross-Speaker Transfer (2 papers)
  - [4] Controllable Multi-Speaker Emotional Speech Synthesis With an Emotion Representation of High Generalization Capability (Junjie Zheng, 2025) View paper
  - [20] Multi-Speaker Expressive Speech Synthesis via Multiple Factors Decoupling (Xinfa Zhu, 2023) View paper
  - Prosody-Controllable Multi-Speaker Synthesis (2 papers)
  - [15] Acquiring conversational speaking style from multi-speaker spontaneous dialog corpus for prosody-controllable sequence-to-sequence speech synthesis (Slava Shechtman, 2021) View paper
  - [16] Cross-Speaker Style Transfer for TTS with Singing Voice Conversion Data Augmentation, Style Filtering, and F0 Matching (Marques, 2024) View paper
  - Fine-Grained Prosody and Emotion Modeling (2 papers)
  - [21] End-to-End Multi-speaker Speech Synthesis with Controllable Stress (Ting Liang, 2022) View paper
  - [23] Multi-Speaker Emotional Speech Synthesis with Fine-Grained Prosody Modeling (Chunhui Lu, 2021) View paper

- Context-Aware Long-Form Speech Synthesis
  - Multicast Audiobook Generation with Context Modeling (1 papers)
  - [12] Audiobook-CC: Controllable Long-context Speech Generation for Multicast Audiobook (Liu Min, 2025) View paper
  - Narrator and Character Voice Modulation (1 papers)
  - [9] Narrator or Character: Voice Modulation in an Expressive Multi-speaker TTS (Tankala Pavan Kalyan, 2023) View paper
- Conversational Speech Processing and Understanding
  - Multi-Speaker Diarization and Recognition (2 papers)
  - [11] Multi-Channel Conversational Speaker Separation via Neural Diarization (Hassan Taherian, 2024) View paper
  - [22] Speech Recognition and Multi-Speaker Diarization of Long Conversations (Huanru Henry Mao, 2020) View paper
  - Emotion-Aware Dialogue Summarization and Understanding (2 papers)
  - [10] Spoken DialogSum: An Emotion-Rich Conversational Dataset for Spoken Dialogue Summarization (Yen-Ju Lu, 2025) View paper
  - [13] A long-duration Speech Semantic Recognition and Summarization Model for multi-speaker Conversations (Mingzhen Song, 2024) View paper
- Dialogue Response Generation with Emotion
  - Emotion-Enhanced Multi-Turn Response Generation (1 papers)
  - [8] Incorporating emotion for response generation in multi-turn dialogues (Yanying Mao, 2022) View paper
  - Listener Vocalization Synthesis (1 papers)
  - [19] Synthesis of listener vocalizations (Pammi, 2011) View paper
- Personalized Speech and Dialog Modeling (1 papers)
  - [17] Deep Generative Models for Personalized Speech and Spoken Dialog Modeling (ê¹□í□¬ì□¹, 2025) View paper

Narrative

Core task: expressive long-form multi-speaker conversational speech synthesis. The field organizes around several complementary branches that address different facets of generating natural, emotionally rich dialogue. Conversational Dialogue Speech Generation focuses on producing multi-turn exchanges with appropriate turn-taking and prosodic cues, while Emotion and Style Transfer in Multi-Speaker TTS emphasizes controlling affective dimensions and speaker identity. Context-Aware Long-Form Speech Synthesis tackles the challenge of maintaining coherence and naturalness over extended durations, and Conversational Speech Processing and Understanding provides the analysis tools needed to model dialogue structure. Dialogue Response Generation with Emotion and Personalized Speech and Dialog Modeling round out the taxonomy by addressing content planning and speaker-specific adaptation. Representative works span from early listener vocalization studies to recent systems like FireRedTTS[2] and DialogueAgents[3] that integrate multiple expressive dimensions.

A particularly active line of work centers on podcast-style long-form dialogue synthesis, where systems must balance naturalness, speaker consistency, and narrative flow over minutes rather than seconds. VibeVoice[0] sits squarely within this emerging cluster, alongside SoulX Podcast[5] and Podagent[6], all of which tackle the challenge of generating extended multi-speaker conversations with appropriate emotional arcs and turn-taking dynamics. Compared to SoulX Podcast[5], which emphasizes content-driven narrative structure, VibeVoice[0] appears to focus more directly on the expressive acoustic modeling required for sustained conversational realism. Meanwhile, works like Controllable Emotional Speech[4] and DialoSpeech[14] explore finer-grained control over prosody and emotion within shorter dialogue contexts, highlighting an ongoing tension between expressiveness at the utterance level versus coherence across long-form interactions. The field continues to grapple with how to scale emotional richness and speaker variability to podcast-length scenarios without sacrificing naturalness.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. FireRedTTS-2: Towards Long Conversational Speech Generation for Podcast and Chatbot

Authors: Xie Kun, Shen Feiyu, Kun Xie, Li Junjie, Feiyu Shen, et al. (14 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Current dialogue generation approaches typically require the complete dialogue text before synthesis and produce a single, inseparable speech containing all voices, making them unsuitable for interactive chat; moreover, they suffer from unstable synthesis, inaccurate speaker transitions, and incoherent prosody. In this work, we present FireRedTTS-2, a long-form streaming TTS system for multi-speaker dialogue generation, delivering stable, natural speech with reliable speaker switching and contex...

