Discrete vs. Continuous Representations in Sample-Efficient Speech Enhancement

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the sample efficiency of discrete token-based speech enhancement models compare to continuous representation approaches when trained on varying amounts of noisy speech data. 9 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Discrete Audio Representation as an Alternative to Mel-Spectrograms for Speaker and Speech Recognition. Research question: How does the sample efficiency of discrete token-based speech enhancement models compare to continuous representation approaches when trained on varying amounts of noisy speech data?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

13 papers retrieved. 9 claims extracted; 0 independently verified. Quality review score: 4.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The EnCodec-trained model achieved a 12% absolute Equal Error Rate (EER) improvement over the Mel-Spectrogram model on t	×	0.10
On the VoxCeleb-clean wide-band trial files, the Mel-Spectrogram model achieved an EER of 2.2%, while the EnCodec-32 mod	×	0.08
When trained on additional narrowband data from the NIST-SRE train set, the Mel-Spectrogram model achieved an EER of 21.	×	0.05
In speaker diarization experiments using embeddings trained on wideband VoxCeleb 1&2, the EnCodec-32 model achieved an a	×	0.06
On the AMI Lapel dataset, the Mel-Spectrogram model achieved a DER of 3.11, outperforming the EnCodec-32 model which ach	×	0.05
Spectral analysis of EnCodec compressed audio reveals an Estimated Transfer Function that attenuates frequencies beyond	×	0.05
Increasing the number of EnCodec codebooks from 1 (0.75 Kbps) to 32 (24 Kbps) reduces the Word Error Rate on LS-Dev Clea	×	0.02
On the CallHome Test set, the Mel-Spectrogram model achieved a score of 43.30, while the EnCodec model achieved a score	×	0.05
On the VoxPopuli Test set, the EnCodec model achieved a score of 16.97, outperforming the Mel-Spectrogram model which sc	×	0.05

References

- http://arxiv.org/abs/2304.09116v3

- http://arxiv.org/abs/2502.20040v2
- http://arxiv.org/abs/2309.10922v1