DM-CODEC: DISTILLING MULTIMODAL REPRESENTATIONS FOR SPEECH TOKENIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advancements in speech-language models have yielded significant improvements in speech tokenization and synthesis. However, effectively mapping the complex, multidimensional attributes of speech into discrete tokens remains challenging. This process demands acoustic, semantic, and contextual information for precise speech representations. Existing speech representations generally fall into two categories: acoustic tokens from audio codecs and semantic tokens from speech self-supervised learning models. Although recent efforts have unified acoustic and semantic tokens for improved performance, they overlook the crucial role of contextual representation in comprehensive speech modeling. Our empirical investigations reveal that the absence of contextual representations results in elevated Word Error Rate (WER) and Word Information Lost (WIL) scores in speech transcriptions. To address these limitations, we propose two novel distillation approaches: (1) a language model (LM)-guided distillation method that incorporates contextual information, and (2) a combined LM and self-supervised speech model (SM)-guided distillation technique that effectively distills multimodal representations (acoustic, semantic, and contextual) into a comprehensive speech tokenizer, termed DM-Codec. The DM-Codec architecture adopts a streamlined encoder-decoder framework with a Residual Vector Quantizer (RVQ) and incorporates the LM and SM during the training process. Experiments show DM-Codec significantly outperforms state-of-the-art speech tokenization models, reducing WER by up to 13.46%, WIL by 9.82%, and improving speech quality by 5.84% and intelligibility by 1.85% on the LibriSpeech benchmark dataset.

1 INTRODUCTION

In recent years, the advent of Large Language Models (LLMs) has revolutionized various domains, offering unprecedented advancements across a wide array of tasks (OpenAI, 2024). A critical component of this success has been the tokenization of input data, enabling vast amounts of information processing (Du et al., 2024; Rust et al., 2021). Inspired by these breakthroughs, significant attention has shifted towards replicating similar successes in the realm of speech understanding and generation (Défossez et al., 2022; Hsu et al., 2021). However, tokenizing speech into discrete units presents unique challenges compared to text, as speech is inherently continuous and multidimensional, requiring various speech attributes such as acoustic properties, semantic meaning, and contextual clues (Ju et al., 2024). Traditional approaches using feature representations such as Mel-Spectrograms (Sheng et al., 2019), Mel-frequency cepstral coefficients (MFCCs) (Juvela et al., 2018), and Waveforms (Kim et al., 2021) have proven inadequate in capturing this full spectrum of information, resulting in suboptimal performance in downstream tasks such as speech synthesis (Ju et al., 2024).

These limitations led researchers to explore various approaches, and one prominent direction leading to audio codecs (Borsos et al., 2023). Notable examples include SoundStream (Zeghidour et al., 2021) and EnCodec (Défossez et al., 2022), which utilize Residual Vector Quantizers (RVQ) within a neural codec framework, iteratively refining quantized vectors to discretize speech into acoustic tokens. Concurrently, self-supervised speech representation learning models such as HuBERT (Hsu et al., 2021) and wav2vec 2.0 (Baevski et al., 2020) facilitated extracting speech representations as semantic tokens (Borsos et al., 2023). Efforts to unify acoustic and semantic representations have led to two notable approaches: SpeechTokenizer (Zhang et al., 2024a), which utilizes semantic dis-

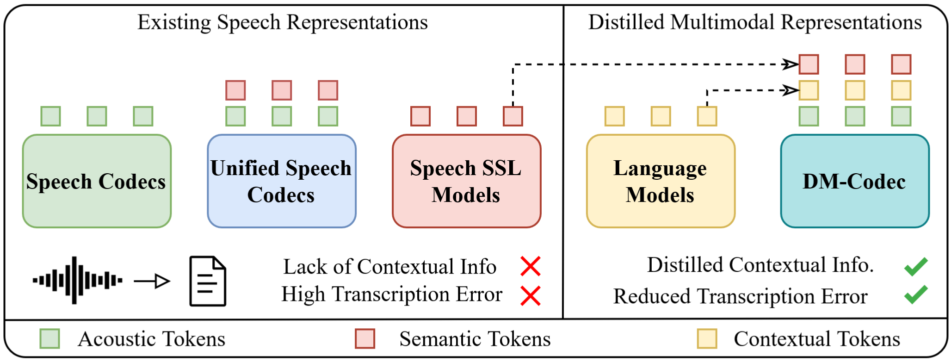

Figure 1: An overview of speech tokenization approaches using discrete acoustic, semantic, and contextual tokens. DM-Codec integrates these multimodal representations for robust speech tokenization, learning comprehensive speech representations.

tillation from HuBERT, and FACodec (Ju et al., 2024), which proposes a factorized vector quantizer to disentangle speech representation into different subspaces using separate RVQs with supervision.

While these approaches have shown promising results, they often overlook a crucial aspect of speech representation: the integration of contextual language information. Language models (LMs) have demonstrated a remarkable ability to learn contextual representations that capture the meaning of tokens based on their broader linguistic context (Devlin et al., 2019). These contextual representations can provide essential insights into speech representation, allowing for a more nuanced understanding of words in varying linguistic contexts. Our empirical investigations also reveal that existing discrete speech representation models struggle to align reconstructed speech with accurate textual form, resulting in elevated Word Error Rates (WER) and Word Information Lost (WIL) scores in speech transcription tasks. This observation underscores the need for a more comprehensive approach to speech tokenization that incorporates contextual language information.