Relationship Analysis

Both papers belong to the Podcast-Style Long-Form Dialogue Synthesis category, focusing on generating extended multi-speaker conversational speech with natural turn-taking and prosody. They overlap in addressing scalability challenges for long-form podcast generation, employing efficient speech tokenizers (VibeVoice uses 7.5Hz continuous tokens, FireRedTTS-2 uses 12.5Hz streaming tokens), and leveraging LLM-based architectures for dialogue modeling. The key difference is that FireRedTTS-2 emphasizes streaming generation for real-time interactive chat applications with a text-speech interleaved format and dual-transformer architecture, while VibeVoice focuses on next-token diffusion with hybrid acoustic-semantic representations for offline podcast synthesis up to 90 minutes with 4 speakers.

2. SoulX-Podcast: Towards Realistic Long-form Podcasts with Dialectal and Paralinguistic Diversity

Authors: LIN Haopeng, Hanke Xie, Cao Wenxiao, Haopeng Lin, Tian Wen-jie, et al. (39 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Recent advances in text-to-speech (TTS) synthesis have significantly improved speech expressiveness and naturalness. However, most existing systems are tailored for single-speaker synthesis and fall short in generating coherent multi-speaker conversational speech. This technical report presents SoulX-Podcast, a system designed for podcast-style multi-turn, multi-speaker dialogic speech generation, while also achieving state-of-the-art performance in conventional TTS tasks. To meet the higher nat...

Relationship Analysis

Both papers belong to the Podcast-Style Long-Form Dialogue Synthesis category, focusing on generating extended multi-speaker conversational audio with naturalness. They overlap in addressing long-form podcast generation (both claim 90+ minute capability), multi-speaker synthesis, and expressive conversational dynamics including turn-taking and paralinguistic cues. The key differences are that VibeVoice employs a next-token diffusion framework with continuous speech tokenizers at 7.5 Hz and hybrid acoustic-semantic representations, while SoulX-Podcast emphasizes dialectal diversity (Mandarin, English, and Chinese dialects like Sichuanese, Henanese, Cantonese) and paralinguistic controls for personalized podcast generation.

3. Podagent: A comprehensive framework for podcast generation

Authors: Xiao Yujia, He Lei, Yujia Xiao, Guo, Haohan, et al. (13 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Existing Existing automatic audio generation methods struggle to generate podcast-like audio programs effectively. The key challenges lie in in-depth content generation, appropriate and expressive voice production. This paper proposed PodAgent, a comprehensive framework for creating audio programs. PodAgent 1) generates informative topic-discussion content by designing a Host-Guest-Writer multi-agent collaboration system, 2) builds a voice pool for suitable voice-role matching and 3) utilizes LL...

Relationship Analysis

Both papers belong to the Podcast-Style Long-Form Dialogue Synthesis category, focusing on generating extended multi-speaker conversational audio with naturalness. They overlap in addressing podcast generation challenges including multi-speaker synthesis, natural turn-taking, and expressive speech quality. However, VibeVoice emphasizes a next-token diffusion framework with continuous speech tokenizers operating at 7.5 Hz for scalable synthesis up to 90 minutes with 4 speakers, while PodAgent focuses on a multi-agent LLM system (Host-Guest-Writer) for content generation, voice-role matching through characteristic analysis, and LLM-enhanced instruction-following TTS for expressiveness.

Contributions Analysis

Overall novelty summary. VibeVoice contributes a framework for generating expressive multi-speaker podcast-style conversations up to 30 minutes with up to four speakers, using ultra-low frame rate continuous speech tokenizers (7.5 fps) and next-token diffusion. The paper sits within the 'Podcast-Style Long-Form Dialogue Synthesis' leaf, which contains four papers total including VibeVoice itself. This represents a relatively sparse but emerging research direction, suggesting the problem of extended multi-speaker conversational synthesis remains under-explored compared to shorter-form or single-speaker tasks.

The taxonomy reveals that podcast-style synthesis is one specialized branch under 'Conversational Dialogue Speech Generation', which also includes interactive dual-speaker dialogue and agent-based multi-party generation. Neighboring branches address emotion/style transfer, context-aware long-form synthesis (audiobooks, storytelling), and conversational processing/understanding. VibeVoice's focus on podcast-length naturalness and turn-taking connects it to context-aware synthesis approaches, but its emphasis on zero-shot multi-speaker generation distinguishes it from audiobook methods that typically assume character-level control. The taxonomy's scope notes clarify that podcast-style work must handle extended duration and naturalness, separating it from shorter interactive dialogue systems.

