Distinctive Feature Codec: An Adaptive Efficient Speech Representation for Depression Detection

Xiangyu Zhang *Student Member, IEEE,*, Fuming Fang, Peng Gao, Bin Qin, Beena Ahmed *Member, IEEE,*, Julien Epps *Senior Member, IEEE*

Abstract—Large Language Models (LLMs) have demonstrated remarkable success across diverse fields, establishing a powerful paradigm for complex information processing. This has inspired the integration of speech into LLM frameworks, often by tokenizing continuous audio via neural speech codecs, enabling powerful speech language models. However, this dominant tokenization strategy relies on uniform frame-based processing at fixed time intervals. This fixed-rate approach, while effective for linguistic content, destroys the temporal dynamics. These dynamics are not noise but are established as primary biomarkers in clinical applications such as depression detection. To address this gap, we introduce the Distinctive Feature Codec (DFC), an adaptive framework engineered to preserve this vital timing information. Drawing from linguistic theory, DFC abandons fixed-interval processing and instead learns to dynamically segment the signal at perceptually significant acoustic transitions. This generates variable-length tokens that efficiently encode the temporal structure. As a key contribution, this work is the first to integrate traditional distinctive features into a modern deep learning codec for a temporally sensitive task such as depression detection. We also introduce the Group-wise Scalar Quantization (GSQ) approach to stably quantize these variable-length segments. Our distinctive feature-based approach offers a promising alternative to conventional frame-based processing and advances interpretable representation learning in the modern deep learning speech depression detection framework.

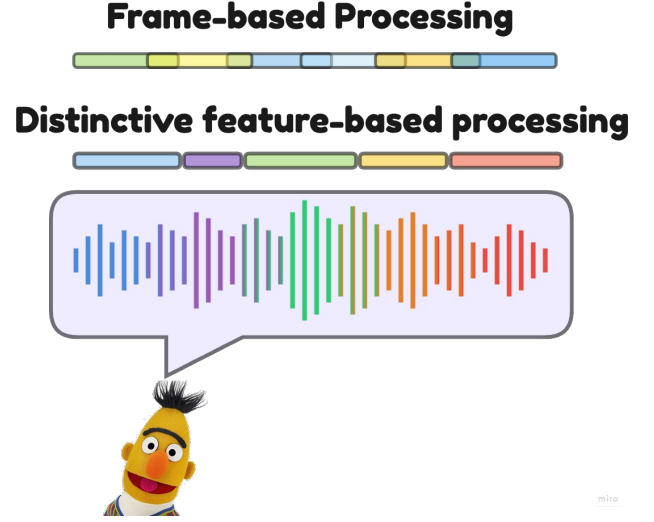*Index Terms*—Depression Detection, Tokenization

Fig. 1: Comparison of segmentation strategies. The Frame-based processing (top) imposes a rigid grid of fixed-length intervals, often disrupting natural speech events. In contrast, Distinctive feature-based processing (bottom) adaptively segments the signal by identifying perceptually significant acoustic transitions as boundaries. This results in variable-length segments that preserve the fine-grained temporal dynamics essential for clinical tasks

I. INTRODUCTION

The remarkable success of large language models (LLMs) in understanding and generating text [1], [2], [3] has inspired researchers to develop similar architectures for speech processing [4], [5], [6]. This expansion is motivated by the rich information encoded in speech signals beyond mere linguistic content, including speaker identity, emotion, and prosody [7], [8], [9], [10]. A fundamental challenge in building these speech-aware models is the tokenization of continuous audio into representations that neural networks can process. Unlike text, which has natural boundaries [11], [12], [13], speech is continuous and complex. Consequently, most current approaches—whether for discrete tokenization [14], [15], [16] or self-supervised feature extraction [17], [18]—primarily rely on **frame-based processing** with fixed time intervals.

