Interspeech 2024
1-5 September 2024, Kos, Greece

Low Bitrate High-Quality RVQGAN-based Discrete Speech Tokenizer

Slava Shechtman, Avihu Dekel

IBM Research

slava@il.ibm.com, avihu.dekel@ibm.com

Abstract

Discrete Audio codecs (or audio tokenizers) have recently regained interest due to the ability of Large Language Models (LLMs) to learn their compressed acoustic representations. Various publicly available trainable discrete tokenizers recently demonstrated impressive results for audio tokenization, yet they mostly require high token rates to gain high-quality reconstruction. In this study, we fine-tuned an open-source general audio RVQGAN model using diverse open-source speech data, considering various recording conditions and quality levels. The resulting wideband (24kHz) speech-only model achieves speech reconstruction, which is nearly indistinguishable from PCM (pulse-code modulation) with a rate of 150-300 tokens per second (1500-3000 bps). The evaluation used comprehensive English speech data encompassing different recording conditions, including studio settings. Speech samples and details on the reproducible trained models are made publicly available[1].
Index Terms: Speech Coding, Discrete Speech Tokenization

1. Introduction

Digital speech and audio codecs were originally used solely for efficient audio compression for data transmission [1, 2]. In such settings, the encoder transforms the digitally sampled audio signal (with a sampling rate that can vary from Narrow Band 8kHz for telephony to Full Band 48-96kHz for high-fidelity studio recordings) into a compressed digital stream that is subsequently transmitted. Before encoding, audio signals typically undergo a two-step process. First, *segmentation* divides the audio data into smaller units known as frames. Next, *feature extraction* represents these frames in a manner that enables efficient lossy signal compression while minimizing perceptual degradation for human listeners. Certain features, such as Spectrograms or Mel-Frequency Cepstral Coefficients, characterize the spectral envelope of the signal frame, while the remaining ones describe the time-domain residual signal. In classic codecs, the spectral features usually undergo multiple trainable Vector Quantizations (VQ), while the rest are represented with a mixture of trainable and rule-based codebooks along with binary representations of various control parameters [1, 2]. Once transmitted, the receiver employs the decoder to reconstruct the original audio from the received digital stream.

Discrete audio tokenization, a special case of digital audio coding, has recently gained renewed interest due to its potential for applying Large Language Modeling (LLM) techniques in audio and mixed text/audio domains for one-shot speech synthesis [3, 4, 5], speech recognition [6], speaker recognition [7] and more. Unlike classic audio coding, which imposes no constraints on the digital-encoder output, discrete audio tokenization [8] (see Figure 1) involves converting an audio signal into a sequence of discrete tokens, represented by integers serving as codeword indices associated with a set of trained codebooks. Audio tokens are designed to capture essential details for precise audio reconstruction using a trainable decoder. This enables the audio tokens generated by a language model to be readily converted into high-quality audio waveforms.

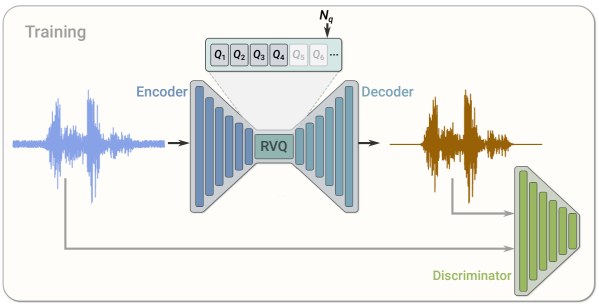

Figure 1: *VQGAN architecture overview [8]*

Discrete tokenizers usually feature a classic bottleneck autoencoder structure with a quantization layer [9]. The Residual Vector Quantization with Generative Adversary Network (RVQGAN) based architecture [8] stands out as one of the most popular choices for tokenization. It deploys residual vector quantization (RVQ) [10] of the bottleneck features, where each quantization layer refines the previous quantization layer, resulting in multiple tokens representing a single audio frame. Several discrete tokenizers, suitable for speech or general audio, have recently become available and gained popularity in the open-source community [11, 12, 13]. They are mostly RVQGAN-based and their high-quality operating points start from 600 tokens per second [14], a demanding high data rate to efficiently train LLM models. To improve the discrete generative modeling of audio, it is highly beneficial to compress further the sequence of tokens representing an audio sample, while preserving the high fidelity of the audio reconstruction.

