Novelty Assessment Report

Paper: StableToken: A Noise-Robust Semantic Speech Tokenizer for Resilient SpeechLLMs

PDF URL: https://openreview.net/pdf?id=17DNmdQ9aU

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

Prevalent semantic speech tokenizers, designed to capture linguistic content, are surprisingly fragile. We find they are not robust to meaning-irrelevant acoustic perturbations; even at high Signal-to-Noise Ratios (SNRs) where speech is perfectly intelligible, their output token sequences can change drastically, increasing the learning burden for downstream LLMs. This instability stems from two flaws: a brittle single-path quantization architecture and a distant training signal indifferent to intermediate token stability. To address this, we introduce StableToken, a tokenizer that achieves stability through a consensus-driven mechanism. Its multi-branch architecture processes audio in parallel, and these representations are merged via a powerful bit-wise voting mechanism to form a single, stable token sequence. StableToken sets a new state-of-the-art in token stability, drastically reducing Unit Edit Distance (UED) under diverse noise conditions. This foundational stability translates directly to downstream benefits, significantly improving the robustness of SpeechLLMs on a variety of tasks.

Core Task Landscape

This paper addresses: **Noise-Robust Semantic Speech Tokenization for Speech Language Models**

A total of **36 papers** were analyzed and organized into a taxonomy with **17 categories**.

Taxonomy Overview

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

- **Semantic Speech Tokenization Architectures**
- **Noise Robustness Enhancement Techniques**
- **Generative Speech Enhancement with Semantic Modeling**
- **Evaluation Methodologies and Benchmarks**
- **Downstream Application Domains**
- **Robustness to ASR Errors and Transcription Noise**
- **Domain-Specific Applications**

Complete Taxonomy Tree

- Noise-Robust Semantic Speech Tokenization for Speech Language Models Survey Taxonomy
- Semantic Speech Tokenization Architectures
  - Consensus-Driven Multi-Branch Tokenization ★ (2 papers)
  - [0] StableToken: A Noise-Robust Semantic Speech Tokenizer for Resilient SpeechLLMs (Anon et al., 2026) View paper
  - [33] FUSE: Universal Speech Enhancement using Multi-Stage Fusion of Sparse Compression and Token Generation Models for the URGENT 2025 Challenge (Goswami, 2025) View paper
  - Single-Path Quantization Approaches (4 papers)
  - [3] Efficient extraction of noise-robust discrete units from self-supervised speech models (Jakob Poncelet, 2024) View paper
  - [6] DC-Spin: A speaker-invariant speech tokenizer for spoken language models (Heng-Jui Chang, 2024) View paper
  - [10] NAST: Noise Aware Speech Tokenization for Speech Language Models (Shoval Messica, 2024) View paper
  - [13] SyllableLM: Learning Coarse Semantic Units for Speech Language Models (Baade, 2024) View paper
  - Dual-Stream Semantic-Acoustic Decomposition (2 papers)
  - [14] SAC: Neural Speech Codec with Semantic-Acoustic Dual-Stream Quantization (Chen Wenxi, 2025) View paper
  - [21] Explainable Disentanglement on Discrete Speech Representations for Noise-Robust ASR (Anshul, 2025) View paper
  - Continuous Representation Learning (3 papers)
  - [11] Continuous Speech Tokens Makes LLMs Robust Multi-Modality Learners (Yuan Ze, 2024) View paper
  - [31] WavJEPA: Semantic learning unlocks robust audio foundation models for raw waveforms (Guetschel, 2025) View paper
  - [32] JEPA as a Neural Tokenizer: Learning Robust Speech Representations with Density Adaptive Attention (Georgios Ioannides, 2025) View paper
- Noise Robustness Enhancement Techniques
  - Denoising via Acoustic Token Prediction (2 papers)
  - [1] Speechx: Neural codec language model as a versatile speech transformer (Xiaofei Wang, 2024) View paper
  - [5] Improving Noise Robustness of LLM-based Zero-shot TTS via Discrete Acoustic Token Denoising (LU Ye-xin, 2025) View paper
  - Codebook-Based Clean Speech Restoration (1 papers)
  - [35] Wav2code: Restore Clean Speech Representations via Codebook Lookup for Noise-Robust ASR (Yu-Chen Hu, 2023) View paper
  - Self-Supervised Noise-Robust Pre-Training (3 papers)
  - [15] A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition (Qiu-Shi Zhu, 2022) View paper
  - [16] Optimizing speech representation learning for enhanced noise robustness in downstream applications (Ng, 2025) View paper