To address these challenges, we propose DM-Codec, a novel speech tokenizer that unifies multimodal language and speech representations within a comprehensive tokenizer for speech. Our approach builds on a neural codec architecture incorporating RVQ with encoder, decoder, and discriminator components. Central to our innovation is the introduction of an LM-guided distillation method that effectively incorporates contextual representations into the speech tokenization process. This technique allows DM-Codec capturing the nuances of linguistic context often missed by existing models. Building upon the LM-guided approach, we further propose a hybrid distillation method combining both LM and speech model (SM)-guided techniques. To the best of our knowledge, we are the first to attempt to integrate all three essential aspects of speech representation—acoustic, semantic, and contextual—within a single codec. See Figure 1 for a depiction.

Through extensive experimentation on the LibriSpeech benchmark dataset (Panayotov et al., 2015), we demonstrate the superiority of DM-Codec, which achieves significantly lower WER and WIL compared to state-of-the-art baseline speech tokenizers. Specifically, DM-Codec achieves a WER of 4.05 and a WIL of 6.61, outperforming SpeechTokenizer (4.49, 7.10), FACodec (4.68, 7.33), and EnCodec (4.53, 7.17). Furthermore, DM-Codec exhibits improved speech quality, as evidenced by its Virtual Speech Quality Objective Listener (ViSQOL) score of 3.26, surpassing the performance of baseline models (EnCodec: 3.08; SpeechTokenizer: 3.09; FACodec: 3.13).

Our research makes the following key contributions:

- We introduce DM-Codec, a novel speech tokenizer that incorporates contextual representations via an LM-guided distillation method.
- We present a novel combined LM and SM-guided representation distillation approach, uniting acoustic, semantic, and contextual representations into a unified framework.
- Through comprehensive experiments and ablation studies, we demonstrate the effectiveness of DM-Codec in preserving increased contextual information and enhancing the retention of acoustic and speech information in reconstructed speech.

2 PROPOSED METHOD

In this section, we present DM-Codec, a novel speech tokenizer designed to encapsulate a comprehensive fusion of multimodal (acoustic, semantic, and contextual) representations. As illustrated in Figure 2, we propose two distinct training approaches to incorporate these representations: (i) a language model (LM)-guided distillation method, and (ii) a combined LM and self-supervised speech model (SM)-guided distillation method. The first approach distills contextual representations from the LM and integrates them with learned acoustic representations. The second approach combines SM and LM to further incorporate semantic representations with contextual and acoustic representations. It ensures that DM-Codec captures the essential elements of speech by harmonizing the acoustic features with contextual and semantic information. The following subsections detail our proposed distillation methods (§2.1), model details (§2.2), and components (Technical Appendix).

2.1 SPEECH AND LANGUAGE MODEL GUIDED DISTILLATION

Our approach first transcribes the raw speech $\mathbf{x}$ into its corresponding text $\mathbf{x}'$ using a Speech-to-Text (STT) model M_{STT}, such that $\mathbf{x}' = M_{STT}(\mathbf{x})$. For simplicity, we omit any post-processing techniques on the $\mathbf{x}'$. Subsequently, we pass the text $\mathbf{x}'$ through a pretrained language model M_{LM} to obtain contextual representations of $\mathbf{x}'$, tokenized into a set of tokens, $\mathcal{T} = \{t_i\}_{i=1}^{n}$. For each token t_i, we extract its corresponding layer-wise hidden representations $\{\mathbf{h}_i^l\}_{l=1}^{L}$, where L denotes the total number of layers in M_{LM}. We utilize all layer representations to derive the representations for each token, as each layer of a pre-trained language model captures hierarchical and contextually distinct information (Niu et al., 2022; Kovaleva et al., 2019; Hao et al., 2019). To obtain the contextual representation $\mathbf{S}_i$ for token t_i, we average the hidden representations across all layers, yielding $\mathbf{S}_i = \frac{1}{L}\sum_{l=1}^{L} \mathbf{h}_i^l$, where $\mathbf{S}_i \in \mathbb{R}^D$ where D is hidden dimension. Consequently, we obtain the contextual representations $\mathbf{S} = \{\mathbf{S}_i\}_{i=1}^{n}$ for the speech input $\mathbf{x}$, which captures the contextually diverse information from M_{LM}.

Simultaneously, we process the raw speech $\mathbf{x}$ through an Encoder $\mathbf{E}(\mathbf{x})$ to obtain the latent feature $\mathbf{v}$. We then pass $\mathbf{v}$ through a Residual Vector Quantizer (RVQ) to obtain quantized features $\mathbf{Q} = \{\mathbf{Q}_k\}_{k=1}^{K}$, where K represents the number of quantization layers in the RVQ, and $\mathbf{Q}_k \in \mathbb{R}^{D'}$ where D' is hidden dimension of k^{th} RVQ layer. These quantized features are subsequently used to reconstruct the audio $\hat{\mathbf{x}}$ via a decoder. To align the quantized feature $\mathbf{Q}_k$ with the LM distilled features $\mathbf{S}_i$, we apply a linear transformation $\mathbf{Q}'_k = \mathbf{W}\mathbf{Q}_k$, where $\mathbf{W} \in \mathbb{R}^{D' \times D}$, ensuring the dimensional consistency for the distillation process.

LM Guided Distillation: In this approach, we distil the LM representations $\mathbf{S}$. To calculate the LM-guided distillation loss, we adopt a *continuous representation distillation* technique, similar to the one employed by SpeechTokenizer (Zhang et al., 2024a), which maximizes the cosine similarity at the dimension level across all time steps. In our case, we calculate the continuous representation distillation of the transformed quantized features $\mathbf{Q}'_k$ and the LM representation features $\mathbf{S}$ as follows:

$$\mathcal{L}_L = -\frac{1}{D}\sum_{d=1}^{D} \log\left(\sigma\left(\frac{\mathbf{Q}_k'^{(:,d)} \cdot \mathbf{S}^{(:,d)}}{\|\mathbf{Q}_k'^{(:,d)}\|\|\mathbf{S}^{(:,d)}\|}\right)\right) \tag{1}$$

Here, the notation $(:, d)$ indicates a vector that includes values from all time steps at the d^{th} dimension. The function $\sigma(\cdot)$ represents the sigmoid activation function, commonly used to squash input values into a range between 0 and 1.