We base this work on the Descript Audio Codec (DAC) [12], an open-source universal discrete tokenizer model capable of preserving high-fidelity audio quality across diverse audio material (including general audio, music, and speech) at bit rates of 6-8 kbps. However, its performance dramatically deteriorates for bitrates of 3 kbps and below [12].

The goal of the current work is to adapt the universal audio DAC model to a high-quality speech-only discrete tokenizer model with reduced operational bitrates. Our contributions are summed up as follows:

[1] ibm.biz/IS24SpeechRVQ

10.21437/Interspeech.2024-2366

- We fine-tune a universal DAC model in low-bitrate settings (1500-3000 bps) with diverse open-source speech data, while carefully balancing various recording conditions and audio quality levels, and focusing on high-fidelity speech data.
- We evaluate the resulting models on various speech datasets, demonstrating high-quality reconstruction for the 1.5-kbps model, and perceptually transparent reconstruction for the 3-kbps model.
- We perform a thorough ablation study to explore how training speech data of various quality levels and recording conditions influences the model performance, as assessed on versatile test data.

2. Method

2.1. RVQ-GAN Model

Residual Vector Quantization (RVQ) is a classic speech-coding technique [15] that has been recently revived as a key element in modern neural discrete tokenization when combined with Generative Adversarial Network (GAN) [16] training techniques [8]. An overview of the ensuing RVQGAN model is presented in Figure 1. In this autoencoder architecture, an encoder downsamples an input signal, creating a more compact latent representation that is incrementally quantized by RVQ, and then reconstructed by a decoder whose structure mirrors that of the encoder. RVQ is a multi-stage VQ technique, where each stage quantizes the residual from the previous VQ stage. During training, the model parameters are optimized using a combination of reconstruction and adversarial losses, where a separate discriminator network is trained concurrently with the autoencoder network [8]. The non-differential quantization layers are optimized using the straight-through estimator [17].

The Descript Audio Codec (DAC) [12] is an RVQGAN-based universal audio codec model that is remarkable in its capability to preserve high-fidelity audio quality across diverse audio material, including general audio, music, and speech [12]. It achieves this by incorporating several techniques, such as periodic activation functions, codebook factorization with L2-normalization, and improved reconstruction and adversarial losses [12]. It also deploys random quantizer dropout to support multiple bitrates for a single model and stabilize the training [8]. Both pretrained DAC models and model source code are released as open-source[2]. They also have been shown to outperform several former popular discrete tokenization models [8, 11].

2.2. Training Data Selection

The quality of discrete audio tokenizers depends not only on their architecture and training protocols but also on the quality of the training data, and among those datasets commonly used for training a discrete tokenizer [14], we observe data of variable speech quality. In general, trainable tokenizers are not language-dependent provided multi-lingual data is used for training. However, in this work, we only focus on English, so we refer below just to English datasets (or English sections of multi-lingual ones), serving for DAC training [12]:

- *ReadSpeech* is a high-quality full-band (48kHz) dataset from the Denoising Challenge [18], containing about 1000 hrs. It is mostly derived from audiobooks with good recording conditions.