- [25] Robust Data2VEC: Noise-Robust Speech Representation Learning for ASR by Combining Regression and Improved Contrastive Learning (Qiu-shi Zhu, 2022) View paper
  - Parameter-Efficient Adaptation for Robustness (1 papers)
  - [24] Thinking Fast and Slow: Robust Speech Recognition via Deep Filter-Tuning (Dianwen Ng, 2025) View paper
- Generative Speech Enhancement with Semantic Modeling
  - Language Model-Guided Enhancement (2 papers)
  - [4] Gense: Generative speech enhancement via language models using hierarchical modeling (Yao, 2025) View paper
  - [27] Tokenized Generative Speech Enhancement With Language Model and Flow Matching (Da-Hee Yang, 2025) View paper
  - Flow and Diffusion-Based Enhancement (2 papers)
  - [9] Simplespeech 2: Towards simple and efficient text-to-speech with flow-based scalar latent transformer diffusion models (Dongchao Yang, 2025) View paper
  - [30] SenSE: Semantic-Aware High-Fidelity Universal Speech Enhancement (Li Xingchen, 2025) View paper
- Evaluation Methodologies and Benchmarks (2 papers)
  - [2] Probing the Robustness Properties of Neural Speech Codecs (Tseng, 2025) View paper
  - [8] CodecBench: A Comprehensive Benchmark for Acoustic and Semantic Evaluation (Gao Qing-hui, 2025) View paper
- Downstream Application Domains
  - Automatic Speech Recognition (4 papers)
  - [7] Large Language Models are Efficient Learners of Noise-Robust Speech Recognition (Hu, 2024) View paper
  - [12] CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions (Mario Zusag, 2024) View paper
  - [28] Noise Robust Speech Recognition for Search and Rescue Domain (Sagar, 2023) View paper
  - [34] QV finder: an accurate Quran verse finder system (Bashar Al-Rfooh, 2025) View paper
  - Speech Translation and Cross-Modal Tasks (2 papers)
  - [19] SBVQA 2.0: Robust End-to-End Speech-Based Visual Question Answering for Open-Ended Questions (Faris Alasmary, 2023) View paper
  - [29] Robust Semantic Communications for Speech Transmission (Zhenzi Weng, 2024) View paper
  - Speech Emotion Recognition (1 papers)
  - [18] Describe where you are: Improving noise-robustness for speech emotion recognition with text description of the environment (Seong-Gyun Leem, 2024) View paper
  - Semantic Communication Systems (3 papers)
  - [17] SNR-Adaptive Multi-Layer Semantic Communication for Speech (Jiejie Guo, 2023) View paper
  - [20] Large Model Empowered Multi-Modal Semantic Communication With Selective Tokens for Training (Jincheng Peng, 2025) View paper
  - [36] KGRAG-SC: Knowledge Graph RAG-Assisted Semantic Communication (Fan Da-yu, 2025) View paper
- Robustness to ASR Errors and Transcription Noise (2 papers)
  - [22] Hearing the Meaning, Not the Mess: Beyond Literal Transcription for Spoken Language (Min Sun, 2025) View paper
  - [26] Telephonetic: Making Neural Language Models Robust to ASR and Semantic Noise (Larson, 2019) View paper
- Domain-Specific Applications (1 papers)
  - [23] Natural Language Processing for Real-Time Transcription in VoIP Systems (Prajwal Hegde N, 2025) View paper

