

Task-Specific Audio Coding for Machines: Machine-Learned Latent Features Are Codes for That Machine

Anastasia Kuznetsova¹, Inseon Jang², Wootae Lim², Minje Kim³

¹Indiana University, Bloomington, IN, USA

²Electronics and Telecommunications Research Institute, Daejeon, Korea

³University of Illinois Urbana-Champaign, Champaign, IL, USA

Abstract—Neural audio codecs, leveraging quantization algorithms, have significantly impacted various speech/audio tasks. While high-fidelity reconstruction is paramount for human perception, audio coding for machines (ACoM) prioritizes efficient compression and downstream task performance, disregarding perceptual nuances. This work introduces an efficient ACoM method that can compress and quantize any chosen intermediate feature representation of an already trained speech/audio downstream model. Our approach employs task-specific loss guidance alongside residual vector quantization (RVQ) losses, providing ultra-low bitrates (i.e., less than 200 bps) with a minimal loss of the downstream model performance. The resulting tokenizer is adaptable to various bitrates and model sizes for flexible deployment. Evaluated on automatic speech recognition and audio classification, our method demonstrates its efficacy and potential for broader task and architectural applicability through appropriate regularization.

1. INTRODUCTION

Audio codecs have been an active area of research due to their versatile nature. The primary use of codecs is signal compression with the minimal loss of the perceptual quality for efficient transmission [1]–[6]. More recently, neural audio codecs [7]–[9] played part in facilitating the success of various generative models, turning inherently continuous audio representations into discrete tokens. Various speech and audio related tasks have benefited from discrete tokens: speech language models (SLMs), text-to-speech (TTS) [10], voice conversion (VC) [11]–[13], automatic speech recognition (ASR), audio classification (AC) [14], speech enhancement (SE) [15] among others [16], [17].

Neural audio coding system typically consists of three parts: an encoder that transforms the input audio into a compact (e.g., low-dimensional) feature space, a quantization module that converts the continuous feature vectors into a discrete representation, and the decoder that recovers the original waveform from the de-quantized feature vectors. In other words, if it were not for quantization it is natural to consider it an autoencoder. Indeed, incorporating the quantization module as a trainable part of the model optimization process has been the key to developing a successful neural audio codec. For example, popular neural codecs, e.g., SoundStream [7], Encodec [9], and DAC [8], are trained using residual vector quantization (RVQ) and guided by reconstruction losses to receive high-fidelity output audio signal. In this scenario, the reconstruction loss is still directly [18] or indirectly aiming at retaining the perceptual quality of the original audio, e.g., for speech communication or music streaming.

However, learning to discretize audio for non-human entities has been underexplored. Discriminative audio tasks, such as ASR or AC, do not require the signal to be reconstructed and perceived by humans, thus fine-grained nuances of audio are redundant and bitrate-inefficient. Instead, codes for a downstream machine learning (ML) task can be further optimized under audio coding for machines (ACoM) paradigm

introduced by [19]. ACoM implies generalized approaches to coding, where the coding efficiency is optimized based on the downstream performance, focusing on the utility of such discrete features for the machines while ignoring its perceptual characteristics. ACoM adheres to a set of principles: a) codes must be efficient in bitrate and size; b) can be used by the machines without degrading the downstream performance; c) need to be optimized for machine consumption [19].

One step towards ACoM can be a stream of foundational models, that can be seen as discrete tokenizers via self-supervised learning, such as Wav2Vec2.0 [20], WavLM [21], or HuBERT [22]. They focus mainly on retrieving the phonetic characteristics of speech input [23]. Meanwhile, more general-purpose SSL-based tokenizers, such as BEATs [14], can learn semantic information from general audio as well, improving sound classification tasks. They are designed to learn machine-useful features rather than audio reconstruction, aligning them to the ACoM paradigm. Indeed, the previously mentioned autoencoder-type codecs are also shown to learn codes that are useful for other semantically driven tasks, such as SLMs [12], [24].

Likewise those discretization methods tend to focus on accuracy and quality, often overlooking the complexity and bitrate redundancy [16], [17], [25], [26]. Moreover, striving for a general-purpose universal tokenizer, the downstream tasks are often trained on frozen tokenizers, leading to performance degradation among various tasks [16] in comparison with the continuous baseline.