Among 24 candidates examined, the VibeVoice framework contribution shows one refutable candidate out of four examined, while the ultra-low frame rate tokenizer contribution has four refutable candidates among ten examined. The annotation pipeline contribution appears more novel, with zero refutable candidates among ten examined. These statistics suggest that while the overall framework and tokenizer design have some overlapping prior work within the limited search scope, the specific approach to generating pseudo transcriptions and turn-taking labels for podcast data may represent a less-explored methodological contribution. The search scale (24 candidates) indicates this assessment is based on top-K semantic matches rather than exhaustive field coverage.

Given the limited search scope and the sparse taxonomy leaf (four papers), VibeVoice appears to address an emerging problem space where prior work is still accumulating. The framework and tokenizer contributions show moderate overlap with examined candidates, while the annotation pipeline shows less. The analysis covers top semantic matches and does not claim exhaustive coverage of all relevant speech synthesis literature, particularly work outside the podcast-style conversational domain.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: VibeVoice framework for expressive multi-speaker podcast generation

Description: The authors introduce VibeVoice, a framework that synthesizes expressive, long-form conversational audio (up to 90 minutes) with up to 4 speakers in a zero-shot setting. It employs a next-token diffusion architecture integrated with an LLM to achieve natural turn-taking, pacing, and subtle non-lexical cues crucial for authentic conversational dynamics.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. CoVoMix2: Advancing Zero-Shot Dialogue Generation with Fully Non-Autoregressive Flow Matching

URL: View paper

Brief Assessment

CoVoMix2[45] focuses on zero-shot multi-talker dialogue generation using flow-matching methods, while VibeVoice employs a next-token diffusion architecture integrated with an LLM. The technical approaches differ fundamentally in their generative frameworks and architectural designs.

2. SLAM-Omni: Timbre-Controllable Voice Interaction System with Single-Stage Training

URL: View paper

Brief Assessment

SLAM-Omni[47] focuses on real-time voice interaction systems with timbre control for dialogue applications, not on expressive multi-speaker podcast generation with conversational dynamics like turn-taking and non-lexical cues that VibeVoice addresses.

3. CoVoMix: Advancing zero-shot speech generation for human-like multi-talker conversations

URL: View paper

Prior Art Analysis

CoVoMix[44] demonstrates that prior work exists for zero-shot, multi-speaker, long-form conversational speech generation with natural turn-taking and expressive features. The candidate paper explicitly addresses the same core challenges: generating multi-speaker dialogues with natural turn-taking, managing speaker consistency across extended conversations, and capturing spontaneous behaviors like laughter and overlapping speech. CoVoMix[44] was published at NeurIPS 2024, predating the ORIGINAL paper's ICLR 2026 submission, and presents a complete system for generating human-like multi-talker conversations in zero-shot settings with similar technical objectives.

Evidence

Evidence 1 - **Rationale**: Both papers claim novelty for zero-shot multi-speaker speech generation systems. CoVoMix[44] explicitly presents this capability before the ORIGINAL paper's submission. - **Original**: we present vibevoice, a novel model designed to synthesize expressive, long-form speech with multiple speakers in a zero-shot manner - **Candidate**: we introduce cov omix:conversational voice mixture generation, a novel model for zero-shot, human-like, multi-speaker, multi-round dialogue speech generation

Evidence 2 - **Rationale**: Both papers identify the same core challenge of generating natural multi-speaker conversations beyond simple concatenation. CoVoMix[44] presents a solution to this problem before the ORIGINAL paper. - **Original**: generating convincing multi-speaker conversations introduces unique hurdles. while traditional systems can technically produce multi-speaker, long-form audio by concatenating individually synthesized utterances (as demonstrated in our experiments), achieving naturalness in speaker interaction remain... - **Candidate**: in this paper, we introduce cov omix:conversational voice mixture generation, a novel model for zero-shot, human-like, multi-speaker, multi-round dialogue speech generation. cov omix first converts dialogue text into multiple streams of discrete tokens, with each token stream representing semantic i...

Evidence 3 - **Rationale**: Both papers present end-to-end architectures for multi-speaker dialogue generation. CoVoMix[44] demonstrates this approach was already established in the field. - **Original**: vibevoiceleverages an end-to-end llm-based architecture with a diffusion head, drawing inspiration from latentlm (sun et al., 2024). the llm (yang et al., 2024) handles robust textual understanding and dialogue flow, while the diffusion head (li et al., 2024b) ensures high-fidelity acoustic generati... - **Candidate**: our proposed cov omix, shown in figure 1, consists of a multi-stream text-to-semantic model, an acoustic model and a vocoder. the text-to-semantic model first generates multi-stream semantic token sequences for each speaker, given the dialogue transcription. then the acoustic model transforms these ...

Evidence 4 - **Rationale**: Both papers describe data processing pipelines for conversational speech with spontaneous behaviors. CoVoMix[44] demonstrates this methodology was already established. - **Original**: moreover, we developed a data processing pipeline that curates and annotates raw podcast data. this provides rich, naturalistic training material, enabling vibevoiceto learn realistic intonation, turn-taking, and subtle expressive cues, thereby enhancing perceived audio realism - **Candidate**: spontaneous behavior such as laughter is labeled by [laughter] token in the transcription. for dialogue, we slice long dialogues into shorter, stereo-channel dialogues containing at least two utterances from distinct speakers. we ensure that the first and last sentences of each processed dialogue do...

