ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11462567

ENTROPY-GUIDED GRVQ FOR ULTRA-LOW BITRATE NEURAL SPEECH CODEC

Yanzhou Ren[*], *Noboru Harada*[†], *Daiki Takeuchi*[†], *Siyu Chen*[*], *Wei Liu*[*], *Xiao Zhang*[*], *Liyuan Zhang*[*], *Takehiro Moriya*[†], *and Shoji Makino*[*]

[*] Waseda University, Japan
[†] NTT, Inc., Japan

ABSTRACT

Neural audio codec (NAC) is essential for reconstructing high-quality speech signals and generating discrete representations for downstream speech language models. However, ensuring accurate semantic modeling while maintaining high-fidelity reconstruction under ultra-low bitrate constraints remains challenging. We propose an entropy-guided group residual vector quantization (EG-GRVQ) for an ultra-low bitrate neural speech codec, which retains a semantic branch for linguistic information and incorporates an entropy-guided grouping strategy in the acoustic branch. Assuming that channel activations follow approximately Gaussian statistics, the variance of each channel can serve as a principled proxy for its information content. Based on this assumption, we partition the encoder output such that each group carries an equal share of the total information. This balanced allocation improves codebook efficiency and reduces redundancy. Trained on LibriTTS and VCTK, our model shows improvements in perceptual quality and intelligibility metrics under ultra-low bitrate conditions, with a focus on codec-level fidelity for communication-oriented scenarios.

***Index Terms*—** Neural audio codec, discrete representation, semantic information, entropy-guided quantization, codebook efficiency.

1. INTRODUCTION

Neural audio codec (NAC) has gained increasing importance in modern speech processing by producing discrete representations that serve different purposes in speech applications [1–6]. Prior studies categorize discrete speech representations into two main types: semantic tokens and acoustic tokens [7]. Semantic tokens capture high-level linguistic information and are widely adopted in downstream speech-language modeling tasks [8]. They are commonly extracted from self-supervised pretrained models such as HuBERT and WavLM [9, 10]. They can subsequently be incorporated into neural codecs through semantic-aware quantization [11–13], enabling accurate content alignment for applications such as speech-to-text, speech translation, and zero-shot text-to-speech [14–17]. Acoustic tokens, typically obtained from neural audio codecs, are designed to preserve fine-grained signal details for waveform reconstruction. Neural codecs such as SoundStream [18], EnCodec [19], and Stable Codec [20] rely on residual vector quantization (RVQ) or finite scale quantization (FSQ), while HIFI-Codec [21] further introduced grouped residual vector quantization (GRVQ) to improve perceptual quality by distributing information across quantization groups. It is worth noting that the vector quantization of coefficients for ultra-low bit-rate codecs was an essential research topic over 35 years ago [22], and the similar topic is now being revisited in the context of neural speech and audio coding.

Although semantic and acoustic tokens provide complementary benefits, ensuring both under ultra-low bitrate constraints remains an open challenge. Recent works have attempted to integrate both aspects within a single codec. For example, Mimi [13] introduces a parallel quantization design with dedicated semantic and acoustic information branches. It utilizes an RVQ to code acoustic information. It also allocates bitrate capacity to semantic tokens, which inevitably reduces the resources for acoustic details [23], thus constraining simultaneous improvement of intelligibility and reconstruction quality, especially for ultra-low bitrate conditions.

In this paper, we adopt Mimi [13] as our baseline, for its causal, low-latency architecture suitable for real-time communication scenarios, with particular emphasis on strengthening the acoustic branch to achieve higher reconstruction fidelity. We propose an entropy-guided group residual vector quantization (EG-GRVQ) strategy, which extends the GRVQ framework originally introduced in HiFi-Codec [21]. The EG-GRVQ is motivated by the information-theoretic link between variance and entropy, where channel variance serves as a proxy for information content. Encoder outputs are partitioned to equalize cumulative variance across groups, leading to a more balanced allocation of information and improved codebook efficiency. While explicitly retaining Mimi's semantic branch, our design ensures that the acoustic branch achieves higher fidelity under low-bitrate conditions. Experiments on LibriTTS [24] and VCTK [25] confirm that our approach yields improvements in both perceptual quality and intelligibility.