In this paper we introduce an efficient ACoM method that preserves the performance of the downstream model close to non-quantized version of the model. Instead of aiming for learning a universally working embedding space, we propose to repurpose the feature transformation part of any existing neural network-based downstream model as a tokenizer, while the remainder still performs the supervised task on the token input. The split is useful when part of the downstream task can be divided into two networked entities, e.g., a user device and the cloud server, where the trade-off between the computational cost and representation quality starts to matter. The proposed method finetunes an existing neural network with a guide by a task-specific loss function along with the RVQ losses, ensuring a task-specific bitstream as an intermediate, transmission-friendly feature representation. The proposed method can discretize any internal feature layer output, providing flexibility to beat the desired quality versus model complexity trade-off.

Supervised learning of discrete tokens is not new: in [27], an ASR model is used to train a discrete tokenizer, which turns out to be useful for a subsequent TTS system. However, it does not focus on the tokenizer’s complexity and compression efficiency. On the other hand, BEATs [14] for audio classification presents low bitrate, but applies a multi-stage iterative approach to tokenizer pretraining, which is not always convenient to optimize in practice. In this paper, we show that RVQ-based quantization with proper regularization can serve as means for efficient compression. We evaluate the proposed approach

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [25ZC1100: The research of the basic media contents technologies].

using two popular downstream tasks: ASR and AC, but our method is expandable to a variety of tasks and is architecture independent provided it is optimized with the appropriate regularization.

2. METHOD

Conventional ML pipelines that involve both on-device signal acquisition and cloud computing consist of several stages as shown in Figure 1a. First, the audio received from the microphone on device is compressed using a conventional DSP-based or a neural codec. Following compression, the signal is sent over the network to the cloud, where it is decoded and further processed by the downstream ML model for prediction of the target variables (e.g., class labels). The conventional approach contains several redundancies that can be eliminated: (a) conventional codecs are designed for speech/audio communication considering perceptual quality, and are unaware of the downstream task, resulting in unnecessarily higher bitrates or sacrificed performance on the downstream task (b) neural codecs tend to be too heavy to be deployed in the device, especially if the quality of the code matters (c) given that the downstream model needs to do feature transformation again, the encoding and decoding processes are computationally redundant.

As an alternative, we propose a task-specific, machine oriented codec, which eliminates the necessity of a standalone audio codec from the ML pipeline. Instead, we propose to repurpose the earlier part of the ML model to transform the raw input signal into a compact feature space, which is what those layers are doing anyway, and then introduce quantization in that learned feature space.

The proposed approach offers the following benefits. First, we reduce the bitrate of the system’s encoder, increasing the transmission speed. The quantization process is guided with the task-specific loss, which specializes discrete tokens discarding irrelevant information (e.g., speaker characteristics from the ASR pipeline) as well as eliminates the need for using extra bits of information. Second, the ML pipeline can save the cost of running a codec entirely, as the encoder is replaced by the existing downstream ML model’s layers. Decoding is unnecessary in this pipeline, too. Finally, the proposed scheme is universal and transferrable to many speech and audio tasks. It can accommodate a variety of resource constraints on the device side, allowing for different space constraints while preserving reasonable trade-off between the downstream performance and space.

2.1. Proposed Pipeline

Figure 1b shows the outline of the proposed approach. Here, the main assumption is that the cloud ML model can be split into two parts, the earlier module and remaining part. For example, in a neural network-based model, the first few layers can be offloaded to the edge device for processing, while the remainder reside in the cloud. This kind of split model can be preferred in various use cases, such as when the private user data is preferred to be processed on the device, in order to reduce cloud computing cost, etc. At any rate, by being able to split the model, the entire ML pipeline can flexibly adapt to a particular test-time user environment.

Based on this assumption, we propose to remove the codec part in the conventional scenario. Instead, the proposed ACoM scenario can push some of the early layers of the ML model to the device for processing, which essentially works like the encoder of a codec. Then, the input audio is transformed into a feature vector by M layers of the encoder. Residual vector quantization (RVQ) [7] follows to quantize the feature vector into trained codewords, whose codeword assignment index is sent over to the cloud as a discrete token. Subsequently, the remaining layers in the cloud are optimized to retrieve information