4. Character-Driven Narrative Generation for Scene-Based Video Synthesis

URL: View paper

Brief Assessment

Character-Driven Narrative[48] focuses on generating character-driven dialogue and speech for scene-based video storytelling, not on zero-shot multi-speaker podcast synthesis. The candidate addresses narrative generation from visual prompts, while the original paper tackles long-form conversational audio generation with natural turn-taking dynamics.

Contribution 2: Ultra-low frame rate continuous speech tokenizers

Description: The authors develop specialized acoustic and semantic tokenizers that both operate at an ultra-low frame rate of 7.5 Hz. The acoustic tokenizer uses a sigma-VAE to preserve audio fidelity while the semantic tokenizer extracts linguistic content, together forming a hybrid representation that significantly boosts computational efficiency for long sequences.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. TASLA: Text-Aligned Speech Tokens with Multiple Layer-Aggregation

URL: View paper

Brief Assessment

TASLA[42] operates at approximately 2.62 Hz with text-aligned discrete tokens using FSQ, while the original paper uses continuous σ-VAE tokens at 7.5 Hz for acoustic representation. These are fundamentally different tokenization approaches (discrete vs. continuous) with different frame rates and architectural goals.

2. Kimi-audio technical report

URL: View paper

Prior Art Analysis

Kimi-audio[36] demonstrates prior work on ultra-low frame rate speech tokenizers operating at 12.5 Hz, which predates the original paper's claim of novelty for 7.5 Hz tokenizers. Both papers employ similar architectural principles: discrete semantic tokens combined with continuous acoustic features for hybrid representation. Kimi-audio[36] explicitly describes using '12.5hz audio tokenizer' with 'discrete semantic tokens' and 'continuous acoustic vectors' to achieve efficient compression while preserving audio fidelity. The original paper's contribution of 7.5 Hz tokenizers with sigma-VAE for acoustic representation and separate semantic tokenizers follows the same design philosophy already established in Kimi-audio[36], which was published earlier (arxiv:2504.18425v1 dated 25 Apr 2025 vs. the original paper under review at ICLR 2026).

Evidence

Evidence 1 - **Rationale**: Both papers claim novelty in ultra-low frame rate tokenizers. Kimi-audio[36] operates at 12.5 Hz with hybrid representation (continuous + discrete), while the original operates at 7.5 Hz. The architectural principle of combining semantic and acoustic features at low frame rates is established in Kimi-audio[36]. - **Original**: a cornerstone of vibevoiceis its efficient hybrid speech representation strategy, derived from specialized acoustic and semantic tokenizers, both operating at an ultra-low frame rate of 7.5 hz. the acoustic tokenizer aggressively compresses audio while preserving remarkable reconstruction fidelity, ... - **Candidate**: we leverage a 12.5hz audio tokenizer, design a novel llm-based architecture with continuous features as input and discrete tokens as output, and develop a chunk-wise streaming detokenizer based on flow matching.

Evidence 2 - **Rationale**: Both papers use hybrid tokenization combining discrete semantic tokens with continuous acoustic features. Kimi-audio[36] explicitly describes this strategy before the original paper's submission, establishing prior art for the hybrid approach. - **Original**: acoustic tokenizeradopts the principles of a variational autoencoder (v ae) (kingma & welling, 2014), specifically drawing inspiration from theσ-v ae variant proposed in latentlm (sun et al., 2024) to mitigate potential variance collapse issues of v aes when used in autoregressive modeling settings. - **Candidate**: our audio foundation model employs a hybrid audio tokenization strategy, integrating discrete semantic tokens and complementary continuous vectors of acoustic information to effectively represent speech signals for downstream tasks. this tokenization allows the model to leverage the efficiency and s...

Evidence 3 - **Rationale**: Both papers use ASR-based proxy tasks for semantic tokenization at ultra-low frame rates. Kimi-audio[36] describes using ASR-derived semantic tokens at 12.5 Hz, establishing this approach before the original paper's 7.5 Hz semantic tokenizer. - **Original**: semantic tokenizermirrors the hierarchical architecture of the acoustic tokenizer's encoder, but without v ae components, as its objective is deterministic content-centric feature extraction. the main difference is the training objective, which uses automatic speech recognition (asr) as the proxy ta... - **Candidate**: this component utilizes a supervised speech tokenizer derived from an automatic speech recognition (asr) model. by introducing a vector quantization layer within the whisper encoder architecture [58], we can transform continuous speech representations into a sequence of discrete tokens at a low fram...

