

An ASR Framework for Robust Recognition in Low-Bitrate Environments

Wootack Lim, Sooyoung Park, Inseon Jang

Electronics and Telecommunications Research Institute (ETRI), Korea

wtlim@etri.re.kr, sooyoung@etri.re.kr, jinsn@etri.re.kr

Abstract— In this paper, we propose an automatic speech recognition (ASR) framework for robust recognition in low-bitrate environments, developed within the context of machine listening. The framework integrates a pre-trained voice activity detector for silence removal, a generalized neural codec for efficient low-bitrate compression, and a neural post-enhancement module to refine the decoded signal prior to recognition by the ASR model. Unlike feature-based transmission systems tailored to specific tasks, this approach leverages generalized codecs to ensure broad applicability and compatibility with existing systems. The post-enhancement model, trained on paired clean and decoded speech, focuses on reducing spectral distortion and suppressing coding noise without relying on perceptual losses. Experiments on the LibriSpeech corpus demonstrate that the proposed framework achieves an average bitrate reduction of 12.53% while maintaining or improving recognition accuracy compared to baseline codec performance, achieving word error rates of 7.18% with Lyra and 6.65% with DAC. These results confirm the effectiveness of the proposed framework for machine-oriented speech communication under low-bitrate conditions.

Keywords— Automatic speech recognition, low-bitrate speech coding, voice activity detection, neural codec, speech enhancement

I. INTRODUCTION

Machine listening is a subfield of artificial intelligence that leverages signal processing and machine learning to enable computational systems to perceive, interpret, and understand sounds in a manner similar to human perception [1]. Recent advances in machine learning have dramatically improved the robustness and generalization of machine listening systems. However, transmitting or storing high-quality audio signals for such models remains a bandwidth-intensive process. For example, uncompressed 16 kHz, 16-bit audio requires approximately 256 kbps, which is impractical for large-scale data or real-time audio applications in bandwidth-constrained environments. Reducing the bitrate while maintaining machine listening accuracy remains a crucial challenge, particularly for emerging machine listening applications such as edge devices, distributed sensing platforms, and large-scale audio understanding systems.

Traditional audio codecs, such as Advanced Audio Coding (AAC) [2], Unified Speech and Audio Coding (USAC) [3], Opus [4] or Enhanced Voice Services (EVS) [5], have been

developed primarily to preserve perceptual quality for human listeners. These codecs perform efficiently at bitrates typically ranging from approximately 6 to 24 kbps, while their optimization objectives remain focused on perceptual quality rather than machine listening. From the perspective of Audio Coding for Machines (ACoM), the focus shifts from subjective auditory quality to preserving the information most relevant for automatic analysis and recognition [6]. Unlike task-specific or feature-transmission approaches that encode only features for a particular model [7], ACoM requires the use of generalized audio codecs that can serve multiple downstream machine tasks such as Automatic Speech Recognition (ASR), speaker recognition, anomalous sound detection, and audio classification without task-dependent retraining or feature adaptation. This task-agnostic design not only simplifies system integration but also ensures compatibility with conventional signal processing and transmission frameworks.

Recent progress in neural audio coding has made it feasible to achieve extremely low bitrates while retaining intelligibility. Neural codecs such as Google's Lyra v2 [8], [9] and Descript Audio Codec (DAC) [10] leverage learned representations and vector quantization to efficiently compress speech signals. Despite these advancements, speech reconstructed at very low bitrates often contains temporal and spectral distortions that can degrade recognition performance, especially when non-speech regions are transmitted unnecessarily.

To address these limitations, this study proposes a low-bitrate ASR framework that integrates voice activity detection (VAD) and neural enhancement within a generalized neural coding pipeline. The proposed system employs a pre-trained VAD to exclude silent segments, thereby improving bitrate efficiency, while neural codecs such as Lyra v2 and DAC are used to compress speech segments into compact representations. A Transformer-based post-enhancement model further refines the decoded signals, recovering acoustic details essential for recognition. Experimental evaluation demonstrates that the proposed framework achieves substantial bitrate reduction without compromising ASR robustness, highlighting its potential as a unified machine-oriented architecture for efficient speech transmission in low-bitrate environments. Although this study focuses on ASR, the proposed framework can be extended to a task-agnostic machine listening system.