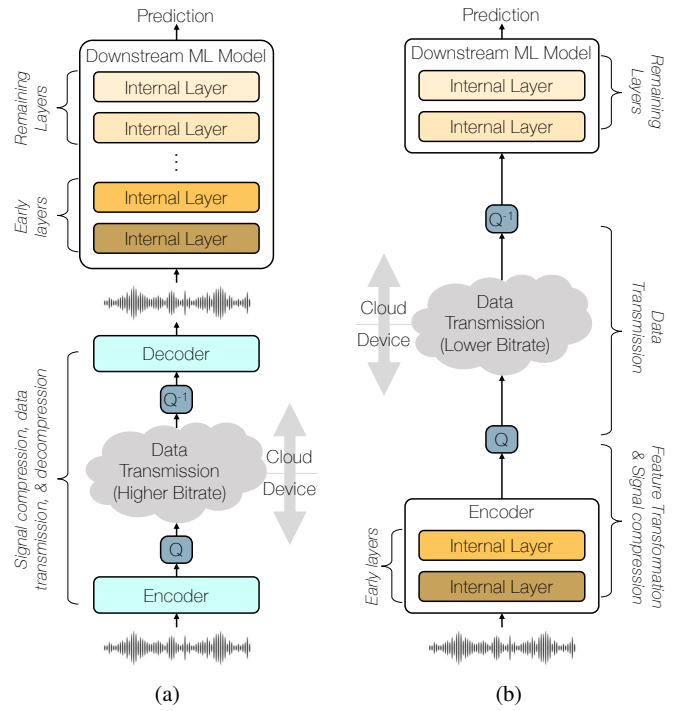


Fig. 1: (a) The conventional ML pipeline involving a codec. The ML model in the cloud is tasked to decompress the signal, perform feature transformation, and the downstream task. (b) The proposed feature coding approach that offloads a few early layers of the downstream model to the device for feature transformation. Trainable quantization on the task-specific feature space ensures more coding gain at a similar downstream performance.

from the dequantized version of the feature vectors. Finally, the downstream ML model performs the original prediction task bearing with the potential quantization error introduced by the RVQ module.

Formally, let $\mathbf{x} \in \mathbb{R}^T$ be raw audio input to the encoder, where T is the number of audio samples. In total the ML model $\mathcal{F}(\cdot)$ contains L layers. Let $\mathcal{F}_i(\cdot)$ represent the i -th layer, for $i = 1, 2, \dots, M$ hosted on device, while $L - M$ layers are hosted in the cloud. Let $\mathbf{h}_i = \mathcal{F}_i(\mathbf{x})$ denote the output of the i -th layer. Then, the on-device encoder is defined as a cascade of M layers: $\mathcal{F}_M \circ \mathcal{F}_{M-1} \circ \dots \circ \mathcal{F}_1(\mathbf{x})$.

2.2. Quantization Mechanism

Subsequently, encoder’s output \mathbf{h}_M is quantized using RVQ [7]. To this end, we define a series of K codebooks, each of which has V D -dimensional codewords: $\mathbf{C} \in \mathbb{R}^{D \times V \times K}$. The first stage of the RVQ process replaces the encoder output \mathbf{h}_M with the closest codeword found in the first codebook $\mathbf{C}_{:, :, 1}$ as follows:

$$\mathbf{h}_M \approx \mathbf{C}_{:, v_1^*, 1}, \text{ where } v_1^* = \arg \min_v \|\mathbf{h}_M - \mathbf{C}_{:, v, 1}\|_2. \quad (1)$$

where v_k^* is the index of the best matching codeword within the k -th codebook. Then, the first residual \mathbf{r}_1 in the code space is calculated, $\mathbf{r}_1 = \mathbf{h}_M - \mathbf{C}_{:, v_1^*, 1}$, to define the new input to the second RVQ stage:

$$\mathbf{r}_1 \approx \mathbf{C}_{:, v_2^*, 2}, \text{ where } v_2^* = \arg \min_v \|\mathbf{r}_1 - \mathbf{C}_{:, v, 2}\|_2. \quad (2)$$

There are K such stages in RVQ to collectively reconstruct the original feature vector $\mathbf{h}_M \approx \hat{\mathbf{h}}_M = \sum_{k=1}^K \mathbf{C}_{:, v_k^*, k}$. We denote this RVQ process by $\mathcal{Q}(\mathbf{x})$, which results in K indices to the closest codewords for K codebooks, respectively: $\mathbf{v}^* = [v_1^*, v_2^*, \dots, v_K^*] = \mathcal{Q}(\mathbf{x})$. In

this way, instead of sending \mathbf{h}_M to the remaining layers in the cloud, the transmission of the discrete token \mathbf{v}^* , suffices for the approximation. The rest of the cloud layers process the dequantized features as input: $\hat{\mathbf{y}} = \mathcal{F}_L \circ \mathcal{F}_{L-1} \circ \dots \circ \mathcal{F}_{M+1}(\hat{\mathbf{h}}_M)$, where $\hat{\mathbf{y}}$ is the prediction of the ground-truth target \mathbf{y} .