Evidence 4 - **Rationale**: Both papers emphasize the efficiency gains from ultra-low frame rate tokenization for long-form audio processing. Kimi-audio[36] demonstrates downsampling to 12.5 Hz for computational efficiency, establishing the principle that the original paper applies at 7.5 Hz. - **Original**: this decoupled design, with both components leveraging the same highly efficient frame rate, allows for optimized acoustic and semantic feature extraction. these features are then combined to form the rich, yet compact, hybrid input essential for managing long-form content within our generative mode... - **Candidate**: complementing the discrete semantic tokens, we incorporate a continuous feature representation derived from a pre-trained whisper model [58] to enhance the perception capability of our model. since the whisper feature has a frame rate of 50hz, we additionally introduce an adaptor upon the whisper fe...

3. LM-SPT: LM-Aligned Semantic Distillation for Speech Tokenization

URL: View paper

Brief Assessment

LM-SPT[43] focuses on discrete semantic tokens at various frame rates (25Hz, 12.5Hz, 6.25Hz) for speech-language alignment, while the original paper develops continuous σ-VAE-based acoustic and semantic tokenizers both operating at 7.5Hz for podcast generation. The architectural approaches and application domains differ fundamentally.

4. SyllableLM: Learning Coarse Semantic Units for Speech Language Models

URL: View paper

Brief Assessment

SyllableLM[40] focuses on discrete syllable-like units at 5Hz for speech language modeling, while the original paper develops continuous σ-VAE-based acoustic tokenizers at 7.5Hz for TTS synthesis. These serve fundamentally different purposes: semantic clustering vs. acoustic reconstruction fidelity.

5. U-Codec: Ultra Low Frame-rate Neural Speech Codec for Fast High-fidelity Speech Generation

URL: View paper

Prior Art Analysis

U-Codec[37] demonstrates prior work on ultra-low frame rate speech tokenizers operating at 5 Hz, which is even lower than the 7.5 Hz claimed in the original paper. The candidate paper explicitly addresses the same technical challenge of developing speech tokenizers at extremely low frame rates while preserving audio fidelity. Both papers employ similar architectural strategies including transformer-based modules for long-term dependency modeling and residual vector quantization (RVQ) for acoustic representation. The candidate's work at 5 Hz predates or is contemporaneous with the original's 7.5 Hz claim, directly challenging the novelty of operating at 'ultra-low' frame rates.

Evidence

Evidence 1 - **Rationale**: Both papers claim to develop 'ultra-low frame rate' speech tokenizers. U-Codec[37] operates at 5 Hz, which is lower than the original's 7.5 Hz, demonstrating that the concept of ultra-low frame rate tokenizers was already explored at even more aggressive compression rates. - **Original**: a core component of our approach is the continuous speech tokenizers operating at an ultra-low frame rate of 7.5. this tokenizer effectively preserves audio fidelity while significantly boosting computational efficiency for processing long sequences. - **Candidate**: we proposeu-codec, anultra low frame-rate neural speechcodecthat achieves high-fidelity reconstruction and fast speech generation at an extremely low framerate of 5hz (5 frames per second).

Evidence 2 - **Rationale**: Both papers address the same technical challenge of preserving audio fidelity under extreme compression. While the original uses sigma-VAE, U-Codec[37] uses transformer-based long-term dependency modeling with RVQ optimization, showing alternative prior approaches to the same problem. - **Original**: acoustic tokenizeradopts the principles of a variational autoencoder (v ae) (kingma & welling, 2014), specifically drawing inspiration from theσ-v ae variant proposed in latentlm (sun et al., under review as a conference paper at iclr 2026 2024) to mitigate potential variance collapse issues of v ae... - **Candidate**: to address loss of speech and spectrum details at extreme compression, we introduce a transformer-based inter-frame long-term dependency model that improves fidelity and carefully optimize residual vector quantization (rvq) layers and codebook sizes.

Evidence 3 - **Rationale**: Both papers recognize and address the challenge of maintaining fidelity at ultra-low frame rates. U-Codec[37] explicitly tackles intelligibility and spectral fidelity degradation through transformer-based long-term dependency modeling, demonstrating prior work on this specific technical challenge. - **Original**: the acoustic tokenizer aggressively compresses audio while preserving remarkable reconstruction fidelity, and semantic tokenizer extracts linguistic content. this decoupled design, with both components leveraging the same highly efficient frame rate, allows for optimized acoustic and semantic featur... - **Candidate**: long-term dependency module.at an ultra low frame-rate of 5hz, each token covers a long speech span, which may degrade intelligibility and spectral fidelity. to address this, we introduce a contextual transformer bottleneck directly after downsampling.

6. GLM-4-Voice: Towards Intelligent and Human-Like End-to-End Spoken Chatbot

URL: View paper

Prior Art Analysis

GLM-4-Voice[38] demonstrates prior work on ultra-low frame rate speech tokenization that predates the original paper's contribution. While the original paper claims novelty in developing specialized acoustic and semantic tokenizers both operating at 7.5 Hz, GLM-4-Voice[38] presents a speech tokenizer operating at an even lower frame rate of 12.5 Hz (compared to the original's 7.5 Hz claim). Both approaches achieve ultra-low frame rates for computational efficiency in processing long sequences. The candidate's tokenizer is derived from an ASR model with a vector-quantized bottleneck, demonstrating that ultra-low frame rate speech tokenization for efficient long-sequence processing was already established in prior work.