Combined LM and SM Guided Distillation: To further enhance the capabilities of DM-Codec, we propose a hybrid approach that utilizes both audio and text modalities. To derive semantic representations from the speech model (SM), we adopt a similar distillation strategy as we used for the LM. We first pass the raw speech $\mathbf{x}$ through the pretrained speech model M_{SM}, which generates its own set of layer-wise hidden representations $\{\mathbf{h}_j^l\}_{l=1}^{L}$. The semantic features are derived by averaging the hidden states across all layers, yielding $\mathbf{A}_j = \frac{1}{L}\sum_{l=1}^{L} \mathbf{h}_j^l$, where $\mathbf{A}_j \in \mathbb{R}^D$. This process results in the semantic representations $\mathbf{A} = \{\mathbf{A}_j\}_{j=1}^{n}$ for the speech input $\mathbf{x}$. The distillation loss in this case considers both the LM and SM representations, jointly optimizing for the alignment

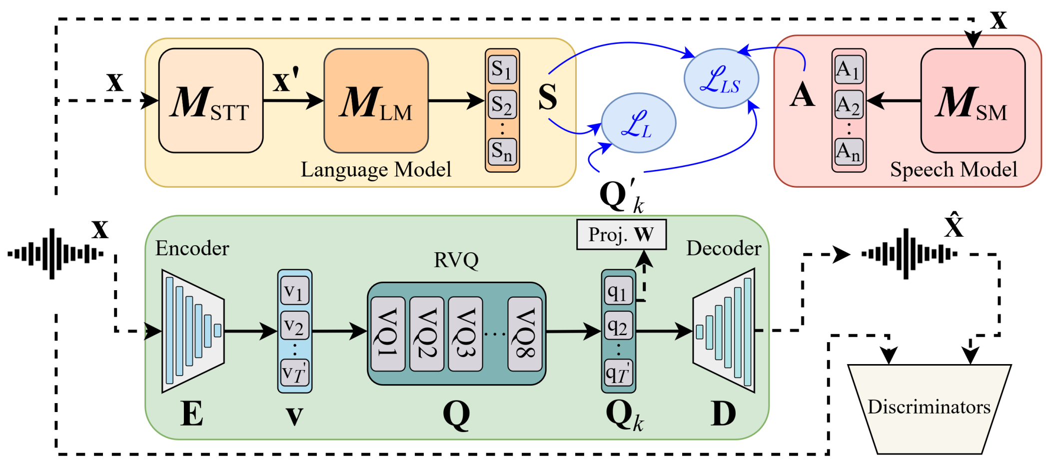

Figure 2: DM-Codec framework consists of an encoder that extracts latent representations from the input speech signal. These latent vectors are subsequently quantized using a Residual Vector Quantizer (RVQ). We designed two distinct distillation approaches: (i) distillation from a language model, and (ii) a combined distillation from both a language model (LM) and a speech model (SM). These approaches integrate acoustic, semantic, and contextual representations into the quantized vectors to improve speech representation for downstream tasks.

of the quantized features $\mathbf{Q}'_k$ with the representations $\mathbf{A}$ and $\mathbf{S}$ derived from M_{SM} and M_{LM}, respectively. Finally, the distillation loss for the SM, $\mathcal{L}_{SM}$, is first computed, followed by averaging with the LM distillation loss, $\mathcal{L}_L$, to ensure a balanced contribution from both losses. The combined distillation loss is computed as:

$$\mathcal{L}_{SM} = -\frac{1}{D}\sum_{d=1}^{D} \log\left(\sigma\left(\frac{\mathbf{Q}_k'^{(:,d)} \cdot \mathbf{A}^{(:,d)}}{\|\mathbf{Q}_k'^{(:,d)}\| \|\mathbf{A}^{(:,d)}\|}\right)\right) \tag{2}$$

$$\mathcal{L}_{LS} = \frac{1}{2}\left(\mathcal{L}_{SM} + \mathcal{L}_L\right) \tag{3}$$

This formulation ensures that DM-Codec effectively integrates both acoustic and semantic knowledge from SM, along with the contextual information provided by LM, resulting in a more robust and comprehensive set of features for speech discretization.

2.2 Model Details

Our framework builds upon the Residual Vector Quantizer with Generative Adversarial Networks (RVQ-GAN) architecture, incorporating state-of-the-art components and novel distillation techniques. The core of our model consists of an Encoder $\mathbf{E}$ and Decoder $\mathbf{D}$ with an RVQ architecture, inspired by Encodec (Défossez et al., 2022) and SpeechTokenizer (Zhang et al., 2024a). Moreover, we employ a multi-discriminator framework, comprising: Multi-Scale Discriminator (MSD), Multi-Period Discriminator (MPD), and Multi-Scale Short-Time Fourier Transform (MS-STFT) Discriminator, adopted from HiFi-Codec (Yang et al., 2023) and HiFi-GAN (Kong et al., 2020). Detailed architectural specifications for these components are provided in the Technical Appendix. This foundation provides a robust basis for speech quantization. To further enhance the quantizer with distilled multimodal representations, we use wav2vec 2.0 (wav2vec2-base-960h) as $\mathcal{M}_{STT}$ (Baevski et al., 2020), BERT (bert-base-uncased) as $\mathcal{M}_{LM}$ (Devlin et al., 2019), and HuBERT (hubert-base-ls960) as $\mathcal{M}_{SM}$ (Hsu et al., 2021). We extract the quantized output from the first layer of the RVQ (RVQ-1) for LM-guided distillation and the average of the quantized features across all eight layers (RVQ-1:8) for SM-guided distillation to calculate the distillation loss.