[2] https://github.com/descriptinc/descript-audio-codec

Figure 2: *MUSHRA results with 95% confidence interval*

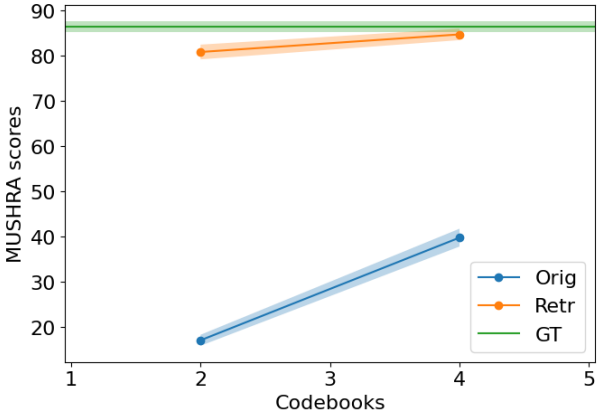

Table 1: *Wilcoxon Rank Sum Test comparing systems to the raw audio*

System	Orig-2	Orig-4	Retr-2	Retr-4
p-value	$<1e{-}5$	$<1e{-}5$	$<1e{-}5$	0.062

- *DAPS* [19] is a small (4.5hrs) dataset of studio quality full band (48khz) recordings.
- *LibriSpeech* [20] is a large (around 1000 hrs) 16kHz-sampled dataset of medium quality that is also based on public-domain audiobooks.
- *LibriTTS* [21] (585 hrs) is also derived from audiobooks, and it is a source of full band (24kHz) speech of medium quality containing a *LibriTTS-clean* subset with better recording conditions, and a *LibriTTS-other* subset with heavily accented speech and/or more challenging recording setup (including noise and reverberation). It is derived from the same material as the LibriSpeech dataset.
- *VCTK* [22] is a medium quality full-band (48kHz) speech dataset (44hrs, 110 speakers).
- *Common Voice* [23] (2500 hrs) is a crowd-collected dataset of speech sampled at 8-16kHz, featuring the most challenging recording conditions and distortions.

For the task of the DAC model adaptation for speech-only settings (see Section 2.3), we added extra sources of speech data:

- *LJ-Speech* [24], a 24-hrs long full-band single-speaker dataset of studio quality.
- *LibriTTS-R* [25], a high fidelity version of *LibriTTS* that is created by passing the whole *LibriTTS* dataset through a Speech Restoration network, simulating clean studio recording conditions.
- *LibriLight* [26] very large (60k hours) mixed-quality public domain dataset of 16khz speech. In our retraining, We used this dataset as an unqualified data source, instead of *Common Voice*.

The training data does not equally represent the various quality levels. To address this, the original training procedure employs balanced data sampling, ensuring an equal mix of datasets from different sources and quality levels within each

training mini-batch [12]. We adopted this approach for the model adaptation, too, and divided our training data into the following categories, which are equally balanced over the training mini-batches:

- *HQ1*: high-quality, clean, contains *ReadSpeech*, *DAPS*, and *LJ-speech* datasets.
- *HQ2*: restored high-quality, clean, contains the *LibriTTS-R-clean* portion of *LibriTTS-R*.
- *HQ3*: restored high-quality, clean, contains the *LibriTTS-R-other* portion of *LibriTTS-R*, where more challenging accentuation is present.
- *MQ1*: medium-quality, clean, contains the *LibriTTS-clean* portion of *LibriTTS*.
- *MQ2*: medium-quality, unclean, contains the *LibriTTS-other* portion of *LibriTTS*.
- *UQ*: unqualified (low-quality/mixed quality), unclean, contains *LibriLight* dataset, upsampled to 24kHz

One can notice that in the proposed setup the low-quality data is under-represented. We found these settings beneficial for high-quality speech reconstruction. Additional data selection trends are explored in Sec 3.3

2.3. Implementation Details

We retrained a 24-kHz universal audio DAC model by strictly following the training procedure proposed in [12] with the balanced training data as detailed above. Unlike the original DAC training, we omitted quantizer dropout (the procedure of random dropping out of some of the later stages in RVQ, during the training [8]). While the quantizer dropout proved beneficial when training from scratch [12], we observed that it had a detrimental effect on model performance when adapting from a pretrained model.