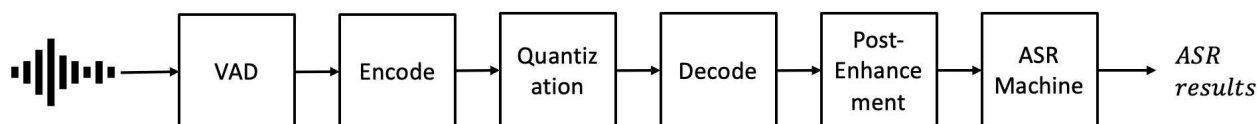


Figure 1. Overall architecture of the proposed ASR framework

II. PROPOSED METHOD

The overall architecture of the proposed framework is illustrated in Figure 1. It is composed of four stages: VAD, neural speech coding, post-enhancement, and the ASR process. The framework aims to reduce the bitrate effectively while maintaining the speech intelligibility and linguistic fidelity necessary for accurate speech recognition. Each component is designed to operate in a unified pipeline, allowing the system to efficiently process speech for machine understanding tasks.

The VAD module identifies active speech regions and removes non-speech segments from the input audio. We employ the pre-trained Silero-VAD [11], a lightweight convolutional neural network that operates on 16 kHz mono waveforms and outputs frame-level probabilities of speech activity. The model analyzes short overlapping frames and produces confidence scores in the range of 0.0 to 1.0. Frames exceeding a threshold are classified as speech and retained for subsequent encoding, while non-speech frames are discarded. This selective transmission reduces the average bitrate by omitting silent intervals, thereby improving efficiency in low-bitrate environments. While the removal of low-energy or transitional regions may slightly affect boundary continuity, the overall design minimizes such distortion and maintains the recognition-relevant content required for stable ASR performance. Through this process, the integration of Silero-VAD achieves substantial bitrate savings while minimizing ASR performance degradation compared to using the original audio.

The detected speech frames are then encoded using a neural audio codec capable of low-bitrate compression. In this work, we adopt two state-of-the-art neural codecs, Lyra and DAC, to validate the effectiveness of the proposed framework. Lyra is based on the SoundStream architecture and employs a convolutional encoder with residual vector quantization to represent speech in a compact form. DAC is a universal neural codec that utilizes a multi-stage residual vector-quantized variational autoencoder (RVQ-VAE) for high-fidelity reconstruction across diverse audio types. Both codecs convert the waveform into discrete latent representations that can be efficiently transmitted or stored while preserving key speech characteristics. On the receiver side, the decoder reconstructs the time-domain waveform from the quantized latent codes. Although these neural codecs achieve highly efficient compression at low bitrates, the reconstructed signals may still exhibit spectral distortion or loss of fine-grained temporal detail, which can affect downstream ASR performance.

To mitigate the distortions introduced by low-bitrate coding, a neural post-enhancement network is applied to the decoded

signal before ASR processing. We employ the MP-SENet architecture [12], which enhances both the magnitude and phase spectra in the time–frequency domain. The model consists of an encoder–decoder Transformer network with parallel branches that estimate a magnitude mask and a wrapped phase correction. By refining both spectral components, the network reconstructs a cleaner and more consistent waveform that better aligns with the distribution expected by the ASR model. The enhancement network is trained on paired data of clean and decoded speech, using a combination of multi-resolution spectrogram losses and waveform-level reconstruction losses to guide optimization. Unlike conventional perceptual enhancement, which focuses on human listening quality, our objective is to maximize machine listening robustness by minimizing distortions that hinder recognition accuracy.

The enhanced speech is finally passed to the ASR backend for transcription. In this study, we utilize the Whisper-base model, an open-source Transformer-based ASR system developed by OpenAI [13]. Whisper is trained on 680,000 hours of multilingual and multitask supervised data collected from the web, enabling it to generalize across a wide variety of acoustic conditions, languages, and noise environments. The model follows an encoder–decoder Transformer architecture, where the encoder processes log-Mel spectrograms of the input audio and the decoder autoregressively predicts text tokens. Unlike conventional ASR systems that rely on domain-specific training or noise adaptation, Whisper learns a universal mapping from speech to text, making it particularly robust to distortions, compression artifacts, and domain mismatches. Among the available model sizes, the Whisper-base version provides a good balance between accuracy and computational efficiency, allowing reliable evaluation of ASR performance under low-bitrate conditions. By combining VAD-driven bitrate reduction, neural speech coding, and Transformer-based post-enhancement, the proposed framework forms a generalized machine-listening pipeline that transmits speech efficiently while maintaining recognition robustness in bandwidth-limited scenarios.