The empirical bitrate of the proposed model depends on the number of codebooks K , the size of each codebook V , and finally, the distribution of codeword assignment. Let R be the frame rate, then the raw bitrate (i.e., before considering the entropy of the code distribution) can be calculated as follows: $\text{Bitrate}_{\text{raw}} = R \cdot K \cdot \lceil \log_2 V \rceil$.

Finally, coding efficiency of the proposed approach can be further improved via entropy coding, e.g., Huffman coding as shown in the soft-to-hard quantization method [28], where the entropy of the codes was regularized by a loss term. In this work, instead of controlling the entropy directly, we adopt the *codebook utilization* concept proposed in DAC [8], where codebook under-utilization corresponds to low empirical entropy. In DAC, there was no specific use of entropy coding, but we bring this concept back to our model, associating it with codebook utilization during our experiments. Eventually, our method benefits from entropy coding, e.g., when a codebook is under-utilized, we assume that entropy coding can follow up for a further bitrate reduction, leveraging the low-entropy code distribution. Hence, we report entropy-based bitrates as the model’s compression performance. To this end, we first calculate the frequency of each codeword of the k -th codebook given the entire test samples $\{\mathbf{x}^{(i)}\}_{i=1}^N$:

$$\rho_v = \frac{1}{N} \sum_{i=1}^N \mathcal{I}(v_k^{*(i)} = v), \quad \forall v = \{1, \dots, V\}, \quad (3)$$

where $\mathcal{I}(\cdot)$ is the indicator function. Then, the frequency vector ρ is used to calculate the empirical entropy of the k -th codebook $\mathcal{H}(\mathbf{C}_{:::,k})$.

$$\mathcal{H}(\mathbf{C}_{:::,k}) = - \sum_{v=1}^V \rho_v \log \rho_v. \quad (4)$$

Eventually, the final entropy of all codebooks is the sum of their empirical entropy values: $\mathcal{H}(\mathbf{C}) = \sum_{k=1}^K \mathcal{H}(\mathbf{C}_{:::,k})$. The empirical entropy is the lower bound of a Huffman coder’s bitrate for a frame. Considering the frame rate, the final bitrate is bounded by

$$\text{Bitrate}_{\text{ent}} \geq R \cdot \mathcal{H}(\mathbf{C}). \quad (5)$$

3. EXPERIMENTAL SETUP

We empirically show the coding efficiency and model complexity reduction properties of the task-specific quantization method. First, we lower the bitrate of the proposed models as much as possible while preserving the quality of the continuous baseline model. Second, we assess the trade-off between the coding efficiency, the downstream performance, and computational complexity. In particular, we define our “encoder” flexibly by varying the number of the on-device layer M to investigate this trade-off: a small M makes on-device processing more affordable at the cost of suboptimal coding gain. On the contrary, a large M lets the quantizer work in a more abstract (i.e., easier to compress) feature space, although it could increase the on-device encoder complexity accordingly. An alternative training strategy is to use discrete RVQ tokens as input features to encoder-decoder model. However, in that case we would have to retrain the whole system from scratch which is computationally inefficient compared to the quantization of fully trained model, and thus defeats the purpose of the proposed method. Additionally, we use DAC [29] to simulate Fig. 1a, whose reconstruction is fed to the continuous baseline models to show that the conventional pipeline may not be optimal both in terms of computational efficiency and quality.

We establish the universality of the proposed model quantization approach by experimenting with speech recognition (ASR) and audio classification (AC) systems, showcasing its effectiveness on both speech data (ASR) and general audio discriminative tasks (AC).