We trained a set of fixed bitrate models (with no quantization dropout) each utilizing a different number of 10-bit RVQ codebooks $Q \in \{1, 2, 4, 8, 16, 32\}$ (corresponding to bitrates $R \in \{0.375, 0.75, 3, 6, 12, 24\}$ kbps), initialized from the publicly available 24kbps DAC model for 24kHz audio [12]. The original model encodes an audio frame with 1-32 RVQ codebooks of 10 bits each at a frame rate of 75Hz. Training took place on two *A100_80g* GPUs for 400k steps with a mini-batch of 72 excerpts of a fixed length of 0.38 sec (randomly extracted from longer speech samples in the datasets).

2.4. Evaluation metrics

We make use of the following metrics for objective reconstruction evaluations:

- *mel loss*: a combined mel-scale loss, serving as a mel reconstruction loss during DAC training [12]. It is evaluated as a sum of $L1$-distances (between the ground truth and the reconstructed) of log mel spectrograms of various spectral resolutions.
- *STFT loss*: a combined Short-Time Fourier Transform (STFT) loss, serving as a linear frequency-domain loss during DAC training [12]. It is evaluated as a sum of $L1$-distances (between the ground truth and reconstructions) of linear spectrograms at various spectral resolutions.
- *PESQ*: Wideband (16kHz) speech quality assessment score [27].
- *STOI*: speech intelligibility metric [28].

While those metrics serve the purpose of tracking trends and model comparisons, they do not reliably predict perceptually significant distortion. Those metrics are therefore complemented with subjective listening evaluations (see Sec 3.2).

Table 2: *Objective metrics for Retrained speech model (*Retr*) vs. Original model (*Orig*) as assessed on various test sets*