Narrative

Core task: noise-robust semantic speech tokenization for speech language models. The field addresses how to convert continuous speech into discrete tokens that preserve semantic content while remaining stable under acoustic degradation. The taxonomy reveals several complementary directions: Semantic Speech Tokenization Architectures explores foundational designs for extracting meaningful units from audio, including consensus-driven multi-branch approaches like StableToken[0] and FUSE[33] that aggregate information from multiple pathways to improve stability. Noise Robustness Enhancement Techniques focuses on methods that explicitly harden tokenizers against environmental corruption, such as Noise-Robust Discrete Units[3] and domain-specific pretraining strategies like Robust Data2VEC[25]. Generative Speech Enhancement with Semantic Modeling investigates how semantic representations can guide enhancement or denoising, exemplified by Gense[4] and Acoustic Token Denoising[5]. Evaluation Methodologies and Benchmarks provide standardized testbeds like CodecBench[8] to measure robustness, while Downstream Application Domains and Domain-Specific Applications examine how these tokenizers perform in real-world settings such as telephony, search-and-rescue, or VoIP transcription.

A central tension in the field is whether to build robustness into the tokenization architecture itself or to rely on separate enhancement modules. Works like Speechx[1] and SimpleSpeech[9] pursue end-to-end designs that jointly model acoustic and semantic features, whereas others such as DC-Spin[6] and SAC[14] emphasize disentangling content from noise at the representation level. StableToken[0] sits within the consensus-driven multi-branch cluster, sharing conceptual ground with FUSE[33] by leveraging multiple encoding pathways to stabilize semantic tokens under noise. Compared to single-pathway methods like Noise-Robust Discrete Units[3], which directly train on noisy data, StableToken[0] emphasizes architectural redundancy to achieve robustness. This positions it as a middle ground between purely data-driven hardening and explicit enhancement preprocessing, offering a design philosophy that balances architectural complexity with generalization across diverse noise conditions.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. FUSE: Universal Speech Enhancement using Multi-Stage Fusion of Sparse Compression and Token Generation Models for the URGENT 2025 Challenge

Authors: Goswami, Nabarun, Harada Tatsuya | **Year/Venue**: 2025 • Interspeech | **URL**: View paper

Abstract

We propose a multi-stage framework for universal speech enhancement, designed for the Interspeech 2025 URGENT Challenge. Our system first employs a Sparse Compression Network to robustly separate sources and extract an initial clean speech estimate from noisy inputs. This is followed by an efficient generative model that refines speech quality by leveraging self-supervised features and optimizing a masked language modeling objective on acoustic tokens derived from a neural audio codec. In the fi...

Relationship Analysis

Both papers belong to the Consensus-Driven Multi-Branch Tokenization category, employing parallel processing architectures to improve robustness in speech processing. However, they address fundamentally different problems: the original paper (StableToken) focuses on creating noise-robust semantic speech tokens for SpeechLLMs through a multi-branch voting mechanism that stabilizes discrete token sequences, while the candidate paper (FUSE) tackles universal speech enhancement by fusing outputs from multiple processing stages (sparse compression, generative refinement, and fusion networks) to improve signal quality rather than tokenization stability. The key distinction is that StableToken produces stable discrete representations for downstream LLM consumption, whereas FUSE generates enhanced continuous speech signals for improved audio quality.

Contributions Analysis

Overall novelty summary. The paper introduces StableToken, a consensus-driven multi-branch tokenizer designed to stabilize semantic speech tokens under acoustic noise. It resides in the 'Consensus-Driven Multi-Branch Tokenization' leaf, which contains only two papers: StableToken itself and FUSE. This is a notably sparse research direction within the broader taxonomy of 36 papers across 17 leaf nodes, suggesting that multi-branch consensus mechanisms for tokenization remain relatively underexplored compared to single-path quantization approaches, which contain four papers.