2.3 Training Objective

Our training strategy employs a GAN-guided framework, following methodologies established in recent work (Zhang et al., 2024a; Yang et al., 2023). In addition to the distillation loss described in Section 2.1, we utilize reconstruction losses, adversarial and feature matching losses, and a commitment loss to guide the learning process. For the original speech $\mathbf{x}$ and the reconstructed speech $\hat{\mathbf{x}}$, we calculate the losses as described below.

Reconstruction Loss. To ensure that the model preserves the key attributes of speech, we employ both time-domain and frequency-domain reconstruction losses. The time-domain loss $\mathcal{L}_t$ is computed as the L1 distance between $\mathbf{x}$ and $\hat{\mathbf{x}}$. For the frequency-domain loss $\mathcal{L}_f$, we combine L1 and L2 losses over 64-bin Mel-spectrograms Mel_i, with varying window sizes of 2^i, hop lengths of $2^i/4$, and scales $e = \{5, \ldots, 11\}$.

$$\mathcal{L}_\text{t} = \|\mathbf{x} - \hat{\mathbf{x}}\|_1 \tag{4}$$

$$\mathcal{L}_f = \sum_{i \in e} (\|\text{Mel}_i(\mathbf{x}) - \text{Mel}_i(\hat{\mathbf{x}})\|_1 + \|\text{Mel}_i(\mathbf{x}) - \text{Mel}_i(\hat{\mathbf{x}})\|_2) \tag{5}$$

Adversarial Loss. The adversarial loss promotes the generator to produce realistic and indistinguishable speech. We apply a hinge loss formulation to compute the adversarial loss for both the generator $\mathcal{L}_g$ and the discriminator $\mathcal{L}_d$. These losses are computed across all three discriminators: the multi-scale discriminator (MSD), multi-period discriminator (MPD), and the multi-scale STFT guided (MS-STFT) discriminator (details are in the Technical Appendix).

$$\mathcal{L}_g = \frac{1}{N} \sum_{n=1}^{N} \max(1 - R_n(\hat{\mathbf{x}}), 0) \tag{6}$$

$$\mathcal{L}_d = \frac{1}{N} \sum_{n=1}^{N} (\max(1 - R_n(\mathbf{x}), 0) + \max(1 + R_n(\hat{\mathbf{x}}), 0)) \tag{7}$$

where N is the number of discriminators and R_n represents the n^{th} discriminator.

Feature Matching Loss. To prevent the generator from overfitting to the discriminator's decisions, we apply a feature matching loss $\mathcal{L}_{fm}$. This loss compares features from each discriminator R_n's internal layers M across all dimensions, promoting stability and better generalization.

$$\mathcal{L}_{fm} = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \frac{\|R_n^m(\mathbf{x}) - R_n^m(\hat{\mathbf{x}})\|_1}{\text{mean}(\|R_n^m(\mathbf{x})\|_1)} \tag{8}$$

RVQ Commitment Loss. To guide the encoder to produce outputs that closely match their corresponding quantized values in the residual vector quantization (RVQ) process, we introduce a commitment loss $\mathcal{L}_w$. For N_q quantization vectors, where $\mathbf{q}_i$ represents the current residual and $\mathbf{q}_{c_i}$ is the closest entry in the corresponding codebook for the i^{th} entry, the $\mathcal{L}_w$ is computed as:

$$\mathcal{L}_w = \sum_{i=1}^{N_q} \|\mathbf{q}_i - \mathbf{q}_{c_i}\|_2^2 \tag{9}$$

Overall Generator Loss. The total generator loss $\mathcal{L}_G$ is a weighted sum of the individual loss components, including the distillation loss $\mathcal{L}_{L/LS}$ (which is either $\mathcal{L}_L$ or $\mathcal{L}_{LS}$ depending on the chosen distillation method). We use the corresponding weighting factors $\lambda_{L/LS}, \lambda_t, \lambda_f, \lambda_g, \lambda_{fm}$, and λ_w to control the influence of each loss component on the overall training objective as:

$$\mathcal{L}_G = \lambda_{L/LS}\mathcal{L}_{L/LS} + \lambda_t\mathcal{L}_t + \lambda_f\mathcal{L}_f + \lambda_g\mathcal{L}_g + \lambda_{fm}\mathcal{L}_{fm} + \lambda_w\mathcal{L}_w \tag{10}$$

This comprehensive training objective ensures DM-Codec learns acoustic speech representations while incorporating semantic and contextual representation through novel distillation approaches.

Table 1: Evaluation of speech reconstruction quality of DM-Codec and comparison with baselines. DM-Codec♠ achieves the best performance in WER, WIL, and ViSQOL, highlighting its enhanced content preservation and speech quality, with competitive intelligibility results. ♡ means the results were reproduced using the official training code. ◇ means the results were obtained using official model checkpoints. ♣ indicates LM-guided Distillation method. ♠ indicates combined LM and SM-guided Distillation method. **Bold** highlights the best result and <u>underline</u> the second-best result.

Tokenizer	WER ↓	WIL ↓	ViSQOL ↑	STOI ↑
Groundtruth	3.78	6.03	-	-
EnCodec◇	4.53	7.17	3.08	0.920
SpeechTokenizer♡	4.49	7.10	3.09	0.923
FACodec◇	4.68	7.33	3.13	**0.949**
DM-Codec ♣	<u>4.36</u>	<u>7.06</u>	<u>3.18</u>	0.935
DM-Codec ♠	**4.05**	**6.61**	**3.26**	<u>0.937</u>

3 Experimental Setup

Dataset. We trained DM-Codec using the LibriSpeech training set of 100 hours of clean speech (Panayotov et al., 2015). This dataset was selected primarily because of its successful use for training and evaluation in various speech tokenizer and modeling tasks (Zhang et al., 2024a; Ju et al., 2024; Hsu et al., 2021). Before training, we made the data uniform by randomly cropping each sample to three seconds and ensuring a consistent sample rate of 16 Hz.