2. PROPOSED METHOD

2.1. Encoder and Decoder

The encoder and decoder of our proposal follow the overall architecture of Mimi [13]. As shown in Fig. 1, the input is 24 kHz waveforms, which are first passed through four residual convolutional blocks and finally a 1D convolution in the encoder. These blocks progressively reduce the temporal resolution and increase the channel dimension, ultimately transforming the waveform into 512-dimensional latent representations at a rate of 12.5 frames per second. Following the convolutional blocks, Transformer blocks are applied in the bottleneck to capture long-range dependencies and enhance the compactness of latent representations. The decoder adopts a symmetric architecture to the encoder, operating in reverse to reconstruct the waveform. It takes the quantized latent features as input and reconstructs the waveform through the transposed convolutional blocks that progressively upsample the temporal resolution and reduce the channel dimensions.

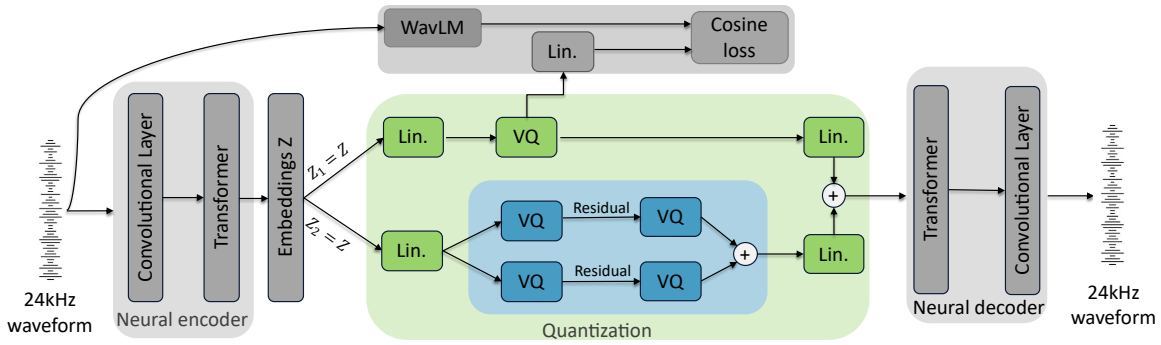

Fig. 1: Structure of the proposed model.

2.2. Quantization

As shown in Fig. 1, after encoding, the latent features are duplicated into two branches. The first branch is processed by a semantic quantizer, designed to capture high-level linguistic features that are critical for speech intelligibility. This semantic layer employs a single codebook to minimize bitrate consumption while still preserving essential semantic information.

The second branch is processed by an entropy-guided grouped residual vector quantizer. Fig. 2 shows the three different quantizer structure configurations for acoustic information: (a) RVQ, (b) GRVQ and (c) EG-GRVQ (Proposal). While RVQ sequentially utilizes a series of residual codebooks to quantize the target acoustic vector Z_2, the conventional GRVQ splits the channels of the acoustic target vector Z_2 into two coding groups.

Unlike conventional GRVQ approaches that divide channels evenly, our method, EG-GRVQ, exploits the global variance distribution of encoder outputs to guide grouping. Let $z_{k,t}$ denote the encoder activation at channel $k \in \{1, \ldots, C\}$ and time frame $t \in \{1, \ldots, T\}$. The variance of channel k over the training set is defined as

$$\sigma_k^2 = \frac{1}{T} \sum_{t=1}^{T} \left(z_{k,t} - \mu_k\right)^2, \tag{1}$$

where μ_k is the average output of channel k.

From an information-theoretic perspective, we assume that each encoder channel, after subtracting its mean constant bias, approximately follows a Gaussian distribution [26]. Accordingly, the differential entropy is given by

$$H(X) = \tfrac{1}{2} \ln \left(2\pi e \sigma^2\right), \tag{2}$$

indicating that entropy is a monotonic function of variance. Hence, channel variance can serve as a proxy for the amount of information carried by each channel.

We then determine the smallest index k^* such that

$$\sum_{i=1}^{k^*} \sigma_i^2 \quad \geq \quad \tfrac{1}{2} \sum_{j=1}^{C} \sigma_j^2, \tag{3}$$

ensuring that the first group accounts for approximately half of the total variance. In our implementation with $C = 512$, the split point is $k = 237$, producing Group 1 (Codebooks 1 and 3 as shown in Fig. 2-(c)) with 237 channels and Group 2 (Codebooks 2 and 4 as shown in Fig. 2-(c)) with 275 channels. The even split at $k = 256$ allocates 55.30 % of the variance to the first half and 44.70 % to the second half. This strategy mitigates the dominance of high-variance channels and the under-utilization of low-variance channels, thereby enhancing quantization efficiency and reconstruction fidelity.