test data	metric	system	number of codebooks					
			32	16	8	4	2	1
studio	mel↓	Orig	0.22	0.29	0.38	0.48	0.58	0.69
		Retr	**0.15**	**0.22**	**0.3**	**0.37**	**0.45**	**0.57**
	STFT↓	Orig	0.47	0.55	0.64	0.75	0.9	1.02
		Retr	**0.42**	**0.49**	**0.53**	**0.59**	**0.67**	**0.79**
	PESQ↑	Orig	4.5	4.3	3.69	2.6	1.67	1.24
		Retr	**4.54**	**4.43**	**4.11**	**3.57**	**2.78**	**1.91**
	STOI↑	Orig	**1**	**0.99**	0.97	0.93	0.87	0.79
		Retr	**1**	**0.99**	**0.98**	**0.97**	**0.95**	**0.89**
DAPS	mel↓	Orig	0.56	0.75	0.98	1.22	1.47	1.79
		Retr	**0.41**	**0.61**	**0.82**	**1.00**	**1.20**	**1.45**
	STFT↓	Orig	1.15	1.31	1.51	1.71	1.91	2.18
		Retr	**1.12**	**1.28**	**1.43**	**1.56**	**1.72**	**1.92**
	PESQ↑	Orig	4.50	4.32	3.65	2.63	1.74	1.28
		Retr	**4.52**	**4.40**	**4.06**	**3.52**	**2.76**	**1.98**
	STOI↑	Orig	1.00	0.98	0.95	0.91	0.86	0.78
		Retr	**1.00**	**0.99**	**0.97**	**0.95**	**0.92**	**0.87**
Libri-TTS-R-clean	mel ↓	Orig	0.25	0.33	0.43	0.53	0.65	0.79
		Retr	**0.17**	**0.25**	**0.32**	**0.39**	**0.47**	**0.58**
	STFT↓	Orig	0.5	0.57	0.66	0.75	0.85	0.96
		Retr	**0.44**	**0.5**	**0.55**	**0.59**	**0.65**	**0.74**
	PESQ↑	Orig	4.5	4.3	3.67	2.62	1.75	1.25
		Retr	**4.54**	**4.44**	**4.15**	**3.74**	**3.06**	**2.17**
	STOI↑	Orig	**1**	**0.99**	0.97	0.93	0.88	0.81
		Retr	**1**	**0.99**	**0.98**	**0.97**	**0.95**	**0.91**
Libri-TTS-R-other	mel ↓	Retr	**0.13**	**0.19**	**0.25**	**0.30**	**0.35**	**0.44**
	STFT↓	Orig	0.38	0.44	0.50	0.56	0.63	0.72
		Retr	**0.34**	**0.39**	**0.42**	**0.45**	**0.50**	**0.56**
	PESQ↑	Orig	4.50	4.31	3.68	2.69	1.83	1.30
		Retr	**4.55**	**4.44**	**4.15**	**3.73**	**3.06**	**2.19**
	STOI↑	Orig	0.99	0.98	0.96	0.93	0.88	0.80
		Retr	**1.00**	**0.99**	**0.98**	**0.97**	**0.95**	**0.90**
Libri-TTS-clean		Orig	0.27	0.35	0.45	0.55	0.66	0.78
	mel ↓	Orig	0.27	0.35	0.45	0.55	0.66	0.78
		Retr	**0.19**	**0.28**	**0.37**	**0.45**	**0.53**	**0.64**
	STFT↓	Orig	0.53	0.61	0.7	0.79	0.88	0.99
		Retr	**0.5**	**0.56**	**0.63**	**0.68**	**0.74**	**0.83**
	PESQ↑	Orig	**4.46**	**4.25**	3.58	2.52	1.64	1.23
		Retr	**4.51**	**4.38**	**4**	**3.46**	**2.67**	**1.88**
	STOI↑	Orig	**0.99**	**0.98**	0.96	0.92	0.86	0.78
		Retr	**0.99**	**0.99**	**0.98**	**0.96**	**0.93**	**0.87**
Libri-TTS-other	mel ↓	Orig	0.21	0.28	0.35	0.43	0.51	0.62
		Retr	**0.15**	**0.22**	**0.3**	**0.36**	**0.43**	**0.52**
	STFT↓	Orig	0.42	0.49	0.56	0.62	0.7	0.79
		Retr	**0.39**	**0.45**	**0.5**	**0.54**	**0.6**	**0.67**
	PESQ↑	Orig	**4.42**	**4.14**	3.39	2.39	1.61	1.22
		Retr	**4.48**	**4.3**	**3.81**	**3.16**	**2.4**	**1.73**
	STOI↑	Orig	**0.99**	**0.97**	0.94	0.9	0.83	0.75
		Retr	**0.99**	**0.98**	**0.96**	**0.94**	**0.9**	**0.84**

2.5. Test Datasets

Both in the final objective evaluation and the ablation studies we assessed the objective metrics on the following test dataset:

- *Studio*: a proprietary set of 1024×2 studio-quality samples for male and female speakers, sampled at 22.05kHz.
- *DAPS*: a held-out set of 128 samples from the full-band high-fidelity *DAPS* dataset
- *LibriTTS-R-clean*: a random set of 1024 samples from the held-out set *LibriTTS-R-test-clean*, with unseen speakers.
- *LibriTTS-R-other*: a random set of 1024 samples from the held-out set *LibriTTS-R-test-other*, with unseen speakers.
- *LibriTTS-clean*: a random set of 1024 samples from the held-out set *LibriTTS-test-other*, with unseen speakers.

Table 3: *Removing high-quality (HQ) and mid/unknown-quality (MQ/UQ) data sources, and measuring objective metrics various speech test sets*