While this fixed-rate processing is effective for tasks focused on linguistic content, such as speech recognition, it presents a fundamental limitation: it overlooks the varying information

density of speech and, critically, **destroys the fine-grained temporal dynamics** [19]. This frame-based segmentation arbitrarily cuts through natural speech events, disrupting prosodic rhythms, distorting pause structures, and obscuring speech rate variations. These temporal dynamics are not noise; they are established as primary biomarkers in critical clinical applications, most notably depression detection [20], [21]. Thus, a fundamental mismatch exists: the dominant speech representation methods are optimized for linguistic content, making them inherently unsuitable for downstream tasks that depend on temporal fidelity.

To bridge this gap, we must identify meaningful units within continuous signals in a way that preserves temporal structure. The linguistic theory of distinctive features provides an insightful alternative [22]. This theory posits that speech can be naturally segmented at points where acoustic characteristics are most differentiated [23], [24]. Instead of arbitrary fixed-length segments, this approach identifies boundaries where acoustic properties undergo significant changes. Such an adaptive, variable-length segmentation process naturally encodes

Xiangyu Zhang, Beena Ahmed, Julien Epps are with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia.

Fuming Fang, Peng Gao, Bin Qin are with Xiaomi Corp, Beijing, China.

the temporal dynamics; for example, a pause is no longer a sequence of identical "silence" frames, but is represented by a single, long segment whose duration is explicitly preserved. However, implementing distinctive feature analysis in neural networks has remained challenging. The irregular, variable-length nature of these features conflicts with the regular, grid-like computations (e.g., convolutions, transformers) that dominate modern deep learning [25], [26], [17], [18]. This technical mismatch has largely confined distinctive feature analysis to traditional signal processing approaches [19], [27].

In this paper, we address this challenge by introducing the **Distinctive Feature Codec (DFC)**, a framework designed to learn an efficient speech representation that preserves the critical temporal information required for depression detection. We propose the first architecture that successfully integrates the theory of distinctive features into a modern, end-to-end neural codec framework [28], [29]. Our approach uses a lightweight, self-supervised boundary detector to identify perceptually significant acoustic transitions, which guide an adaptive encoder-decoder (based on SEANet [30]) to process variable-length segments. This marks a departure from conventional fixed-interval processing [14], [15], [16]. Furthermore, our investigation reveals that standard quantization methods like Finite Scalar Quantization (FSQ) [31] become unstable when applied to such variable-length segments at low bitrates. To solve this, we develop a novel **Group-wise Scalar Quantization (GSQ)** approach, which ensures robust and stable quantization. Our work validates distinctive features as a promising direction for codec design, offering new perspectives on efficient speech representation for depression detection.

II. PRELIMINARY

A. Distinctive Features

Distinctive features, first proposed in linguistic theory [22], [23], [32], [24], characterize speech by identifying regions with acoustically distinctive properties that help differentiate speech segments from one another. As illustrated in Fig.1, this approach fundamentally differs from conventional frame-based processing: while frame-based methods uniformly segment speech signals into fixed-length overlapping windows, distinctive feature analysis identifies boundaries where acoustic characteristics undergo significant changes. This approach has proven valuable in early automatic speech recognition systems [23], [32], [33] and has been successfully applied to various healthcare applications [34], [35], [36], [21], [37]. However, despite its theoretical advantages, the development of distinctive feature-based methods has faced significant limitations in the deep learning era. Traditional implementations of distinctive features heavily rely on linguistic expertise and hand-crafted rules [23], [24], leading to limited training data and difficulties in scaling across different acoustic conditions and languages. Additionally, while frame-level processing naturally aligns with convolutional neural networks and enables efficient batch processing, the variable-length nature of distinctive features poses challenges for modern deep learning architectures. The success of frame-level processing in various deep learning systems [17], [18] has led to a rich ecosystem

TABLE I: Analysis of Sylber Codec performance. The table shows mel_error, stft, pseq, and stoi metrics for different KM (K-Means) cluster sizes.

Sylber Codec	mel_error ↓	stft ↓	PESQ ↑	stoi ↑
KM=5K	0.6204	2.0246	1.0025	0.7066
KM=10K	0.6155	2.2049	1.2501	0.7136
KM=20K	0.6727	2.1525	0.8135	0.7003

of pre-trained models and established practices, making it the predominant choice for modern speech processing systems despite its inherent inefficiencies. This technical mismatch, combined with the limited availability of labeled data for distinctive feature analysis, has constrained its adoption in contemporary deep learning approaches.