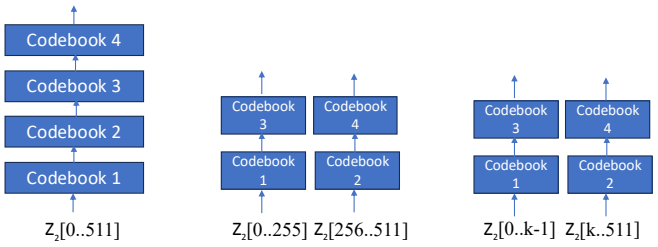

Fig. 2: Quantizer structure configuration for (a) RVQ, (b) GRVQ, and (c) EG-GRVQ (Proposal).

Note that when some signaling bits are allocated, the split position k can be adaptive in each frame. However, considering the trade-off for spending additional bits and coding gain improvement, no adaptation has been applied in this proposal. The resulting grouping remains fixed as a hyperparameter for all data during both training and inference.

Finally, the quantized features from semantic branch and acoustic branch are summed and passed to the decoder for waveform reconstruction.

3. EXPERIMENTS

3.1. Baselines

In our experiments, we consider four systems as baselines for comparison. We denote the original Mimi codec without retraining as Mimi (official), the retrained version with residual vector quantization as Mimi (retrain), and our re-implementation of grouped residual vector quantization within the Mimi framework as Mimi (GRVQ). Our proposed method, based on entropy-guided grouping, is referred to as Proposal (EG-GRVQ).

3.2. Datasets and Experimental Setup

We trained our proposed method on the combined LibriTTS [24] and VCTK [25], including the train-clean-100, train-clean-360, and train-other-500 subsets of LibriTTS, as well as the full VCTK corpus. To ensure a fair and objective comparison, we retrained the

Table 1: Objective evaluation results at 0.6875 kbps.

Method	VQ scheme (Structure)	SDR↑	PESQ↑	STOI↑	ViSQOL↑
Mimi (official)	RVQ (1x4)	**3.451**	1.872	0.876	2.010
Mimi (retrain)	RVQ (1x4)	-6.969	1.779	0.886	**2.546**
Mimi (GRVQ)	GRVQ (2x2)	-7.294	1.852	0.889	2.464
Proposal	EG-GRVQ (2x2)	-7.309	**1.881**	**0.890**	2.496

Table 2: Normalized Mean Squared Error (NMSE) across acoustic quantizers.

Method	Codebook1	Codebook2	Codebook3	Codebook4	Total
RVQ	0.928	0.908	0.894	0.884	0.884
GRVQ	0.890	0.895	0.842	0.862	0.852
Proposal (EG-GRVQ)	**0.865**	**0.855**	**0.831**	**0.813**	**0.819**

Mimi baseline under the same data conditions. For comparison, we also re-implemented the previously proposed GRVQ quantization [21] within the Mimi framework, denoted as Mimi (GRVQ). All trainings and evaluations were conducted using the 5-codebook setting under the same conditions. All models were trained on 8 NVIDIA A6000 GPUs (48 GB each), with a batch size of 12 per GPU.

3.3. Training Strategy

To train the proposed model, we adopt a multiobjective strategy that integrates feature loss and adversarial loss, with a discriminator-based scheme to enhance perceptual quality. The generator is optimized by minimizing the following composite loss:

$$\mathcal{L}_{\text{gen}} = \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{feat}}\mathcal{L}_{\text{FM}}, \tag{4}$$

where:

- $\mathcal{L}_{\text{adv}}$ is the adversarial loss, computed as the mean squared error (MSE) between discriminator predictions and the target label 1 (real):

$$\mathcal{L}_{\text{adv}} = \frac{1}{N}\sum_{i=1}^{N}(D(\hat{x}_i) - 1)^2. \tag{5}$$

- $\mathcal{L}_{\text{FM}}$ is the feature matching loss, defined as the L_1 distance between real and generated intermediate discriminator features:

$$\mathcal{L}_{\text{FM}} = \frac{1}{N}\sum_{l=1}^{N}\left\|D^{(l)}(x) - D^{(l)}(\hat{x})\right\|_1, \tag{6}$$

  where N is the total number of feature maps across all layers and scales.