3.1. Models

For ASR, we chose a Conformer-Transformer architecture with 12 encoder and 6 decoder layers as defined in SpeechBrain toolkit [30]. The continuous model serves as a baseline, which is also used to load the pretrained weights that are later finetuned along with the RVQ loss functions. The ASR model is trained using 5,000 BPE subwords [31], Connectionist Temporal Classification (CTC) [32] loss as well as KL-divergence as a label smoothing loss, guided by Adam optimizer [33]. In the quantized version of the ASR model, we add codebook and commitment losses as regularizers as in the original VQ-VAE method [34] that the SoundStream model [7] used, thus the total ASR loss is defined as (6), where λ and β are the coefficients for ASR-related losses and the RVQ regularizer, respectively.

$$\mathcal{L}_{\text{ASR}} = \lambda \mathcal{L}_{\text{CTC}} + (1 - \lambda) \mathcal{L}_{\text{KL}} + \beta (\mathcal{L}_{\text{code}} + \mathcal{L}_{\text{commit}}) \quad (6)$$

As the baseline for the AC task, we also use SpeechBrain’s ECAPA TDNN [35] architecture with dilated convolutions, squeeze and excitation blocks, and attentive statistical pooling. The main objective is the additive angular margin softmax loss [36] accompanied by the codebook loss regularizer as in Eq. 6. The model contains 4 encoder layers followed by attentive statistical pooling and a linear classifier. Before applying RVQ, we apply average pooling across the time dimension to lower the original frame rate $R = 160$ to 40.

We perform hyperparameter optimization using the tree-structured Parzen estimator (TPE) algorithm [37], [38]. For ASR, we tune batch sizes, learning rates, and the code vector dimension D . For AC, we additionally tune the β hyperparameter for quantizer regularization.

3.2. Data

For ASR experiments, we use the conventional LibriSpeech [39] dataset containing 960h of training data with a maximum audio length of 10 seconds. Only *train-clean-100* is used for hyperparameter search.

For AC UrbanSound8k dataset [40] is used. It contains 8,732 sound excerpts ≤ 4 seconds long that are annotated for 10 sound classes, such as air conditioner, gun shot, siren, children playing, etc. We use the original 10-fold split and perform cross-validation as suggested by authors. Both datasets are sampled at 16 kHz.

3.3. Evaluation

The quality of the ASR system is measured in the word error rate (WER) reported on *test-clean* and *test-other* subsets of LibriSpeech [39]. Note that the lower WER is the better, which we denote by \downarrow in the result tables. The performance of the AC is assessed via classification accuracy (Acc), which is better if it is higher, i.e., \uparrow . To evaluate the coding gain after entropy coding we compute entropy based bitrate ($\text{Bitrate}_{\text{ent}, \downarrow}$) defined in Eq. (5). Additionally, to analyze model complexity, we compute the number of giga multiply-accumulate operations (GMAC) for a one-second audio sample using PyFlops¹. We compute GMACs \downarrow of our M encoder layers as well as the RVQ modules to compare them with the 16 kHz version of DAC’s encoding, RVQ, and decoding processes, as a simulation of the conventional scenario when the audio needs to go through the neural codec first (Fig. 1a). We also report raw and entropy-based bitrates for both ASR and AC datasets.

¹<https://github.com/sovrsov/flops-counter.pytorch/tree/master>

Table 1: Results for ASR on LibriSpeech [39] with different depths of encoder quantization. GMACs \downarrow are reported to measure the on-device portion of the processing pipeline, in comparison to conventional coding approach based on DAC quantization. We only compute MACs for the DAC bitrates corresponding to quantized ASR models’ bitrates. WER \downarrow is reported on *test-clean* and *test-other* subsets.

No. Codebooks	Quant. layer	Codebook size&dim	BR _{raw} (bps)	WER (test-clean)	WER (test-other)	Entropy (frame)	BR _{ent} (bps)	GMACs (on-device)
Cont. baseline	-	-	-	2.01	4.52	-	-	-
Cont. DAC 250	-	-	-	2.94	19.91	44.99	174.68	12.30
Cont. DAC 500	-	-	-	2.94	4.53	13.99	730.74	12.30
12	4	1024, 512	3000	4.06	10.1	4.76	1428.30	2.85
12	6	1024, 512	3000	3.04	7.28	5.14	1542.45	4.22
12	8	1024, 512	3000	2.87	6.79	5.21	1563.36	5.48
2	4	1024, 64	500	2.25	5.33	9.98	499.06	2.95
2	6	1024, 64	500	2.23	5.01	10.85	542.52	4.32
2	8	1024, 64	500	2.21	4.99	10.95	547.59	5.69
1	4	8192, 8	325	2.87	7.53	6.38	159.41	2.75
1	6	8192, 8	325	3.21	8.23	5.76	144.12	4.12
1	8	8192, 8	325	3.03	7.33	6.05	151.33	5.49
1	4	1024, 64	250	2.93	7.66	5.34	133.49	2.75
1	6	1024, 64	250	2.9	7.34	5.27	131.68	4.12
1	8	1024, 64	250	2.79	6.89	5.26	131.62	5.49