B. Speech Codecs

Speech codecs compress speech signals into discrete tokens while preserving essential acoustic and linguistic information [38], [39]. These discrete representations serve as inputs for downstream tasks such as large language models [5], [6], [4]. Current approaches can be broadly categorized into two paradigms: **end-to-end trained codecs** that directly learn discrete representations through reconstruction objectives [15], [14], [16], and **two-stage approaches** that first extract semantic features using self-supervised models [18], [17], then apply generative modeling techniques such as flow matching [40], [41] or diffusion [42], [43].

Two-stage approaches (e.g., Sylber [44]) achieve impressive compression by operating at syllable-level granularity [45], [46], reducing token rates to as low as 5-10 tokens per second. These methods prioritize *semantic preservation*—retaining linguistic content that overlaps with text representations—while relying on powerful generative models. However, this design philosophy is fundamentally misaligned with depression detection, where diagnostic information resides not in semantic content but in *acoustic characteristics*: voice quality, spectral energy distribution, and prosodic micro-variations [21].

Table I, based on our experimental evaluation of the Sylber model, empirically demonstrates this limitation. As Sylber's model capacity increases (K-means clusters from 10K to 20K), perceptual quality (PESQ) paradoxically *degrades* from 1.25 to 0.81, while mel-spectral error increases from 0.616 to 0.673. This pattern indicates that the generative model produces perceptually plausible speech by hallucinating acoustic details rather than preserving them from the original signal. While the semantic content ("what was said") remains intact, the acoustic substrate ("how it was said") diverges from the input—erasing the subtle spectral and prosodic deviations that constitute depression biomarkers. Multiple studies confirm that such semantically-focused representations underperform on tasks requiring acoustic fidelity [47], [48].

Our work therefore focuses on **end-to-end trained codecs**, which optimize for faithful acoustic reconstruction rather than semantic compression, explicitly preserving the spectral and temporal dynamics essential for clinical applications.
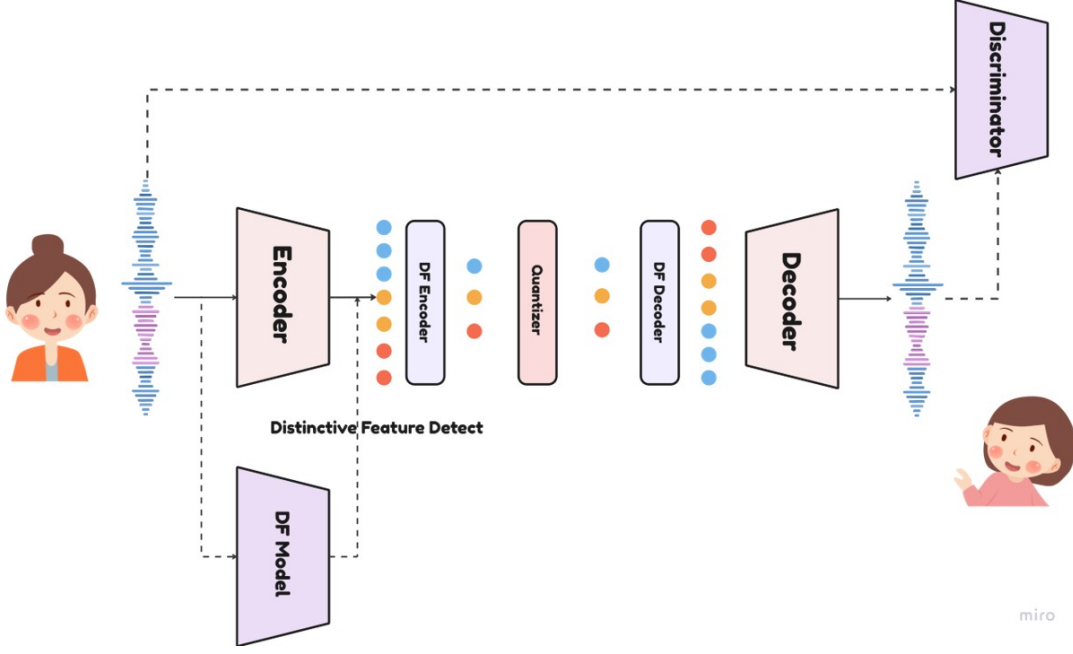
Fig. 2: Overview of the Distinctive Feature Codec framework. The Distinctive Feature Detector (top) identifies acoustic boundaries through contrastive learning to guide variable-length segmentation. The codec pipeline segments encoded speech based on these boundaries, compresses each segment via the DF Encoder, applies quantization, and reconstructs through the DF Decoder and main decoder.