The taxonomy reveals that StableToken sits at the intersection of architectural innovation and noise robustness. Neighboring leaves include 'Single-Path Quantization Approaches' with methods like Noise-Robust Discrete Units and 'Dual-Stream Semantic-Acoustic Decomposition' exemplified by Speechx. The 'Noise Robustness Enhancement Techniques' branch explores complementary strategies such as denoising via acoustic token prediction and self-supervised pre-training. StableToken diverges from these by embedding robustness directly into the tokenizer architecture through parallel branches and bit-wise voting, rather than relying on separate enhancement modules or data augmentation alone.

Among the three contributions analyzed, the literature search examined 13 candidates total. The voting-LFQ module was compared against 3 candidates with no refutations found, while the consensus-driven training strategy was assessed against 10 candidates, also yielding no clear prior work overlap. The core StableToken system itself was not directly compared to candidates in the refutation analysis. This limited search scope—focused on top-K semantic matches—suggests that while no immediate overlaps were detected, the analysis does not claim exhaustive coverage of all possible prior work in multi-branch tokenization or voting mechanisms.

Based on the 13 candidates examined, StableToken appears to occupy a relatively novel position within the sparse consensus-driven tokenization space. However, the limited search scope and the presence of only one sibling paper (FUSE) in the taxonomy leaf indicate that a broader literature review—particularly in adjacent fields like ensemble methods or multi-view learning—would be necessary to fully assess the originality of the bit-wise voting mechanism and multi-branch design.

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

Contribution 1: StableToken: a noise-robust semantic speech tokenizer

Description: The authors propose StableToken, a new semantic speech tokenizer designed to be robust against acoustic noise. It uses a multi-branch architecture with a bit-wise voting mechanism to produce stable token sequences even under noisy conditions, addressing the fragility of existing semantic tokenizers.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 2: voting-LFQ module with bit-level majority voting

Description: The authors introduce a novel multi-branch quantization architecture that extends the LFQ algorithm with a differentiable bit-level majority voting mechanism. This provides fine-grained error correction at the bit level rather than coarse token level, enabling robust representation learning with negligible inference overhead.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Multi-Branch Integrated Model for Respiratory Disease Screening Using Cough Sounds

URL: View paper

Brief Assessment

Respiratory Cough Screening[38] applies voting to multi-branch cough sound classification for respiratory disease screening, not to quantization or representation learning in speech tokenizers. The voting operates on final classification results rather than bit-level quantization codes.

---

2. GAP-CoT: A Multi-role Multi-path Game-theoretic Chain-of-Thought Reasoning Framework for Industrial Intelligent Decision-making

URL: View paper

Brief Assessment

GAP-CoT[39] focuses on game-theoretic chain-of-thought reasoning for industrial decision-making with voting mechanisms at the answer level, not multi-branch quantization with bit-level voting for speech representation learning.

---

3. Quantization Aware Matryoshka Adaptation: Leveraging Matryoshka Learning, Quantization, and Bitwise Operations for Reduced Storage and Improved Retrieval Speed

URL: View paper

Brief Assessment

Quantization Matryoshka[37] focuses on embedding quantization for information retrieval using bitwise operations (XOR, NOT, popcount) for efficient retrieval, not on multi-branch quantization with bit-level majority voting for error correction in speech tokenization.

Contribution 3: consensus-driven training strategy with multi-view inputs

Description: The authors develop a training strategy that presents clean audio to most branches and perturbed audio to a minority, creating a stable reference. A consensus loss is applied to enforce agreement across branches, addressing the problem of distant supervisory signals in tokenizer training.

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. Imagined Speech Decoding by Learning Consensus Graph From RKHS-Based Multi-View EEG Features

URL: View paper

Brief Assessment

Imagined Speech Decoding[48] focuses on EEG-based brain-computer interfaces using multi-view feature learning in kernel space for speech imagery decoding, not audio tokenization or speech LLM training. The domains, objectives, and technical approaches are fundamentally different.

---

2. ViewSRD: 3D Visual Grounding via Structured Multi-View Decomposition

URL: View paper

Brief Assessment

ViewSRD[46] addresses 3D visual grounding using multi-view scene decomposition for spatial understanding, not audio tokenization robustness. The 'multi-view' concept refers to different 3D spatial perspectives of scenes, not multiple audio processing branches for consensus-based noise robustness.