III. EXPERIMENTS

For the experiment, we employed the LibriSpeech corpus [14] to evaluate the proposed low-bitrate ASR framework. The training data for the enhancement network consisted of the train-clean-100 dataset, while the evaluation used the test-clean dataset. To ensure the evaluation better reflects real-world conditions, we selected only utterances longer than 5 s from test-clean dataset, yielding 1,531 samples in total.

TABLE 1. AN EVALUATION OF ASR PERFORMANCE AND BITRATE ON LIBRISPEECH TEST-CLEAN WITH THE WHISPER-BASE

System	Bitrate (kbps)	WER (%)
Ground Truth	256.0	4.64
Lyra v2 (Codec only)	3.20	7.28
Lyra v2 + VAD	2.80	7.52
Lyra v2 + VAD + Post	2.80	7.18
DAC (Codec only)	3.00	6.70
DAC + VAD	2.62	7.38
DAC + VAD + Post	2.62	6.65

All ASR results were measured using the pre-trained Whisper-base model [13] without fine-tuning. The uncompressed 16 kHz PCM audio served as the ground truth, with a word error rate (WER) of 4.64%. This value represents the baseline performance achievable when no compression is applied. To validate the proposed framework, we utilized two state-of-the-art pre-trained neural codecs, Lyra v2 [8] and DAC [10]. Both codecs were originally trained on large-scale audio corpora using end-to-end neural architectures that jointly optimize the encoder, quantizer, and decoder components. In this study, Lyra operated at 3.2 kbps and DAC at 3.0 kbps, representing comparable low-bitrate conditions. Both models were used in their pre-trained form to ensure consistency and reproducibility. The only component trained specifically for this work was the enhancement network, while all other modules, including the neural codec, VAD model, and ASR model remained frozen. The enhancement module can be seamlessly integrated into various backend machine listening systems, allowing flexible adaptation to different recognition or classification tasks.

The proposed framework first applied a pre-trained Silero-VAD [11] to remove silent or non-speech intervals prior to encoding. The model processed 16 kHz audio using a window size of 32 ms with a 4 ms look-back context, a positive threshold of 0.5, a negative threshold of 0.35, and a minimum silence duration of 100 ms. This configuration provided reliable frame-level detection with minimal temporal discontinuity, resulting in an average bitrate reduction of 12.53% across the evaluation dataset. By discarding silent frames, the effective bitrate was significantly reduced without substantial degradation in ASR performance. To compensate for the quality loss introduced by low-bitrate codecs, we trained a neural post-enhancement model based on MP-SENNet. The model was trained using paired data consisting of clean and coded waveforms generated from the LibriSpeech train-clean-100 dataset. The enhancement networks were optimized for each codec, enabling the model to learn distortion patterns and reconstruction characteristics. The training objectives followed the original MP-SENNet formulation, which combines magnitude, phase, and complex spectral losses along with an STFT consistency loss. These objectives ensure spectral coherence and phase accuracy between the enhanced and clean signals. Unlike perceptually driven enhancement systems, our training excluded GAN-based loss functions, as

the goal was not to improve human listening quality but to enhance features important for machine listening. Instead, the model focused on suppressing coding artifacts and reconstructing the waveform to closely resemble the original clean speech. This approach allows the enhanced signal to retain the acoustic and linguistic information necessary for accurate speech recognition.