C. Speech Representation for Depression Detection

Clinical research has long established that depression profoundly affects speech production mechanisms, manifesting in distinctive acoustic biomarkers known as "psychomotor retardation" [49]. These manifestations primarily include reduced speech rate, monotonic prosody, prolonged pauses, and blunted articulation [50], [51]. Consequently, capturing these **temporal dynamics** and **prosodic variations** has been a central focus in feature extraction for depression detection.

Early approaches predominantly relied on hand-crafted Low-Level Descriptors (LLDs) to explicitly model these characteristics. Standard feature sets, such as eGeMAPS and INTERSPEECH ComParE [52], aggregate frame-level acoustic properties (e.g., pitch, energy, formants) using statistical functionals to capture global prosodic trends. More distinctively, some studies explored **landmark-based** or **distinctive feature** analysis [20], [37], which focuses on specific acoustic events—such as the onset of bursts or glottal transitions—to detect subtle articulatory coordination deficits associated with depression [20], [34], [21]. These methods offered high interpretability and preserved the temporal integrity of speech events but were limited by their inability to model complex, high-level abstractions.

In the deep learning era, the paradigm shifted towards learning data-driven representations directly from raw audio or spectrograms using Convolutional Neural Networks (CNNs) and Transformers [25], [26]. Self-supervised models (SSL) like Wav2Vec 2.0 and HuBERT have achieved state-of-the-art performance by encoding speech into continuous or discrete representations [17], [18], [15], [16]. However, these modern approaches predominantly employ **fixed-rate frame-based processing**, where speech is segmented into uniform time intervals regardless of the underlying acoustic content. While effective for linguistic tasks like ASR, recent studies suggest that this rigid segmentation may disrupt the fine-grained temporal structures—such as variable-length pauses and rhythmic patterns—that serve as critical diagnostic cues [19], [37], [21]. This creates a fundamental mismatch: the dominant feature extraction methods are optimized for semantic continuity, potentially at the cost of the temporal fidelity required for reliable depression assessment.

III. DISTINCTIVE CODEC FRAMEWORK

As illustrated in Fig. 2, our approach consists of two key stages: First, a lightweight boundary detector is trained to identify perceptually significant transitions in speech signals. Second, these detected boundaries guide the codec model to adaptively merge or separate speech segments, leading to more efficient tokenization that aligns with the natural structure of speech.

A. Distinctive Features Detector

The core idea of distinctive features lies in identifying regions where speech segments exhibit maximal acoustic contrast with their neighbors. This naturally aligns with the objective of contrastive learning, which aims to learn representations by maximizing the differences between distinct samples while minimizing differences between similar ones [53], [54], [55], [56]. We leverage this connection to design a self-supervised boundary detector that learns to identify distinctive features without requiring phoneme-level annotations.