3. Facial Video-based Remote Physiological Measurement via Self-supervised Learning

URL: View paper

Brief Assessment

Facial Remote Physiological[49] focuses on self-supervised learning for remote physiological measurement from facial videos, not speech tokenization. The multi-view approach in [49] involves spatial augmentation of video frames for contrastive learning, which is fundamentally different from the audio-based multi-branch consensus mechanism for speech token stability described in the original paper.

4. Efficient Generalization via Multimodal Co-Training under Data Scarcity and Distribution Shift

URL: View paper

Brief Assessment

Multimodal Co-Training[45] focuses on multimodal learning with different modalities (e.g., text and images) using co-training and agreement loss between views. The original paper addresses noise robustness in speech tokenization through multi-branch quantization with consensus mechanisms. These are fundamentally different problem domains and technical approaches.

5. Odyssey: reconstructing evolution through emergent consensus in the global proteome

URL: View paper

Brief Assessment

Odyssey[44] applies consensus mechanisms to protein structure modeling with finite scalar quantization and iterative propagation between residues, not to speech tokenization with multi-branch audio processing and noise perturbation training.

6. Mix before align: Towards zero-shot cross-lingual sentiment analysis via soft-mix and multi-view learning

URL: View paper

Brief Assessment

Mix Before Align[40] focuses on cross-lingual sentiment analysis using multi-view learning between original and code-switched sentences, not on robust speech tokenization with perturbed audio inputs and consensus loss for token stability.

7. Perla: Perceptive 3d language assistant

URL: View paper

Brief Assessment

Perla[43] focuses on 3D scene understanding with point clouds using local-global representation aggregation and a consensus loss for spatial connectivity. The original paper addresses speech tokenization with multi-branch audio processing and consensus across perturbed/clean audio branches—fundamentally different domains and mechanisms.

8. Multi-view document clustering with joint contrastive learning

URL: View paper

Brief Assessment

Multi-view Document Clustering[42] focuses on document clustering with multi-view representations for capturing cluster-wise information, not speech tokenization with consensus mechanisms for noise robustness. The technical domains and objectives are fundamentally different.

9. Deterministic Reversible Data Augmentation for Neural Machine Translation

URL: View paper

Brief Assessment

Deterministic Reversible Augmentation[47] focuses on machine translation with multi-granularity subword segmentations and multi-view learning for translation consistency, not on robust speech tokenization with consensus mechanisms for handling acoustic perturbations in speech LLMs.

10. MCL-NER: cross-lingual named entity recognition via multi-view contrastive learning

URL: View paper

Brief Assessment

MCL-NER[41] addresses cross-lingual NER using multi-view contrastive learning with code-switched data for semantic alignment, not audio tokenization with consensus mechanisms for noise robustness. The domains (NLP entity recognition vs. speech tokenization) and technical approaches differ fundamentally.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] StableToken: A Noise-Robust Semantic Speech Tokenizer for Resilient SpeechLLMs View paper
- [1] Speechx: Neural codec language model as a versatile speech transformer View paper
- [2] Probing the Robustness Properties of Neural Speech Codecs View paper
- [3] Efficient extraction of noise-robust discrete units from self-supervised speech models View paper
- [4] Gense: Generative speech enhancement via language models using hierarchical modeling View paper
- [5] Improving Noise Robustness of LLM-based Zero-shot TTS via Discrete Acoustic Token Denoising View paper
- [6] DC-Spin: A speaker-invariant speech tokenizer for spoken language models View paper
- [7] Large Language Models are Efficient Learners of Noise-Robust Speech Recognition View paper
- [8] CodecBench: A Comprehensive Benchmark for Acoustic and Semantic Evaluation View paper
- [9] Simplespeech 2: Towards simple and efficient text-to-speech with flow-based scalar latent transformer diffusion models View paper
- [10] NAST: Noise Aware Speech Tokenization for Speech Language Models View paper
- [11] Continuous Speech Tokens Makes LLMs Robust Multi-Modality Learners View paper