Evidence

Evidence 1 - **Rationale**: Both papers claim ultra-low frame rate speech tokenizers as a core contribution for computational efficiency. GLM-4-Voice[38] operates at 12.5 Hz, demonstrating that ultra-low frame rate tokenization was already achieved in prior work, refuting the novelty of the 7.5 Hz claim. - **Original**: a core component of our approach is the continuous speech tokenizers operating at an ultra-low frame rate of 7.5. this tokenizer effectively preserves audio fidelity while significantly boosting computational efficiency for processing long sequences. - **Candidate**: glm-4-voice uses an ultra-low bitrate (175bps), single-codebook speech tokenizer with 12.5hz frame rate derived from an automatic speech recognition (asr) model by incorporating a vector-quantized bottleneck into the encoder.

7. LongCat-Audio-Codec: An Audio Tokenizer and Detokenizer Solution Designed for Speech Large Language Models

URL: View paper

Brief Assessment

LongCat Audio Codec[39] operates at 16.67 Hz frame rate, which is more than 2x higher than the original paper's 7.5 Hz. The candidate uses discrete coding (0.43-0.87 kbps bitrate) rather than continuous σ-VAE representations, representing a fundamentally different technical approach to tokenization.

8. Tadicodec: Text-aware diffusion speech tokenizer for speech language modeling

URL: View paper

Prior Art Analysis

Tadicodec[35] demonstrates that prior work achieved ultra-low frame rate speech tokenization before the original paper. Tadicodec[35] operates at 6.25 Hz frame rate with a single-layer codebook, achieving a bitrate of 0.0875 kbps for 24 kHz speech. This is even lower than the 7.5 Hz claimed in the original paper. Both papers use continuous representations (sigma-VAE in the original, diffusion autoencoder in Tadicodec[35]) rather than discrete codebooks, and both emphasize the computational efficiency benefits for long-form speech processing. The original paper's claim to novelty in developing 'specialized acoustic and semantic tokenizers that both operate at an ultra-low frame rate of 7.5 Hz' is refuted by Tadicodec[35]'s earlier achievement of 6.25 Hz.

Evidence

Evidence 1 - **Rationale**: This pair demonstrates that Tadicodec[35] achieved an even lower frame rate (6.25 Hz vs 7.5 Hz) using continuous representations, directly refuting the novelty claim of ultra-low frame rate continuous tokenizers. - **Original**: a core component of our approach is the continuous speech tokenizers operating at an ultra-low frame rate of 7.5. this tokenizer effectively preserves audio fidelity while significantly boosting computational efficiency for processing long sequences. - **Candidate**: we introduce the text-aware diffusion transformer speech codec (tadicodec), a novel approach designed to overcome these challenges. tadicodec employs end-to-end optimization for quantization and reconstruction through a diffusion autoencoder, while integrating text guidance into the diffusion decode...

Evidence 2 - **Rationale**: This shows that the concept of decomposing speech into semantic and acoustic components at low frame rates was already established in prior work cited by Tadicodec[35], indicating this was not a novel contribution. - **Original**: the acoustic tokenizer aggressively compresses audio while preserving remarkable reconstruction fidelity, and semantic tokenizer extracts linguistic content. this decoupled design, with both components leveraging the same highly efficient frame rate, allows for optimized acoustic and semantic featur... - **Candidate**: recent studies [2, 3, 4, 10, 11, 25, 26] emphasize that effective speech tokens for language modeling should exhibit low frame rates and semantic richness, which criteria that directly shape the design of modern speech tokenizers. to achieve this, several works [10, 25, 26, 27] decompose speech int...

Evidence 3 - **Rationale**: This pair directly compares the frame rates, showing Tadicodec[35] achieved lower rates while maintaining quality metrics, refuting the claim of being first to achieve ultra-low frame rates. - **Original**: a cornerstone of vibevoiceis its efficient hybrid speech representation strategy, derived from specialized acoustic and semantic tokenizers, both operating at an ultra-low frame rate of 7.5 hz. - **Candidate**: tadicodec achieves an extremely low frame rate of 6.25 hz and a corresponding bitrate of 0.0875 kbps with a single-layer codebook for 24 khz speech, while maintaining superior performance on critical speech generation evaluation metrics such as word error rate (wer), speaker similarity (sim), and sp...

Evidence 4 - **Rationale**: Both papers use continuous latent representations (sigma-VAE vs diffusion autoencoder) at ultra-low frame rates, but Tadicodec[35] achieves lower rates, demonstrating prior art in this design space. - **Original**: acoustic tokenizeradopts the principles of a variational autoencoder (v ae) (kingma & welling, 2014), specifically drawing inspiration from theσ-v ae variant proposed in latentlm (sun et al., under review as a conference paper at iclr 2026 2024) to mitigate potential variance collapse issues of v ae... - **Candidate**: to address the limitations of current speech tokenizers, we propose the text-aware diffusion transformer speech codec (tadicodec), a novel model that achieves an exceptionally low frame rate of 6.25 hz using a single codebook, corresponding to a bitrate of 0.0875 kbps for 24 khz speech.