Training. We trained DM-Codec utilizing 2 to 4 A100 GPUs until the model converged within 100 epochs. The batch size ranged from 6 to 20, depending on GPU resource availability. We applied a learning rate of 1×10^{-4} using the Adam optimizer with a 0.98 learning rate decay. The embedding size was set to 1024 for RVQ and 768 for the LM and SM. For all experiments, we used a random seed of 42 to ensure reproducibility. We also share our training code with the entire configuration file and a docker file to reproduce the training environment in the Technical Appendix.

Baselines. We compared DM-Codec with the baseline speech tokenizers: EnCodec (Défossez et al., 2022), SpeechTokenizer (Zhang et al., 2024a), and FACodec (NaturalSpeech3) (Ju et al., 2024). We reproduced SpeechTokenizer using the official training code and used official model checkpoints of EnCodec (EnCodec_24khz_6kpbs) and FACodec as the baselines.

Evaluation Dataset. To evaluate DM-Codec, we randomly selected 300 audio samples from the LibriSpeech test subset, following a similar practice of sampling test data used in our baselines (Zhang et al., 2024a; Zeghidour et al., 2021) and to align the experimental setup with that of SpeechTokenizer. In our experiments, we sampled the test subset of LibriSpeech using a random seed of 42. We also evaluated the baseline models with the same sampled test dataset for a fair comparison.

Evaluation Metrics. To evaluate DM-Codec, we employed different metrics suited to get insights into various aspects of information and quality preservation in the reconstructed speech. First, we used the Word Error Rate (WER) and Word Information Lost (WIL) metrics to evaluate context preservation by calculating the amount of word-level transcription errors and key information missing in transcription, respectively. For these metrics, we used the Whisper (whisper-medium) (Radford et al., 2023) model to extract the transcription from the reconstructed speech. To provide a fairer comparison and indicate the level of transcription error by the Whisper model, we also included the Groundtruth WER and WIL scores for the Whisper's transcribed text from the original speech versus the true text. Next, we assessed the acoustic and semantic information preservation using the ViSQOL (Virtual Speech Quality Objective Listener) (Hines et al., 2012) and Short-Time Objective Intelligibility (STOI) metrics, respectively. The ViSQOL metric measures the similarity between a reference and a test speech sample using a spectro-temporal measure and produces a MOS-LQO (Mean Opinion Score - Listening Quality Objective) score ranging from 1 (worst) to 5 (best). For this metric, we used the wideband model suited for speech evaluation. Lastly, the STOI metric evaluates the perceived intelligibility of speech by analyzing short-time correlations between original and reconstructed speech, with scores ranging from 0 to 1.

Table 2: Significance Analysis of DM-Codec compared to baselines EnCodec (**E**), SpeechTokenizer (**S**), and FACodec (**F**). Results reveal DM-Codec consistently achieves significantly better scores in key metrics across all individual samples. ✓ indicates that DM-Codec is significantly better, a ★ denotes dominance, and a ✗ means no significant improvement over the baseline. Avg and Std mean the average and standard deviation of each score.

WER					**WIL**					**ViSQOL**					**STOI**				
Avg	**Std**	**E**	**S**	**F**	**Avg**	**Std**	**E**	**S**	**F**	**Avg**	**Std**	**E**	**S**	**F**	**Avg**	**Std**	**E**	**S**	**F**
0.053	0.113	✓	✓	✓	0.082	0.157	✓	✓	✓	3.258	0.184	★	✓	✓	0.937	0.019	✓	✓	✗

4 Experimental Results and Discussion

We conducted extensive experiments to evaluate DM-Codec's reconstructed speech using WER and WIL for contextual information retention, and ViSQOL and STOI for semantic-acoustic information preservation. To demonstrate the effectiveness of our two distillation approaches, we present results for DM-Codec♣ (LM-guided Distillation) and DM-Codec♠ (LM and SM-guided Distillation).

4.1 Comparison of Speech Tokenization models

We compared the quality of DM-Codec's discrete speech representations by reconstructing speech from quantized vector features and comparing it with state-of-the-art (SOTA) speech tokenization models: EnCodec, SpeechTokenizer, and FACodec. For LM-guided distillation, we utilize quantized features from the first Residual Vector Quantizer layer (RVQ-1), and for the combined LM and SM-guided distillation, we average all layers (RVQ-1:8).

Results: The results in Table 1 show that DM-Codec outperformed all evaluated SOTA speech tokenization models across most metrics. Specifically, DM-Codec with only LM-guided distillation exceeds the SOTA models, achieving improved scores: WER 4.36, WIL 7.06, and ViSQOL 3.18. Furthermore, DM-Codec's with combined LM and SM-guided distillation outscore LM-guided distillation and all previous scores with 4.05 WER, 6.61 WIL, 3.26 ViSQOL, and achieved highly compatible 0.937 STOI scores compared to SOTA models.

Discussion: The observed performance gains stem from the proposed LM-guided distillation, which enhances the quantized features by leveraging LM's contextual representations. This process aligns the speech with its overall context and word relation, resulting in more accurate reconstructions, as reflected in the reduced WER and WIL scores. By embedding contextual cues, the method effectively grounds isolated phonetic units within their overall context, reconstructing speech that aligns with human expectations, as demonstrated by the higher ViSQOL and STOI scores.