- [12] CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions View paper
- [13] SyllableLM: Learning Coarse Semantic Units for Speech Language Models View paper
- [14] SAC: Neural Speech Codec with Semantic-Acoustic Dual-Stream Quantization View paper
- [15] A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition View paper
- [16] Optimizing speech representation learning for enhanced noise robustness in downstream applications View paper
- [17] SNR-Adaptive Multi-Layer Semantic Communication for Speech View paper
- [18] Describe where you are: Improving noise-robustness for speech emotion recognition with text description of the environment View paper
- [19] SBVQA 2.0: Robust End-to-End Speech-Based Visual Question Answering for Open-Ended Questions View paper
- [20] Large Model Empowered Multi-Modal Semantic Communication With Selective Tokens for Training View paper
- [21] Explainable Disentanglement on Discrete Speech Representations for Noise-Robust ASR View paper
- [22] Hearing the Meaning, Not the Mess: Beyond Literal Transcription for Spoken Language View paper
- [23] Natural Language Processing for Real-Time Transcription in VoIP Systems View paper
- [24] Thinking Fast and Slow: Robust Speech Recognition via Deep Filter-Tuning View paper
- [25] Robust Data2VEC: Noise-Robust Speech Representation Learning for ASR by Combining Regression and Improved Contrastive Learning View paper
- [26] Telephonetic: Making Neural Language Models Robust to ASR and Semantic Noise View paper
- [27] Tokenized Generative Speech Enhancement With Language Model and Flow Matching View paper
- [28] Noise Robust Speech Recognition for Search and Rescue Domain View paper
- [29] Robust Semantic Communications for Speech Transmission View paper
- [30] SenSE: Semantic-Aware High-Fidelity Universal Speech Enhancement View paper
- [31] WavJEPA: Semantic learning unlocks robust audio foundation models for raw waveforms View paper
- [32] JEPA as a Neural Tokenizer: Learning Robust Speech Representations with Density Adaptive Attention View paper
- [33] FUSE: Universal Speech Enhancement using Multi-Stage Fusion of Sparse Compression and Token Generation Models for the URGENT 2025 Challenge View paper
- [34] QV finder: an accurate Quran verse finder system View paper
- [35] Wav2code: Restore Clean Speech Representations via Codebook Lookup for Noise-Robust ASR View paper
- [36] KGRAG-SC: Knowledge Graph RAG-Assisted Semantic Communication View paper
- [37] Quantization Aware Matryoshka Adaptation: Leveraging Matryoshka Learning, Quantization, and Bitwise Operations for Reduced Storage and Improved Retrieval Speed View paper
- [38] Multi-Branch Integrated Model for Respiratory Disease Screening Using Cough Sounds View paper
- [39] GAP-CoT: A Multi-role Multi-path Game-theoretic Chain-of-Thought Reasoning Framework for Industrial Intelligent Decision-making View paper
- [40] Mix before align: Towards zero-shot cross-lingual sentiment analysis via soft-mix and multi-view learning View paper
- [41] MCL-NER: cross-lingual named entity recognition via multi-view contrastive learning View paper
- [42] Multi-view document clustering with joint contrastive learning View paper
- [43] Perla: Perceptive 3d language assistant View paper
- [44] Odyssey: reconstructing evolution through emergent consensus in the global proteome View paper
- [45] Efficient Generalization via Multimodal Co-Training under Data Scarcity and Distribution Shift View paper
- [46] ViewSRD: 3D Visual Grounding via Structured Multi-View Decomposition View paper
- [47] Deterministic Reversible Data Augmentation for Neural Machine Translation View paper
- [48] Imagined Speech Decoding by Learning Consensus Graph From RKHS-Based Multi-View EEG Features View paper
- [49] Facial Video-based Remote Physiological Measurement via Self-supervised Learning View paper