We set $\lambda_{\text{adv}} = 1$ and $\lambda_{\text{feat}} = 15$. In addition, a vector quantization commitment loss $\mathcal{L}_{\text{commit}}$, weighted by $\lambda_{\text{commit}} = 1$, is included to stabilize codebook usage.

For semantic learning, we followed Speechtokenizer [12] by using a distillation loss between projected semantic quantizer outputs and WavLM-derived embeddings. The latent features from the semantic quantizer are projected to a 1024-dimensional space and then compared with WavLM-derived semantic embeddings using a cosine similarity loss L_{sem}. The semantic distillation loss L_{sem} is computed as:

$$\mathcal{L}_{\text{sem}} = 1 - \cos(z_s, z_{\text{WavLM}}). \tag{7}$$

3.4. Objective Evaluation

For objective evaluation, we randomly selected 200 samples from the LibriTTS [24] test-clean subset, which was strictly excluded from training. To prevent information leakage, we ensured that there was no data overlap between the training and evaluation sets. We employ four objective metrics to assess the reconstruction quality: Perceptual Evaluation of Speech Quality (PESQ) [27], Short-Time Objective Intelligibility (STOI) [28], Virtual Speech Quality Objective Listener (ViSQOL) [29], and Signal-to-Distortion Ratio (SDR) [30]. Specifically, we adopt the wideband version of PESQ for perceptual quality, STOI for intelligibility under low-bitrate conditions, and ViSQOL for reference-based perceptual correlation. SDR, although widely used in separation and enhancement tasks, is sensitive to phase misalignment; therefore, we report it only as a supplementary metric. During evaluation, all model outputs, originally at 24 kHz and 16-bit resolution, are downsampled to 16 kHz using a 7.5 kHz low-pass filter to meet the input requirements of PESQ and STOI.

As shown in Table 1, our proposed EG-GRVQ achieves the best performance in PESQ and STOI, while maintaining competitive results in ViSQOL. Compared with Mimi (official), our method improves PESQ by 0.01, ViSQOL by 0.49, and STOI by 0.01, demonstrating substantial gains in perceptual quality. Relative to Mimi (retrain), which is trained on the same data, our method achieves an increase of 0.1 in PESQ and 0.01 in STOI while yielding comparable SDR (-7.31 vs. -6.97). Compared with Mimi (GRVQ), our method further improves PESQ from 1.85 to 1.88 and ViSQOL from 2.46 to 2.50. These results validate the effectiveness of the proposed entropy-guided design in enhancing perceptual quality under low bitrate conditions.

In addition, we evaluate the normalized mean squared error (NMSE) between the quantized reconstruction and the oracle encoder output, which provides a scale-invariant measure of distortion. Table 2 shows the NMSE results across acoustic quantizers. The proposed method consistently achieves lower NMSE, confirming that our grouping strategy yields more accurate reconstruction relative to latent variance.

3.5. Codebook Efficiency and Grouping Analyses

We compare codebook utilization across the acoustic branch to evaluate the effectiveness of different quantization strategies. As shown in Fig. 3, the RVQ baseline exhibits a sharp decline in utilization

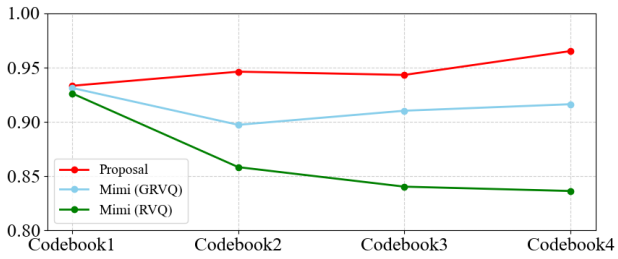

Fig. 3: Codebook utilization rate in acoustic branch.

Table 3: Objective results of different grouping numbers at 0.6875 kbps.