Specifically, we train a lightweight encoder network $f_\theta : \mathbb{R}^L \to \mathbb{R}^D$ that maps raw speech input segments $\mathbf{x} = \{x_1, ..., x_T\}$ into a latent representation space, where L is the segment length and D is the latent dimension. For a given segment x_t at position t, we compute its latent representation $\mathbf{z}_t = f_\theta(x_t)$ and compare it with subsequent segments at different positions $t + k_{k=1}^K$. The similarity score is computed along the feature dimension:

$$s(t, k) = -\alpha \cdot \cos(\mathbf{z}_t, \mathbf{z}_{t+k})_D \tag{1}$$

where α is a scaling coefficient. For each positive pair (x_t, x_{t+k}), we construct a set of negative samples by randomly shuffling segments from the same batch. The model is trained to minimize the contrastive loss:

$$\mathcal{L} = -\mathbb{E}_{t,k}[\log \frac{\exp(s(t,k)/\tau)}{\exp(s(t,k)/\tau) + \sum_{n=1}^{N} \exp(s(t,n)/\tau)}] \tag{2}$$

where τ is a temperature parameter and N is the number of negative samples. This contrastive objective encourages the model to learn representations that capture the inherent acoustic differences between speech segments. The resulting similarity scores naturally highlight regions where acoustic characteristics undergo significant changes, corresponding to distinctive feature boundaries. These boundaries then guide the subsequent merging of segments in our codec model.

B. Distinctive Codec

To evaluate our distinctive feature-based approach, we build our codec model on top of the SpeechTokenizer framework [16], which leverages the SEANet architecture [30] for encoder-decoder operations. The key innovation of our approach lies in how we process the encoded representations based on the distinctive features detected by our boundary detector. Given input speech $\mathbf{x} \in \mathbb{R}^{1 \times L}$, the encoder E_θ first maps it to a latent representation:

$$\mathbf{e} = E_\theta(\mathbf{x}) \in \mathbb{R}^{D \times T} \tag{3}$$

where D is the feature dimension and $T = L/r$ represents the temporal dimension after downsampling with ratio r. Traditional frame-based codecs would typically process this representation uniformly across time. In contrast, our distinctive codec first identifies segment boundaries $\mathcal{B} = b_1, b_2, \ldots, b_M$ using the boundary detector described in Section III-A. With these boundaries, we partition the feature sequence into variable-length segments:

$$\mathbf{S} = \mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{M+1} \tag{4}$$

where $\mathbf{s}_i = \mathbf{e}[:, b_{i-1} : b_i]$ represents the i-th segment ($b_0 = 0$ and $b_{M+1} = T$ for notational convenience). For each segment, we apply a Distinctive Feature encoder (DFE) to compress the variable-length representation into a fixed-length embedding:

$$\mathbf{z}_i = \text{DFE}(\mathbf{s}_i) \in \mathbb{R}^{H \times 1} \tag{5}$$

where H is the hidden dimension. This operation effectively merges temporal information within each segment into a single token, guided by the distinctive feature boundaries. During
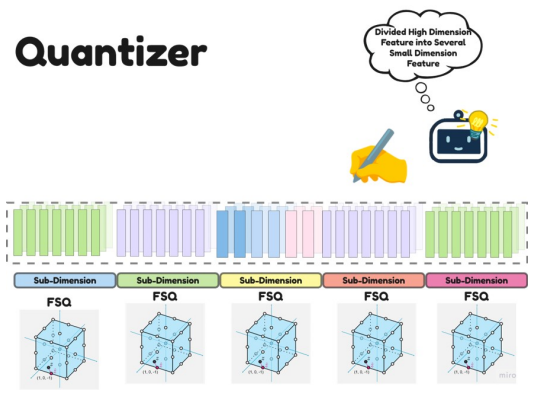

Fig. 3: Features are divided into smaller groups for independent quantization, enhancing stability and representation quality.

decoding, we expand each compressed segment embedding back to its original length using a Distinctive Feature decoder (DFD):

$$\hat{\mathbf{s}}_i = \text{DFD}(\mathbf{z}_i, l_i) \tag{6}$$

where $l_i = b_i - b_{i-1}$ is the original segment length. The full sequence is reconstructed by concatenating the expanded segments:

$$\hat{\mathbf{e}} = \text{Concat}(\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \ldots, \hat{\mathbf{s}}_{M+1}) \tag{7}$$

Finally, the decoder D_ϕ transforms the reconstructed latent representation back to the waveform domain:

$$\hat{\mathbf{x}} = D_\phi(\hat{\mathbf{e}}) \tag{8}$$

C. Group-wise Scalar Quantization

Finite Scalar Quantization (FSQ) [31] has emerged as an effective approach for discrete representation learning due to its computational efficiency and strong performance across various tasks [6], [57], [58]. Unlike vector quantizers that require nearest neighbor search in high-dimensional spaces, FSQ directly quantizes each dimension of the latent representation independently, significantly reducing computational complexity.