Test Data		full	Reduce HQ Data			Reduce MQ/UQ Data				
LJ-speech		✓	✗	✓	✗	✓	✓	✓	✓	✓
Libri-TTS-R-clean		✓	✓	✗	✗	✓	✓	✓	✓	✓
Libri-TTS-R-others		✓	✓	✗	✗	✓	✓	✓	✗	✗
Libri-TTS-clean		✓	✓	✓	✓	✓	✓	✗	✓	✗
Libri-TTS-others		✓	✓	✓	✓	✓	✗	✗	✗	✗
UQ		✓	✓	✓	✓	✗	✗	✗	✗	✗
Studio	mel loss↓	**0.37**	0.38	0.38	0.38	0.38	0.38	0.40	0.38	0.40
	STFT loss↓	**0.59**	0.62	0.59	0.64	0.62	0.62	0.71	0.60	0.69
	PESQ↑	3.57	**3.58**	3.48	3.53	3.51	3.50	3.43	3.51	3.42
DAPS	mel loss↓	**1.00**	1.02	1.02	1.08	1.02	1.03	1.02	1.01	1.02
	STFT loss↓	1.56	1.56	**1.55**	1.71	1.57	1.58	1.57	1.57	1.56
	PESQ↑	**3.52**	3.48	3.44	3.38	3.46	3.47	3.46	3.49	3.47
Libri-TTS-R-clean	mel loss↓	**0.39**	0.40	0.43	0.48	0.40	0.40	0.39	0.39	0.39
	STFT loss↓	**0.59**	0.60	0.65	0.74	0.61	0.60	0.60	0.60	0.60
	PESQ↑	3.74	3.69	3.47	3.29	3.69	3.71	**3.75**	3.70	3.72
Libri-TTS-R-other	mel loss↓	**0.30**	0.30	0.33	0.36	0.30	0.30	0.30	0.30	0.30
	STFT loss↓	**0.45**	0.46	0.49	0.55	0.46	0.46	0.46	0.46	0.46
	PESQ↑	**3.73**	3.68	3.47	3.31	3.69	3.70	3.74	3.69	3.71
Libri-TTS-clean	mel loss↓	**0.45**	0.45	0.46	0.47	0.46	0.46	0.47	0.45	0.47
	STFT loss↓	**0.68**	0.68	0.69	0.71	0.69	0.69	0.71	0.68	0.71
	PESQ↑	**3.46**	3.40	3.39	3.36	3.40	3.39	3.32	3.40	3.34
Libri-TTS-other	mel loss↓	**0.36**	0.37	0.36	0.38	0.37	0.37	0.38	0.37	0.38
	STFT loss↓	**0.54**	0.55	0.55	0.58	0.55	0.56	0.57	0.55	0.58
	PESQ↑	**3.16**	3.12	3.12	3.06	3.09	3.06	3.00	3.09	3.02

- *LibriTTS-other*: a random set of 1024 samples from the held-out set *LibriTTS-test-other*, with unseen speakers.

3. Results

3.1. Objective metrics

The objective metrics for the retrained models are presented in Table 2. As one can notice, all the objective metrics consistently improve over all the test sets, and the improvement becomes more perceptually significant for smaller quantization codebooks. The original high-rate models were found virtually transparent to the recordings when perceptually assessed, so their improvement was not perceptually significant. However, when the number of codebooks is reduced, the deterioration of the original model becomes apparent and the improvement becomes more perceptually salient.

3.2. Subjective evaluations

We selected the 4-codebook (3 kbps) and 2-codebook (1.5 kbps) models for further subjective listening test evaluation. 16 subjects participated in the MUSHRA [29] test, assessing the 4 systems (2- and 4-codebook systems, original, and retrained) and a hidden PCM reference signal. 30 test stimuli were randomly selected from *LibriTTS-r-clean*, *LibriTTS-r-other*, *LibriTTS-clean*, *LibriTTS-other*, *DAPS*, and *Studio* test sets, 6 stimuli for each set. The MUSHRA average scores with 95% confidence intervals are presented in Figure 2. Statistical significance of the difference from PCM for each system was assessed via the Wilcoxon Rank Sum test [30], revealing that the perceptual difference between the retrained 4-codebook system outputs and the original recordings is statistically insignificant (Table 1)[3].