Method	SDR↑	PESQ↑	STOI↑	ViSQOL↑
1×4 (RVQ)	**-6.969**	1.779	0.886	**2.546**
4×1	-7.420	1.736	0.875	2.026
2×2 (GRVQ)	-7.294	**1.852**	**0.889**	2.464

in deeper layers, indicating inefficient codeword usage and potential room for improvement. The Mimi (GRVQ) baseline alleviates this problem to some extent by balancing the allocation across groups, but still suffers from layer-wise imbalance. In contrast, the proposed EG-GRVQ maintains consistently high and stable utilization across all layers, demonstrating more effective use of quantizer capacity. This indicates that our entropy-guided grouping not only preserves reconstruction accuracy but also maximizes representation efficiency under constrained bitrate.

To evaluate the impact of grouping numbers in the acoustic branch, we compared three configurations under a fixed total of 4 quantizers: $2groups \times 2quantizers, 4groups \times 1quantizer$, $1group \times 4quantizers$. As shown in Table 3, the 2×2 configuration consistently outperforms the 4×1 variant across all the metrics. The 2-group structure enables deeper quantization per group, leading to better reconstruction quality, particularly in perceptual similarity. These results indicate that, under fixed bitrate, grouping fewer but deeper quantizers is more effective than shallowly quantizing more groups. Compared to the 1×4 configuration, the 2×2 variant achieves significantly higher PESQ and STOI, with equal ViSQOL but slightly lower SDR. This suggests that simply stacking quantizers in a single group may cause inefficiencies in information allocation, leading to degraded perceptual quality. These results indicate that, under a fixed bitrate, grouping fewer but deeper quantizers is more effective than both shallow multi-group quantization and single-group quantization without structural separation.

3.6. Subjective Evaluation

In addition to objective metrics, we conducted the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) test to evaluate perceptual quality from a human listener's perspective. MUSHRA is widely adopted in audio coding and speech enhancement research. It provides a reliable subjective assessment by asking participants to rate the quality of multiple signal variants relative to a hidden reference and an explicitly degraded anchor.

We randomly selected eight speech samples, four male and four female, from the LibriTTS test-clean subset, with durations ranging from 1.4 to 13 seconds. These speech samples were recorded

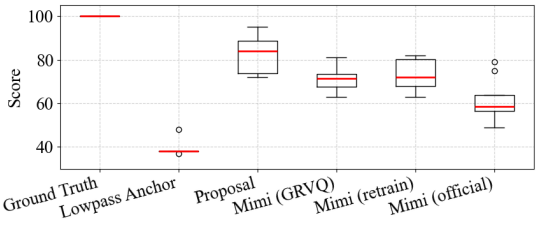

Fig. 4: MUSHRA score distributions of different methods.

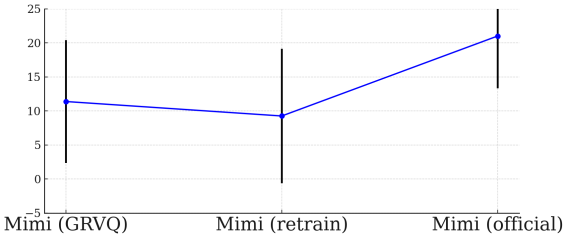

Fig. 5: Mean MUSHRA score differences between Proposal (EG-GRVQ) and baselines with 95% confidence intervals

at 24 kHz with 16-bit resolution for evaluation. The tested systems included an oracle clean reference, a low-pass anchor created by applying a 3.4 kHz low-pass filter to the reference, Mimi (GRVQ), Proposal (EG-GRVQ), and Mimi (retrain) and Mimi (official). The listening test was conducted with eight participants (five male and three female), all of whom had prior experience in perceptual audio evaluation.

As shown in Fig. 4, the proposed method significantly improves the MUSHRA score by 21 points over the official Mimi, and by 11 points over Mimi with GRVQ, indicating strong subjective preference. Fig. 5 further shows the mean differential MUSHRA scores with 95% confidence intervals. The results indicate that the proposed method achieves statistically significant improvements over the official baseline and Mimi (GRVQ), as the CI ranges do not overlap with zero. These findings highlight that our proposed method offers clear advantages in perceptual preference under low-bitrate conditions.