However, during our experiments with distinctive feature-based tokenization, we discovered that standard FSQ becomes unstable when operating at high downsampling rates. This instability manifests as training divergence and poor reconstruction quality, particularly when compressing longer, variable-length segments into single tokens. We hypothesize that this issue stems from the increased difficulty of directly quantizing high-dimensional features with varying temporal characteristics.

To address this issue, we propose Group-wise Scalar Quantization (GSQ) as shown in Figure 3, which decomposes the high-dimensional quantization problem into multiple lower-dimensional sub-problems. Given a compressed segment representation $\mathbf{z}_i \in \mathbb{R}^{H \times 1}$, we divide it into G groups, each with dimension $H_g = H/G$:

$$\mathbf{z}_i = [\mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^G] \tag{9}$$

Each group is then processed independently through quantization. The key design choice lies in how we parameterize the

quantization function for each group $Q_g(\cdot)$. We explore two variants:

Direct Quantization (M2M): Each group is quantized directly without dimensional change:

$$\hat{\mathbf{z}}_i^g = \mathrm{FSQ}_g(\mathbf{z}_i^g), \quad \mathbf{z}_i^g, \hat{\mathbf{z}}_i^g \in \mathbb{R}^{H_g} \qquad (10)$$

This preserves the distributed representation across all H_g dimensions but requires a codebook of size L^{H_g} per group (where L is the number of quantization levels per dimension).

Projection-based Quantization (M2O): Each group is first projected to a lower-dimensional space (in our case, a scalar), quantized, then projected back:

$$p_i^g = \mathbf{W}_g^{(c)} \mathbf{z}_i^g, \quad \hat{p}_i^g = \mathrm{FSQ}_g(p_i^g), \quad \hat{\mathbf{z}}_i^g = \mathbf{V}_g^{(e)} \hat{p}_i^g \qquad (11)$$

where $\mathbf{W}_g^{(c)} \in \mathbb{R}^{k \times H_g}$ is the compression matrix, $\mathbf{V}_g^{(e)} \in \mathbb{R}^{H_g \times k}$ is the expansion matrix, and k is the projection dimension (we use $k = 1$, i.e., scalar projection). This reduces the codebook size to L^k per group while maintaining expressiveness through learned projections.

The full representation is reconstructed by concatenating all quantized groups:

$$\hat{\mathbf{z}}_i = [\hat{\mathbf{z}}_i^1, \hat{\mathbf{z}}_i^2, \ldots, \hat{\mathbf{z}}_i^G] \qquad (12)$$

This group-wise decomposition is particularly beneficial for quantizing variable-length segments with diverse acoustic characteristics. By processing features in smaller groups with dedicated quantizers, we reduce quantization complexity through lower-dimensional operations. Our primary experiments employ the M2O variant with scalar projection ($k = 1$) due to its superior stability at low token rates and dramatically reduced codebook size. Algorithm 1 presents the unified framework, with the projection step (lines 3-4, 6) being optional for the M2M variant.

IV. EXPERIMENTAL SETUP

We implement our Distinctive Codec using the widely-adopted SEANet-based [30] encoder-decoder architecture, which has become the de facto standard in modern neural speech codecs including SoundStream [15], EnCodec [59], DAC [59], and SpeechTokenizer [16]. This architectural choice enables direct comparison with existing methods while isolating the impact of our distinctive feature-based adaptive segmentation versus conventional fixed-rate processing. The encoder employs strided convolutional layers for downsampling, bidirectional LSTM layers for temporal modeling, and residual blocks for feature refinement. The decoder mirrors this structure to reconstruct the waveform from the latent representation.