Moreover, the integration of LM and SM-based distillation further amplifies these improvements. The addition of SM distillation contributes to enhanced semantic-acoustic fidelity, as SM models capture phonetic nuances alongside prosodic and tonal characteristics. This dual representation—context from LM and phonetic detail from SM—produces a more coherent and natural speech reconstruction, yielding superior results across all metrics.

4.2 Significance Analysis of Speech Tokenizer performance

We conducted a significance analysis at $\alpha = 0.05$, following the approach of Dror et al. (2019), to measure the stochastic dominance of DM-Codec over the baselines: EnCodec, SpeechTokenizer, and FACodec. Specifically, we computed inverse cumulative distribution functions (CDFs) for all reconstructed speech samples' individual WER, WIL, ViSQOL, and STOI scores. Notably, the average WER and WIL are calculated from each sentence individually, while the Table 1 scores are calculated by concatenating all sentences into one. Significance was evaluated using the ϵ value and categorized as: significantly better when $0.0 < \epsilon \leq 0.5$, significantly dominant when $\epsilon = 0.0$, and not significantly better when $\epsilon > 0.5$. For this analysis, we selected DM-Codec♠, trained with combined LM and SM-guided distillation. **To the best of our knowledge, we are the first to conduct significance analysis to measure the effectiveness of different speech tokenizers.** Here, we visualize the significance analysis for DM-Codec compared to each baseline, and we report the full significance analysis for each baseline compared to others in the Technical Appendix.

Results and Discussion: The results in Table 2 show that DM-Codec significantly outperforms the baselines in WER, WIL, ViSQOL, and STOI scores. The improved average values (0.053 WER, 0.082 WIL, 3.258 ViSQOL, 0.937 STOI) and consistent standard deviations (0.113 WER, 0.157 WIL, 0.193 ViSQOL, 0.019 STOI) further demonstrate the statistical significance. Notably, DM-Codec's performance in WER and WIL underscores the importance of contextual representation distillation for enhanced speech reconstruction. Additionally, its dominance in ViSQOL and STOI, especially over EnCodec, highlights the benefits of combining LM and SM distillation for retaining semantic-acoustic fidelity. While DM-Codec does not achieve significant dominance over FACodec in terms of STOI, it significantly outperforms the baselines across all other metrics.

4.3 Ablation Studies

We conducted a thorough analysis of DM-Codec's performance and the impact of each methodological choice in LM-guided and combined LM and SM-guided distillation. Unless otherwise stated, we use distillation for both LM and SM from the first Residual Vector Quantizer layer (RVQ-1) for comparison consistency and simplicity.

4.3.1 Ablation Study: Impact of Combined Semantic Distillation

We conducted experiments with different weighted combinations of LM and SM distillation loss to evaluate their impact on reducing WER. The combined distillation loss from Equation 3 was updated using SM and LM weights (λ_{SM} and λ_{LM}), ranging from 0.0 to 1.0, with the constraint $\lambda_{SM} + \lambda_{LM} = 1$.

$$\mathcal{L}_{LS} = \frac{1}{2} \left(\lambda_{SM} \cdot \mathcal{L}_{SM} + \lambda_{LM} \cdot \mathcal{L}_{L} \right) \tag{11}$$

Table 3: Effects of weights on combined representation distillation: Higher LM weight enhances content preservation, leading to lower WER. λ_{SM} is the SM weight, λ_{LM} is the LM weight.

λ_{SM}	λ_{LM}	**WER ↓**
1.0	0.0	4.83
0.9	0.1	4.63
0.8	0.2	4.44
0.7	0.3	4.23
0.6	0.4	4.76
0.5	0.5	4.18
0.4	0.6	4.54
0.3	0.7	4.34
0.2	**0.8**	**4.07**
0.1	0.9	4.33
0.0	1.0	4.36

Results and Discussion: The experimental results are presented in Table 3, showing the speech reconstruction results with WER scores for different weighted combinations. From the values, we notice a trend showing that incorporating LM representations significantly improves WER, especially when LM distillation is dominant. The lowest WER score of 4.07 occurs with a weight of $\lambda_{LM} = 0.8$ for LM, while $\lambda_{SM} = 0.2$ for SM, highlighting the strong influence of LM distillation on capturing contextual information. A balanced weighting of $\lambda_{SM} = 0.5$ and $\lambda_{LM} = 0.5$ produces a WER of 4.18, confirming that distillation from both LM and SM is beneficial. However, as the weighting shifts more in favor of SM ($\lambda_{SM} > 0.7$), WER deteriorates, reaching 4.83 when relying entirely on SM. This underscores that over-reliance on SM distillation compromises contextual accuracy in favor of raw speech features. Thus, an LM-dominant approach yields optimal results, while using SM alone is less effective in preserving content.

4.3.2 Ablation Study: Impact of Distillation on Different RVQ Layers

We evaluated the effect of applying distillation at various Residual Vector Quantizer (RVQ) layers, including the first layer (RVQ-1), the average of eight layers (RVQ-1:8), and the last layer (RVQ-8).

Results and Discussion: In LM-guided distillation, RVQ-1:8 achieves the best WER and WIL scores (4.23 and 6.94), though with lower ViSQOL and STOI scores (3.12 and 0.929) compared to RVQ-8 (3.28 and 0.935). The RVQ-1 layer provides the best overall balance between content preservation and perceptual quality, with WER, WIL, ViSQOL, and STOI scores of 4.36, 7.06, 3.18, and 0.935. This demonstrates RVQ-1:8 prioritizes contextual integrity, while RVQ-8 favors perceptual quality. Thus, we select RVQ-1 for LM-guided distillation due to its balanced performance.