4. CONCLUSIONS

In this paper, we propose an entropy-guided group residual vector quantization framework, which demonstrates clear advantages for ultra-low bitrate neural speech coding. By leveraging channel variance as a proxy for entropy, our method ensures a more balanced allocation of information across quantization groups, thereby enhancing both reconstruction fidelity and codebook utilization rate. Objective evaluations show consistent improvements in PESQ and STOI while maintaining competitive ViSQOL and SDR scores, and NMSE analysis further confirms more accurate reconstruction relative to latent variance. Subjective MUSHRA tests also indicate a significant perceptual preference over baseline systems. Moreover, analysis of codebook usage reveals that the proposed design increased codebook utilization rate, leading to more efficient representation under constrained bitrate conditions. These results highlight the potential of EG-GRVQ for ultra-low bitrate neural speech codecs.

5. REFERENCES

[1] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[2] H. Yang, K. Zhen, S. Beack, and M. Kim, "Source-aware neural speech coding for noisy speech compression," in *Proc. IEEE ICASSP*, 2021, pp. 706–710.

[3] Y. Yang, F. Shen, C. Du, Z. Ma, K. Yu, D. Povey, and X. Chen, "Towards universal speech discrete tokens: A case study for ASR and TTS," in *Proc. IEEE ICASSP*, 2024, pp. 10 401–10 405.

[4] S. Chen, C. Wang, 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," *IEEE Trans. Audio, Speech, Lang. Process.*, 2025.

[5] Y.-C. Wu, I. D. Gebru, D. Marković, and A. Richard, "Audiodec: An open-source streaming high-fidelity neural audio codec," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[6] Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi, "SoundStorm: Efficient parallel audio generation," *arXiv preprint arXiv:2305.09636*, 2023.

[7] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi *et al.*, "AudioLM: a language modeling approach to audio generation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2523–2533, 2023.

[8] T. Brychcín and M. Konopík, "Semantic spaces for improving language modeling," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 192–209, 2014.

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 12 449–12 460, 2020.

[10] 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 J. Sele. Topics in Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

[11] H. Liu, X. Xu, Y. Yuan, M. Wu, W. Wang, and M. D. Plumbley, "SemantiCodec: An ultra low bitrate semantic audio codec for general sound," *IEEE J. Sele. Topics in Signal Process.*, 2024.

[12] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "SpeechTokenizer: Unified speech tokenizer for speech large language models," in *Proc. Int. Conf. Learn. Represent.*, 2024.

[13] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," *arXiv preprint arXiv:2410.00037*, 2024.

[14] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[15] X. Zhu, Y. Lv, Y. Lei, T. Li, W. He, H. Zhou, H. Lu, and L. Xie, "Vec-Tok speech: speech vectorization and tokenization for neural speech generation," *IEEE Trans. Audio, Speech, Lang. Process.*, 2025.

[16] Y. Wang, H. Zhan, L. Liu, R. Zeng, H. Guo, J. Zheng, Q. Zhang, X. Zhang, S. Zhang, and Z. Wu, "MaskGCT: Zero-shot text-to-speech with masked generative codec transformer," in *Proc. Int. Conf. Learn. Represent.*, 2025.

[17] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities," in *Proc. Conf. Empirical Methods in Natural Language Processing*, 2023.

[18] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 495–507, 2021.

[19] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[20] J. D. Parker, A. Smirnov, J. Pons, C. Carr, Z. Zukowski, Z. Evans, and X. Liu, "Scaling transformers for low-bitrate high-quality speech coding," in *Proc. Int. Conf. Learn. Represent.*, 2025.

[21] D. Yang, S. Liu, R. Huang, J. Tian, C. Weng, and Y. Zou, "Hifi-codec: Group-residual vector quantization for high fidelity audio codec," *arXiv preprint arXiv:2305.02765*, 2023.

[22] Y. Shiraki and M. Honda, "LPC speech coding based on variable-length segment quantization," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 9, pp. 1437–1444, 1988.

[23] S. Ji, Z. Jiang, W. Wang, Y. Chen, M. Fang, J. Zuo, Q. Yang, X. Cheng, Z. Wang, R. Li, Z. Zhang, X. Yang, R. Huang, Y. Jiang, Q. Chen, S. Zheng, and Z. Zhao, "WavTokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling," in *Proc. Int. Conf. Learn. Represent.*, 2025.

[24] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[25] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, pp. 271–350, 2019.

[26] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[27] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE ICASSP*, vol. 2, 2001, pp. 749–752.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE ICASSP*, 2010, pp. 4214–4217.

[29] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "ViSQOL: an objective speech quality model," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 1, p. 13, 2015.

[30] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *Proc. IEEE ICASSP*, 2019, pp. 626–630.