9. Lower Frame Rate Neural Network Acoustic Models.

URL: View paper

Brief Assessment

Lower Frame Rate[41] focuses on conventional neural network acoustic models for ASR with lower frame rates (30-40ms), not on speech tokenizers for generative TTS. The paper addresses frame-level acoustic modeling for recognition, whereas the original contribution concerns tokenization for synthesis with VAE-based compression.

10. Say More with Less: Variable-Frame-Rate Speech Tokenization via Adaptive Clustering and Implicit Duration Coding

URL: View paper

Brief Assessment

Variable Frame-Rate Tokenization[34] focuses on adaptive, variable-rate tokenization (adjusting tokens based on local feature similarity), whereas the original paper uses a fixed ultra-low frame rate of 7.5 Hz with separate acoustic and semantic tokenizers. These are fundamentally different approaches to token rate design.

Contribution 3: Annotation pipeline for podcast data with conversational dynamics

Description: The authors propose a novel automatic annotation pipeline tailored for extended multi-speaker speech data. This pipeline generates pseudo transcriptions and speaker turn-taking labels for large-scale podcast datasets, enabling the model to learn realistic intonation, turn-taking, and subtle expressive cues from authentic conversational material.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Automatic segmentation of multiparty dialogue

URL: View paper

Brief Assessment

Multiparty Dialogue Segmentation[32] focuses on automatic segmentation of meeting dialogues into topic boundaries, not on creating annotation pipelines for podcast transcription and turn-taking labels. The candidate addresses dialogue segmentation as an analysis task, whereas the original contribution is about data preparation infrastructure for training speech synthesis models.

2. DeepCon: An end-to-end multilingual toolkit for automatic minuting of multi-party dialogues

URL: View paper

Brief Assessment

DeepCon[27] focuses on meeting transcription and minuting using existing ASR services (Amazon Transcribe), not on developing novel annotation pipelines for podcast data with turn-taking labels and pseudo transcriptions for training conversational TTS models.

3. A multi-modal explainability approach for human-aware robots in multi-party conversation - Data

URL: View paper

Brief Assessment

Multi-Modal Explainability[28] focuses on addressee estimation in human-robot multi-party conversations, not on automatic annotation pipelines for podcast transcription and turn-taking.

4. An End-to-End Multilingual System for Automatic Minuting of Multi-Party Dialogues

URL: View paper

Brief Assessment

Multilingual Minuting[29] focuses on automatic minuting of multi-party dialogues using ASR and meeting summarization, not on creating annotation pipelines for podcast data with turn-taking labels and pseudo transcriptions for training TTS models.

5. Proactive Hearing Assistants that Isolate Egocentric Conversations

URL: View paper

Brief Assessment

Proactive Hearing Assistants[33] focuses on real-time egocentric conversation partner identification using binaural audio and turn-taking behavior, not on automatic annotation pipelines for podcast datasets with transcription and speaker labels.

6. Towards multi-party conversation modeling

URL: View paper

Brief Assessment

Multi-Party Conversation Modeling[24] focuses on collecting asynchronous multi-party conversations through a social media tool for everyday talk, not on automatic annotation pipelines for podcast audio with turn-taking labels and transcriptions.

7. Diarization-Aware Multi-Speaker Automatic Speech Recognition via Large Language Models

URL: View paper

Brief Assessment

Diarization-Aware ASR[26] focuses on multi-speaker ASR with diarization for meeting transcription, not on creating annotation pipelines for podcast training data with turn-taking labels and pseudo transcriptions.

8. Identifying introductions in podcast episodes from automatically generated transcripts

URL: View paper

Brief Assessment

Podcast Introductions[31] focuses on identifying introduction segments in podcast episodes using ASR transcripts, not on generating pseudo transcriptions and turn-taking labels for training conversational TTS models. The candidate's annotation pipeline labels introduction boundaries for segmentation tasks, while the original paper's pipeline generates training data for speech synthesis with conversational dynamics.

9. Spoken Language Processing: Conversational AI for Spontaneous Human Dialogues

URL: View paper

Brief Assessment

Conversational AI Spontaneous[25] provides insufficient detail about annotation pipelines for multi-speaker podcast data. The candidate's context only contains fragmentary mentions of transcript engineering and multi-speaker challenges without describing a complete automatic annotation pipeline for turn-taking and transcription generation.

10. NaturalVoices: A Large-Scale, Spontaneous and Emotional Podcast Dataset for Voice Conversion

URL: View paper

Brief Assessment

NaturalVoices[30] focuses on voice conversion tasks with emotion and speaker annotations, while the original paper targets podcast generation with turn-taking and transcription for TTS training.