For LM and SM-based distillation, the RVQ-1 and RVQ-1:8 combination achieves the best WER and WIL scores (4.05 and 6.61), with RVQ-1 and RVQ-1 as the second-best (4.18 and 6.84). In contrast, the RVQ-1 and RVQ-8 combination yields the highest ViSQOL and STOI scores (3.33 and 0.939),

Table 4: Analysis of different RVQ layers effect on speech reconstruction. LM-guided distillation on RVQ-1 layer ensures greater content preservation, while SM-guided distillation on RVQ-1:8 layer is more effective at preserving semantic representation. LM-layer and SM-layer indicate the RVQ layer used for respective distillation. ♣ indicates LM-guided Distillation. ♠ indicates combined LM and SM-guided Distillation. **Bold** highlights the best result and <u>underline</u> the second-best result.

Tokenizer	LM-Layer	SM-Layer	WER ↓	WIL ↓	ViSQOL ↑	STOI ↑
DM-Codec ♣	RVQ-1	-	4.36	7.06	3.18	0.935
DM-Codec ♣	RVQ-1:8	-	4.23	6.94	3.12	0.929
DM-Codec ♣	RVQ-8	-	4.44	7.22	3.28	0.935
DM-Codec ♠	RVQ-1	RVQ-1	<u>4.18</u>	<u>6.84</u>	3.13	0.933
DM-Codec ♠	RVQ-1:8	RVQ-1	4.59	7.34	3.21	0.937
DM-Codec ♠	RVQ-8	RVQ-1	4.49	7.24	<u>3.30</u>	<u>0.938</u>
DM-Codec ♠	RVQ-1	RVQ-1:8	**4.05**	**6.61**	3.26	0.937
DM-Codec ♠	RVQ-1	RVQ-8	4.39	7.08	**3.33**	**0.939**

followed by RVQ-8 and RVQ-1 (3.30 and 0.938). RVQ-1 captures contextual representation more effectively due to its simpler quantized vector, while RVQ-1:8 incorporates more nuanced semantic and acoustic aspects. Overall, this ablation shows that selecting RVQ layers for LM and SM-based distillation greatly affects the balance between contextual accuracy and semantic-acoustic fidelity, allowing layer combinations to be tailored to task requirements.

4.3.3 Ablation Study: Impact of Different Models on Distillation

We experimented with different LM and SM distillations to analyze performance variations based on different model selections. In addition to our selected BERT (Devlin et al., 2019) and HuBERT (Hsu et al., 2021), we experiment with ELECTRA (electra-base-discriminator) (Clark et al., 2020) as the LM and wav2vec 2.0 (wav2vec2-base-960h) (Baevski et al., 2020) as the SM.

Results and Discussion: In LM-guided distillation, the ELECTRA model significantly enhances performance, achieving WER and WIL scores of 4.12 and 6.63, respectively, compared to BERT's scores of 4.36 and 7.06. This indicates the architecture of ELECTRA's effectiveness for the proposed LM-guided distillation, demonstrating its superior contextual representation. These results are consistent with ELECTRA's better performance in general natural language processing tasks. However, we select BERT for its simplicity and established performance.

In LM and SM-guided distillation, the combination of BERT and wav2vec 2.0 achieves the highest overall performance, with scores of WER 4.13, WIL 6.77, ViSQOL 3.15, and STOI 0.942. However, the combination of BERT and HuBERT closely follows with second-best scores of WER 4.18, WIL 6.84, and ViSQOL 0.933. These findings demonstrate that different speech models can be effectively integrated with the BERT model.

Table 5: Analysis of representation distillation from different models. BERT can be effectively combined with HuBERT or wav2vec 2.0, however, ELECTRA in LM-guided distillation outperforms BERT. ♣ indicates LM-guided Distillation. ♠ indicates combined LM and SM-guided Distillation. **Bold** highlights the best result and <u>underline</u> the second-best result.

Tokenizer	LM	SM	WER ↓	WIL ↓	ViSQOL ↑	STOI ↑
DM-Codec ♣	BERT	-	4.36	7.06	<u>3.18</u>	0.935
DM-Codec ♣	ELECTRA	-	**4.12**	**6.63**	3.10	<u>0.936</u>
DM-Codec ♠	BERT	HuBERT	4.18	6.84	3.13	0.933
DM-Codec ♠	BERT	wav2vec 2.0	<u>4.13</u>	<u>6.77</u>	**3.15**	**0.942**
DM-Codec ♠	ELECTRA	wav2vec 2.0	4.70	7.51	3.14	0.933
DM-Codec ♠	ELECTRA	HuBERT	4.67	7.58	2.94	0.932

Table 6: Analysis of different distillation layers representation on speech reconstruction. Average layer provides more comprehensive representations. ♣ indicates LM-guided Distillation. ♠ indicates combined LM and SM-guided Distillation. **Bold** highlights the best result and <u>underline</u> the second-best result.

Tokenizer	Distillation Layer(s)	WER ↓	WIL ↓	ViSQOL ↑	STOI ↑
DM-Codec ♣	Average	<u>4.36</u>	<u>7.06</u>	**3.18**	**0.935**
DM-Codec ♣	Last	4.62	7.56	2.95	0.926
DM-Codec ♣	9^{th}	4.75	7.80	2.88	0.925
DM-Codec ♠	Average	**4.18**	**6.84**	<u>3.13</u>	<u>0.933</u>
DM-Codec ♠	Last	4.68	7.55	3.03	0.933
DM-Codec ♠	9^{th}	4.52	7.43	3.00	0.933

4.3.4 Ablation Study: Impact of Different Distillation Layer(s)

We evaluated speech reconstruction using different distillation layers of the LM and SM, examining which combination of layers yields the most relevant representations of semantic and contextual information. For this ablation, we considered the average of all layer representations, the 9^{th} layer representations, and the last layer representations. Table 5 shows the full results.