4. RESULTS

Table 1 presents WERs for the ASR task on the *test-clean* and *test-other* subsets of LibriSpeech, along with the corresponding bitrates, the on-device portion of the model’s GMACs, and corresponding quantization configurations. The table compares a continuous baseline ASR model with models with different quantization strategies. As a simulation of Fig. 1a scenario, we finetune an ASR model using waveforms reconstructed by a DAC at 250 bps and 500 bps, to evaluate the neural codec’s effects on the ASR performance. In this case, DAC’s encoder and its RVQ routine is the only on-device operations, amounting to a static 12.3 GMACs.

The continuous baseline ASR model achieves WERs of 2.01 and 4.52 on *test-clean* and *test-other* respectively, which is the performance upper bound of any systems involving quantization. Retraining the same continuous model on DAC reconstructed waveforms at 250 bps leads to a degradation in performance: 2.94 for *test-clean* and 7.58 for *test-other* at 250 bps. At a higher rate of 500 bps, the WERs are similar (2.94 and 7.12), which deems the perceptual quality improvement inefficient for ACoM.

The proposed method (Fig. 1b) introduces RVQ to any selected ASR encoder layer instead of a separate waveform coding. In doing so, we also tested various hyperparameters that affect the compression ratio and the loss of information. As for the number of codebooks that linearly increases the raw bitrate, we found that two codebooks are enough for the ASR task, while 12 codebooks are adding additional burden to the optimization process, resulting in lower ASR performance. Of the three chosen encoder layers, 4th, 6th, and 8th, we found a steady trend that the quantization in the higher layer results in better performance (4.99 in 8th), especially for the noisy test set *test-other*. This improvement comes at the cost of increased on-device run-time complexity from 2.95 to 5.69 GMACs. Overall, we observe the best WERs, 2.21 and 4.99, at the 8th layer with two codebooks, that are close to the continuous baseline’s. Compared to the Fig. 1a pipeline, ours achieves much better WERs, especially on the noisy input (7.12 vs. 4.99), even without the computational

Table 2: Results for audio classification (Acc \uparrow) on UrbanSound8k [40]. The accuracy Acc \uparrow is reported on the provided test set. The second row Cont. DAC 750 shows the result for DAC reconstructed signals at 750 bps, closest to 800 bps quantized AC system.

No. Codebooks	Quantized layer	Codebook size&dim	BR _{raw} (bps)	Test Acc	Entropy (frame)	BR _{ent} (bps)	MMACs (on-device)
Cont. baseline	-	-	-	0.797	-	-	-
Cont. DAC 250	-	-	-	0.591	6.23	155.86	12,300
Cont. DAC 750	-	-	-	0.749	20.37	1528.95	12,300
2	1	1024, 8	800	0.761	11.84	947.02	42.34
2	2	1024, 8	800	0.762	11.55	924.00	287.02
2	4	1024, 64	800	0.782	3.14	251.44	784.99
1	1	1024, 8	400	0.761	5.62	224.60	41.7
1	2	1024, 8	400	0.758	5.43	217.25	286.38
1	4	1024, 512	400	0.793	9.92	396.56	814.46
1	1	32, 8	200	0.694	3.50	139.93	41.5
1	2	32, 8	200	0.743	2.71	108.53	286.18
1	4	32, 512	200	0.802	4.21	168.44	801.76

overhead (12.3 GMACs) that the DAC module comes with.

The raw bitrate of the two-codebook quantizers is 500 bps, which is already significantly low. In theory, we can further compress the bitstream via an entropy coding scheme, although in this particular case, due to the relatively high per-frame entropy, entropy coding did not result in additional compression. However, it is worth mentioning the very low entropy bitrates of one codebook cases, that achieve 131.62 bps, reaching the theoretical bound of speech communication, 100 bps [41]. The ultra-low bitrate coding scheme sacrifices the WER performance, accordingly (2.79 and 6.89), although they could be an acceptable performance considering the amount of bitrate saving. In addition, they are still a better solution than the DAC and the continuous ASR pipeline in terms of both WER and complexity.