Appendix: Text Similarity Detection

Textual similarity detection checked 28 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. MoonCast: High-quality zero-shot podcast generation

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] VibeVoice: Expressive Podcast Generation with Next-Token Diffusion View paper
- [1] Interactive Conversational Head Generation View paper
- [2] FireRedTTS-2: Towards Long Conversational Speech Generation for Podcast and Chatbot View paper
- [3] DialogueAgents: A Hybrid Agent-Based Speech Synthesis Framework for Multi-Party Dialogue View paper
- [4] Controllable Multi-Speaker Emotional Speech Synthesis With an Emotion Representation of High Generalization Capability View paper
- [5] SoulX-Podcast: Towards Realistic Long-form Podcasts with Dialectal and Paralinguistic Diversity View paper
- [6] Podagent: A comprehensive framework for podcast generation View paper
- [7] Boosting Multi-Speaker Expressive Speech Synthesis with Semi-Supervised Contrastive Learning View paper
- [8] Incorporating emotion for response generation in multi-turn dialogues View paper
- [9] Narrator or Character: Voice Modulation in an Expressive Multi-speaker TTS View paper
- [10] Spoken DialogSum: An Emotion-Rich Conversational Dataset for Spoken Dialogue Summarization View paper
- [11] Multi-Channel Conversational Speaker Separation via Neural Diarization View paper
- [12] Audiobook-CC: Controllable Long-context Speech Generation for Multicast Audiobook View paper
- [13] A long-duration Speech Semantic Recognition and Summarization Model for multi-speaker Conversations View paper
- [14] DialoSpeech: Dual-Speaker Dialogue Generation with LLM and Flow Matching View paper
- [15] Acquiring conversational speaking style from multi-speaker spontaneous dialog corpus for prosody-controllable sequence-to-sequence speech synthesis View paper
- [16] Cross-Speaker Style Transfer for TTS with Singing Voice Conversion Data Augmentation, Style Filtering, and F0 Matching View paper
- [17] Deep Generative Models for Personalized Speech and Spoken Dialog Modeling View paper
- [18] Multi-Speaker Expressive Speech Synthesis via Semi-supervised Contrastive Learning View paper
- [19] Synthesis of listener vocalizations View paper
- [20] Multi-Speaker Expressive Speech Synthesis via Multiple Factors Decoupling View paper
- [21] End-to-End Multi-speaker Speech Synthesis with Controllable Stress View paper
- [22] Speech Recognition and Multi-Speaker Diarization of Long Conversations View paper
- [23] Multi-Speaker Emotional Speech Synthesis with Fine-Grained Prosody Modeling View paper

- [24] Towards multi-party conversation modeling View paper
- [25] Spoken Language Processing: Conversational AI for Spontaneous Human Dialogues View paper
- [26] Diarization-Aware Multi-Speaker Automatic Speech Recognition via Large Language Models View paper
- [27] DeepCon: An end-to-end multilingual toolkit for automatic minuting of multi-party dialogues View paper
- [28] A multi-modal explainability approach for human-aware robots in multi-party conversation - Data View paper
- [29] An End-to-End Multilingual System for Automatic Minuting of Multi-Party Dialogues View paper
- [30] NaturalVoices: A Large-Scale, Spontaneous and Emotional Podcast Dataset for Voice Conversion View paper
- [31] Identifying introductions in podcast episodes from automatically generated transcripts View paper
- [32] Automatic segmentation of multiparty dialogue View paper
- [33] Proactive Hearing Assistants that Isolate Egocentric Conversations View paper
- [34] Say More with Less: Variable-Frame-Rate Speech Tokenization via Adaptive Clustering and Implicit Duration Coding View paper
- [35] Tadicodec: Text-aware diffusion speech tokenizer for speech language modeling View paper
- [36] Kimi-audio technical report View paper
- [37] U-Codec: Ultra Low Frame-rate Neural Speech Codec for Fast High-fidelity Speech Generation View paper
- [38] GLM-4-Voice: Towards Intelligent and Human-Like End-to-End Spoken Chatbot View paper
- [39] LongCat-Audio-Codec: An Audio Tokenizer and Detokenizer Solution Designed for Speech Large Language Models View paper
- [40] SyllableLM: Learning Coarse Semantic Units for Speech Language Models View paper
- [41] Lower Frame Rate Neural Network Acoustic Models. View paper
- [42] TASLA: Text-Aligned Speech Tokens with Multiple Layer-Aggregation View paper
- [43] LM-SPT: LM-Aligned Semantic Distillation for Speech Tokenization View paper
- [44] CoVoMix: Advancing zero-shot speech generation for human-like multi-talker conversations View paper
- [45] CoVoMix2: Advancing Zero-Shot Dialogue Generation with Fully Non-Autoregressive Flow Matching View paper
- [46] MoonCast: High-quality zero-shot podcast generation View paper
- [47] SLAM-Omni: Timbre-Controllable Voice Interaction System with Single-Stage Training View paper
- [48] Character-Driven Narrative Generation for Scene-Based Video Synthesis View paper