The experimental results are summarized in Table 1. For Lyra, the codec-only condition produced a WER of 7.28%. When the VAD was introduced, the effective bitrate dropped from 3.2 kbps to 2.8 kbps, while WER marginally increased to 7.52%. After applying the post-enhancement model, WER improved to 7.18%, achieving comparable performance to the codec-only case while maintaining the reduced bitrate. For DAC, the codec-only system achieved a WER of 6.70% at 3.0 kbps. Applying VAD reduced the bitrate to 2.62 kbps, with a slight increase in WER to 7.38%. Once post-enhancement was applied, the WER decreased again to 6.65%, indicating effective recovery of recognition-relevant features. Overall, both codecs achieved equivalent ASR accuracy to their baselines bitrates while operating at lower bitrates. Compared to the uncompressed ground truth (256 kbps), the proposed framework achieves compression ratios of approximately 91.46x for Lyra and 97.71x for DAC, with less than a 3% absolute degradation in WER. These results confirm that the integration of pre-trained VAD and neural post-enhancement enables efficient, machine-oriented speech transmission under extreme bitrate constraints while maintaining robust ASR performance.

IV. CONCLUSIONS

In this work, we proposed an ASR framework for robust recognition in low-bitrate environments, designed within the context of machine listening. The framework integrates four complementary components in a unified processing chain: a pre-trained VAD model that removes non-speech regions to prevent unnecessary transmission, a generalized neural codec that efficiently compresses speech at low bitrates, a neural post-enhancement module that refines the decoded signal, and an ASR backend that performs final recognition. Unlike feature-based transmission schemes that are limited to a single task, the proposed method demonstrates the potential of generalized codecs for flexible and compatible use across multiple applications. The post-enhancement module focuses on reducing spectral distortion and restoring details lost in heavily quantized speech. It effectively suppresses coding noise, particularly in spectral valleys and transient regions, thereby enhancing clarity and intelligibility crucial for recognition. Experimental evaluation on the LibriSpeech corpus demonstrates that the framework maintains ASR accuracy comparable to higher bitrate systems, achieving WER of 7.18% with Lyra and 6.65% with DAC. These results verify that the combination of silence-aware bitrate reduction and coded signal enhancement can sustain robust recognition performance under severe bandwidth constraints, establishing a practical and scalable approach to machine-oriented speech communication.

ACKNOWLEDGMENT

This work was supported by Institute for Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2017-II170072, Development of Audio/Video Coding and Light Field Media Fundamental Technologies for Ultra Realistic Tera-media)

REFERENCES

- [1] J. E. Parker and S. Dockray, "All possible sounds: speech, music, and the emergence of machine listening," *Sound Studies*, vol. 9, no. 2, pp. 253–281, 2023.
- [2] J. Herre and M. Dietz, "MPEG-4 high-efficiency AAC coding [standards in a nutshell]," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 137–142, 2008.
- [3] M. Neuendorf, M. Multus, N. Rettelbach et al., "MPEG Unified Speech and Audio Coding - the ISO/MPEG standard for high-efficiency audio coding of all content types," in *Proceedings of the 132nd Audio Engineering Society Convention*, Audio Engineering Society, 2012.
- [4] J.-M. Valin, K. Vos, and T. Terriberry, "Definition of the Opus audio codec," *IETF*, September, 2012.
- [5] M. Dietz, M. Multus, V. Eksler et al., "Overview of the EVS codec architecture," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5698–5702.
- [6] W. Lim, I. Jang, S. Beack, and J. Kang, "Standardization trends in audio coding for machines," *Journal of Broadcast Engineering*, vol. 30, no. 3, pp. 279–288, 2025.
- [7] A. Kuznetsova, I. Jang, W. Lim, and M. Kim, "Task-specific audio coding for machines: Machine-learned latent features are codes for that machine," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2025.
- [8] Google Research Team, "Lyra v2: A better, faster, and more versatile speech codec," *Google AI Blog*, September 2022. [Online]. Available: <https://opensource.googleblog.com/2022/09/lyra-v2-a-better-faster-and-more-versatile-speech-codec.html>
- [9] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 30, pp. 495–507, 2021.
- [10] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," in *Advances in neural information processing systems (NeurIPS)*, vol. 36, 2023, pp. 27980–27993.
- [11] Silero Team, "Silero-VAD: pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier," <https://github.com/snakers4/silero-vad>, 2024.
- [12] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra," in *Proceedings of the Interspeech*, 2023, pp. 3834–3838.
- [13] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the International conference on machine learning (ICML)*. PMLR, 2023.
- [14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.