Audio classification results outlined in Table 2 confirm our observations. Our quantizer with one codebook, when applied to the 4th classifier layer, results in slightly higher than the continuous baseline classification accuracy (80.2 vs. 79.9%). Given the raw bitrate of 200 bps, after entropy coding the lower bound goes down to 168.44 bps. As opposed to ASR results, continuous baseline trained from the DAC reconstructed audio at 750 bps does not deteriorate the quality of the AC much. However, there still exists a significant gap in computational efficiency due to DAC’s computational overhead.

Overall, we observe the tradeoff between the choice of quantization layer, bitrates, and classification accuracies, which is a design choice depending on how much computation the device can afford. However, the overall trend is that the classification performance can be maintained, while the coded feature representations can be extremely compressed, which is a huge advantage in split architecture.

5. CONCLUSION

We explored task-specific quantization strategy that improved computational efficiency of encoder-decoder tasks (ASR and AC) under the ACoM paradigm. Leaving out the factor of human perception in the training allowed to balance quality and computational complexity of the models via making the RVQ based representations more suitable for machine ‘understanding’. Splitting the model between the device and the cloud offers flexibility for edge devices of various constraints, and showed the potential for achieving ultra-low bitrates while preserving reasonable performance. Source codes can be found at: <https://minjekim.com/research-projects/acom#waspaa2025>.