Results and Discussion: In LM-guided distillation, the use of the average layer achieves superior overall performance, with a WER of 4.36, WIL of 7.06, ViSQOL of 3.18, and STOI of 0.935, compared to the variants utilizing the last and 9th layers. Similarly, in LM and SM-guided distillation, the average layer yields superior results compared to the last and 9^{th} layer variants.

The results indicate that averaging all layers leads to more comprehensive representations of semantic or contextual information. In the case of LM, the averaging process provides greater contextual representation and synergizes syntactic information from earlier layers and abstract word relations from higher layers. In combined LM and SM-guided distillation, averaging all SM layers provides a more nuanced understanding of the earlier layer's phonetic information and the higher layers' richer semantic information. Conversely, relying solely on the last layer or the 9^{th} layer fails to capture the overall context and semantic information, yielding less relevant representation distillation.

5 Related Work

The adoption of textual LMs for speech-related tasks is a promising direction. Generally, an audio encoder converts audio signals into discrete representations, which are passed to pre-trained textual LLMs. This approach has been explored by (Hassid et al., 2024), (Wang et al., 2024), (Zhang et al., 2023), (Fathullah et al., 2023), (Shu et al., 2023), and (Rubenstein et al., 2023). Another method involves the corresponding text to feed directly into an LM (Zhang et al., 2024b). Most of these approaches aim to extract representations through LMs while focusing on speech reconstruction training objectives. Recently, LAST (Turetzky & Adi, 2024), explored a language model to tokenize speech toward improved sequential modeling, using the LLM to perform the next token prediction of quantized vectors. However, these approaches significantly differ from our method and do not focus on combining multimodal representations. More details are reported in the Technical Appendix.

6 Conclusion

In this work, we introduced a speech tokenizer DM-Codec, with two novel distillation methods to leverage multimodal (acoustic, semantic, and contextual) representations from a language model and speech self-supervised learning model. Our extensive experimental results and ablation studies suggest that distilling multimodal representations enables DM-Codec to introduce salient speech information in discrete speech tokens. Our significance analysis further revealed that DM-Codec with comprehensive multimodal representations consistently outperforms existing speech tokenizers. This approach highlights the potential of multimodal representations to enhance speech tokenization in various domains, including multilingual and code-switched speech processing.

REFERENCES

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL `https://arxiv.org/abs/2006.11477`.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020. URL `https://arxiv.org/abs/2003.10555`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL `https://arxiv.org/abs/1810.04805`.

Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance - how to properly compare deep neural models. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2773–2785, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1266. URL `https://aclanthology.org/P19-1266`.

Tianqi Du, Yifei Wang, and Yisen Wang. On the role of discrete tokenization in visual representation learning, 2024. URL `https://arxiv.org/abs/2407.09087`.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression, 2022. URL `https://arxiv.org/abs/2210.13438`.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. Towards general-purpose speech abilities for large language models using unpaired data. *arXiv preprint arXiv:2311.06753*, 2023.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Visualizing and understanding the effectiveness of bert. *arXiv preprint arXiv:1908.05620*, 2019.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, et al. Textually pretrained speech language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Andrew Hines, Jan Skoglund, Anil C. Kokaram, and Naomi Harte. Visqol: The virtual speech quality objective listener. In *International Workshop on Acoustic Signal Enhancement*, 2012. URL `https://api.semanticscholar.org/CorpusID:14792040`.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models, 2024. URL `https://arxiv.org/abs/2403.03100`.

Lauri Juvela, Bajibabu Bollepalli, Xin Wang, Hirokazu Kameoka, Manu Airaksinen, Junichi Yamagishi, and Paavo Alku. Speech waveform synthesis from mfcc sequences with generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5679–5683, 2018. doi: 10.1109/ICASSP.2018.8461852.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, 2021. URL `https://arxiv.org/abs/2106.06103`.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*, 2019.

Jingcheng Niu, Wenjie Lu, and Gerald Penn. Does bert rediscover a classical nlp pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3143–3153, 2022.

OpenAI. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/radford23a.html`.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3118–3135, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL `https://aclanthology.org/2021.acl-long.243`.

Leyuan Sheng, Dong-Yan Huang, and Evgeniy N. Pavlovskiy. High-quality speech synthesis using super-resolution mel-spectrogram, 2019. URL `https://arxiv.org/abs/1912.01167`.

Yu Shu, Siwei Dong, Guangyao Chen, Wenhao Huang, Ruihua Zhang, Daochen Shi, Qiqi Xiang, and Yemin Shi. Llasm: Large language and speech model. *arXiv preprint arXiv:2308.15930*, 2023.

Arnon Turetzky and Yossi Adi. Last: Language model aware speech tokenization, 2024. URL `https://arxiv.org/abs/2409.03701`.

Ziqian Wang, Xinfa Zhu, Zihan Zhang, YuanJun Lv, Ning Jiang, Guoqing Zhao, and Lei Xie. Selm: Speech enhancement using discrete tokens and language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11561–11565. IEEE, 2024.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. Hifi-codec: Group-residual vector quantization for high fidelity audio codec, 2023. URL `https://arxiv.org/abs/2305.02765`.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30:495–507, nov 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3129994. URL `https://doi.org/10.1109/TASLP.2021.3129994`.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*, 2023.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech large language models, 2024a. URL `https://arxiv.org/abs/2308.16692`.

Ziqiang Zhang, Sanyuan Chen, Long Zhou, Yu Wu, Shuo Ren, Shujie Liu, Zhuoyuan Yao, Xun Gong, Lirong Dai, Jinyu Li, et al. Speechlm: Enhanced speech pre-training with unpaired textual data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024b.