[3]Sample page is available here: `ibm.biz/IS24SpeechRVQ`

3.3. Data Ablation Studies

In a series of training-data ablation studies, we investigated the impact of excluding portions of the training data on the quality of speech reconstruction, as assessed on six held-out test datasets of various quality levels, as described in Section 2.5. The results for the retrained 3-kbps model after 200k training steps (with a mini-batch of size $B = 72$) are presented in Table 3. We omitted STOI metric in the ablation table, as it had almost identical values in all the columns. The studies revealed several interesting observations. One can notice that, in general, medium to low-quality data is important for the reconstruction of most of the datasets, including the high-fidelity data (*Studio*, *DAPS*), although for some test sets (*LibriTTS-R-clean*, *LibriTTS-R-other*) it is not the case. We also observed that eliminating the high-quality *LibriTTS-R* negatively impacted all scores, including those of its corresponding medium-quality counterpart. On the other hand, the high-quality *LibriTTS-R* test set did not seem to benefit from the presence of medium-quality *LibriTTS* training data.

4. Summary

In this paper we (i) presented an improved version of the RVQ-GAN audio codec, specialized to speech-only data; (ii) showcased the importance of balanced speech data for indistinguishable reconstruction quality; and (iii) provided an ablation study showing the significant impact of data selection for training. Our pretrained model can be useful for speech synthesis, speech continuation, and various other tasks due to its lower bitrate and superior quality. While the current model underwent training and testing exclusively in English, we intend to extend its applicability to achieve consistent performance across multiple languages.

5. References

[1] K. Vos, K. V. Sørensen, S. S. Jensen, and J.-M. Valin, "Voice coding with opus," in *Audio Engineering Society Convention 135*. Audio Engineering Society, 2013.

[2] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, L. Miao, Z. Wang, L. Laaksonen, A. Vasilache *et al.*, "Overview of the evs codec architecture," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5698–5702.

[3] 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.

[4] Z. Borsos *et al.*, "AudioLM: a language modeling approach to audio generation," 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2209.03143

[5] ——, "SoundStorm: Efficient parallel audio generation," 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.09636

[6] T. Wang, L. Zhou, Z. Zhang, Y. Wu, S. Liu, Y. Gaur, Z. Chen, J. Li, and F. Wei, "Viola: Unified codec language models for speech recognition, synthesis, and translation," *arXiv preprint arXiv:2305.16107*, 2023.

[7] K. C. Puvvada, N. R. Koluguri, K. Dhawan, J. Balam, and B. Ginsburg, "Discrete audio representation as an alternative to mel-spectrograms for speaker and speech recognition," *arXiv preprint arXiv:2309.10922*, 2023.

[8] 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*, vol. 30, pp. 495–507, 2021.

[9] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6309–6318.

[10] C. F. Barnes, S. A. Rizvi, and N. M. Nasrabadi, "Advances in residual vector quantization: A review," *IEEE transactions on image processing*, vol. 5, no. 2, pp. 226–262, 1996.

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

[12] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[13] Z. Du, S. Zhang, K. Hu, and S. Zheng, "Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec," *arXiv preprint arXiv:2309.07405*, 2023.

[14] H. Wu, X. Chen, Y.-C. Lin, K.-w. Chang, H.-L. Chung, A. H. Liu, and H.-y. Lee, "Towards audio language modeling-an overview," *arXiv preprint arXiv:2402.13236*, 2024.

[15] A. Vasuki and P. Vanathi, "A review of vector quantization techniques," *IEEE Potentials*, vol. 25, no. 4, pp. 39–47, 2006.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[17] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[18] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," 2020.

[19] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006–1010, 2014.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[21] 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," *Interspeech 2019*, 2019.

[22] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2019.

[23] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[24] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[25] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "Libritts-r: A restored multi-speaker text-to-speech corpus," *Interspeech 2023*.

[26] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673, https://github.com/facebookresearch/libri-light.

[27] I. Union, "Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union, Recommendation P*, vol. 862, 2007.

[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 *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[29] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

[30] F. Wilcoxon, "Individual comparisons by ranking methods. biom. bull., 1, 80–83," 1945.