REFERENCES

- [1] ISO/IEC 11172-3:1993, “Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s,” 1993.
- [2] T. Painter and A. Spanias, “Perceptual coding of digital audio,” *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.
- [3] ISO (2006) ISO/IEC 13818-7:2006, “Information technology — Generic coding of moving pictures and associated audio information — Part 7: Advanced Audio Coding (AAC),” 2006.
- [4] ISO/IEC - Information technology – MPEG audio technologies – Part 2: Spatial Audio Object Coding (SAOC), ISO/IEC IS 23 003-2, 2010.
- [5] ISO/IEC 14496-3:2009/PDAM 3, “Transport of unified speech and audio coding (USAC),” 2011.
- [6] ITU-T Recommendation G.722.2, “Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB),” 2003.
- [7] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, “Soundstream: An end-to-end neural audio codec,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 30, p. 495–507, jan 2022.
- [8] 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)*, 2023.
- [9] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High fidelity neural audio compression,” *Transactions on Machine Learning Research*, 2023.
- [10] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, “Neural codec language models are zero-shot text to speech synthesizers,” *arXiv preprint arXiv:2301.02111*, 2023.
- [11] K. Lakhota, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. and Mohamed *et al.*, “On generative spoken language modeling from raw audio,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.
- [12] W. Cui, D. Yu, X. Jiao, Z. Meng, G. Zhang, Q. Wang, Y. Guo, and I. King, “Recent advances in speech language models: A survey,” *arXiv preprint arXiv:2410.03751*, 2024.
- [13] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, “Neural codec language models are zero-shot text to speech synthesizers,” *arXiv preprint arXiv:2301.02111*, 2023.
- [14] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, “Beats: Audio pre-training with acoustic tokenizers,” *arXiv preprint arXiv:2212.09058*, 2022.
- [15] H. Yang, J. Su, M. Kim, and Z. Jin, “Genhancer: High-fidelity speech enhancement via generative modeling on discrete codec tokens,” in *Proc. Interspeech*, 2024.
- [16] P. Mousavi, L. D. Libera, J. Duret, A. Ploujnikov, C. Subakan, and M. Ravanelli, “DASB - discrete audio and speech benchmark,” 2024.
- [17] H. Wu, H.-L. Chung, Y.-C. Lin, Y.-K. Wu, X. Chen, Y.-C. Pai, H.-H. Wang, K.-W. Chang, A. Liu, and H. Lee, “Codec-SUPERB: An in-depth analysis of sound codec models,” in *Findings of the Association for Computational Linguistics ACL 2024*, Aug. 2024, pp. 10 330–10 348.
- [18] K. Zhen, M. S. Lee, J. Sung, S. Beack, and M. Kim, “Psychoacoustic calibration of loss functions for efficient end-to-end neural audio coding,” *IEEE Signal Processing Letters*, vol. 27, pp. 2159–2163, 2020.
- [19] MPEG, “Use Cases and Requirements on Audio Coding for Machines,” International Organisation for Standardisation, Tech. Rep. N0046. [Online]. Available: <https://www.mpeg.org/standards/Explorations/46/>
- [20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [21] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
- [23] S. Arora, K.-W. Chang, C.-M. Chien, Y. Peng, Y. Wu, H. and Adi, E. Dupoux, H.-Y. Lee, K. Livescu, and S. Watanabe, “On The Landscape of Spoken Language Models: A Comprehensive Survey,” *arXiv preprint arXiv:2504.08528*, 2025.
- [24] H. Wu, X. Chen, Y.-I. Lin, K. Chang, H.-L. Chung, A. H. Liu, and H. Lee, “Towards audio language modeling—an overview,” *arXiv preprint arXiv:2402.13236*, 2024.
- [25] J. Shi, J. Tian, Y. Wu, J.-W. Jung, J. Yip, Y. Masuyama, W. Chen, Y. Wu, Y. Tang, M. Baali, D. Alharthi, D. Zhang, R. Deng, T. Srivastava, H. Wu, A. Liu, B. Raj, Q. Jin, R. Song, and S. Watanabe, “Espnet-codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech,” in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 562–569.
- [26] Y. Guo, Z. Li, H. Wang, B. Li, C. Shao, H. Zhang, C. Du, X. Chen, S. Liu, and K. Yu, “Recent advances in discrete speech tokens: A review,” *arXiv preprint arXiv:2502.06490*, 2025.
- [27] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma *et al.*, “CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens,” *arXiv preprint arXiv:2407.05407*, 2024.
- [28] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. V. Gool, “Soft-to-hard vector quantization for end-to-end learning compressible representations,” in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 1141–1151.
- [29] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, “High-fidelity audio compression with improved RVQGAN,” *arXiv preprint arXiv:2306.06546*, 2023.
- [30] M. Ravanelli, T. Parcollet, A. Moumen, S. de Langen, C. Subakan, P. Plantinga, Y. Wang, P. Mousavi, L. D. Libera, A. Ploujnikov, F. Paissan, D. Borra, S. Zaiem, Z. Zhao, S. Zhang, G. Karakasidis, S.-L. Yeh, P. Champion, A. Rouhe, R. Braun, F. Mai, J. Zuluaga-Gomez, S. M. Mousavi, A. Nautsch, H. Nguyen, X. Liu, S. Sagar, J. Duret, S. Mdhaffar, G. Laperrière, M. Rouvier, R. D. Mori, and Y. Estève, “Open-Source Conversational AI with SpeechBrain 1.0,” *Journal of Machine Learning Research*, vol. 25, no. 333, 2024.
- [31] T. Kudo and J. Richardson, “SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71.
- [32] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.
- [33] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2015.
- [34] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning,” in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 6306–6315.
- [35] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” *arXiv preprint arXiv:2005.07143*, 2020.
- [36] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, “Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition,” in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1652–1656.
- [37] S. Watanabe, “Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance,” *arXiv preprint arXiv:2304.11127*, 2023.
- [38] X. Bouthillier, C. Tsirigotis, F. Corneau-Tremblay, T. Schweizer, L. Dong, P. Delaunay, F. Normandin, M. Bronzi, D. Suhubdy, R. Askari, M. Nounkovich, C. Xue, S. Ortiz-Gagné, O. Breuleux, A. Bergeron, O. Bilaniuk, S. Bocco, H. Bertrand, G. Alain, D. Serdyuk, P. Henderson, P. Lamblin, and C. Beckham, “Epistimio/orion: Asynchronous Distributed Hyperparameter Optimization,” 2022. [Online]. Available: <https://doi.org/10.5281/zenodo.3478592>
- [39] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [40] J. Salamon, C. Jacoby, and J. Bello, “A Dataset and Taxonomy for Urban Sound Research,” in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, p. 1041–1044. [Online]. Available: <https://doi.org/10.1145/2647868.2655045>
- [41] S. V. Kuyk, W. B. Kleijn, and R. C. Hendriks, “On the information rate of speech communication,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5625–5629.