For training and evaluation, we use the LibriSpeech dataset [60], a widely-adopted benchmark containing 960 hours of read English speech with diverse speakers and acoustic conditions. We train on the standard train set and evaluate codec reconstruction quality on 500 randomly selected samples from the test-clean set. This evaluation protocol maintains consistency with prior codec research and enables direct performance comparison across methods. For downstream evaluation of depression detection, we use the DAIC-WOZ dataset [61], which contains clinical interviews specifically

---

Algorithm 1 Group-wise Scalar Quantization (GSQ)

Require: Input tensor $x \in \mathbb{R}^{B \times T \times D}$
Require: Number of groups G, group size $d = D/G$
Require: Use projection: `use_proj` $\in \{\mathrm{true}, \mathrm{false}\}$
Require: Projection dimension: k (typically $k = 1$ for M2O)
Require: *If* `use_proj`: compression matrices $\{W_i^{(c)} \in \mathbb{R}^{k \times d}\}_{i=1}^G$, expansion matrices $\{W_i^{(e)} \in \mathbb{R}^{d \times k}\}_{i=1}^G$
Require: FSQ quantizers $\{\mathrm{FSQ}_i\}_{i=1}^G$
Ensure: Quantized output $\hat{x} \in \mathbb{R}^{B \times T \times D}$

1:
2: **for** $i = 1$ **to** G **do**
3:    $x_i \leftarrow x[:, :, i \cdot d : (i+1) \cdot d]$ {Extract group i}
4:    **if** `use_proj` **then** {M2O: project to lower dim}
5:       $x_i \leftarrow W_i^{(c)} x_i$
6:    **end if**
7:    $\hat{x}_i \leftarrow \mathrm{FSQ}_i(x_i)$ {Quantize (M2M or M2O)}
8:    **if** `use_proj` **then** {M2O: project back}
9:       $\hat{x}_i \leftarrow W_i^{(e)} \hat{x}_i$
10:   **end if**
11: **end for**
12:
13: $\hat{x} \leftarrow \mathrm{Concat}([\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_G], \dim = -1)$
14: **return** $\hat{x}$

---

designed for mental health assessment and provides naturalistic speech with diagnostically relevant temporal dynamics.

A. Details of Distinctive Feature Detector

The Distinctive Feature Detector, a core component of our framework, was implemented as a lightweight CNN-based architecture designed to identify perceptually significant transitions in speech signals using contrastive learning.

a) Model Architecture: The detector uses a 5-layer CNN structure processing raw audio input directly. Each layer consists of a 1D convolutional operation followed by batch normalization and LeakyReLU activation. The network employs variable kernel sizes and strides to progressively downsample the input while capturing acoustic patterns at different time scales. For our primary configuration which yields 9.5 tokens per second (as shown in Table II), the convolutional layers use kernel sizes of 10, 8, 8, 4, and 4, with corresponding stride values of 5, 4, 4, 2, and 2. For our higher frame rate configuration (15.7 tokens per second), we adjusted the stride values to 5, 4, 4, 4, and 2, demonstrating the adaptability of our approach. This flexibility allows our distinctive feature-based method to operate effectively across different token rates, unlike approaches such as Sylber Codec that are limited to specific syllable-level rates. Our configurable architecture provides an effective receptive field capable of capturing both local and broader acoustic transitions while allowing token rate adjustments based on application requirements.

The network's final embedding dimension was set to 256, with an optional projection layer that could further reduce this dimension to 64 for more compact representations. We found that applying this projection with a linear transformation worked well in practice, so we set `z_proj_linear` to true in

our experiments. To enhance the model's robustness and prevent overfitting, we incorporated an optional dropout mechanism in the projection layers, though we found that for our primary experiments, setting the dropout rate to 0 yielded optimal results.

b) Contrastive Learning Approach: For training, we employed a contrastive learning objective where the model learned to identify acoustic boundaries by predicting future frames. We used a single-step prediction horizon (`pred_steps=1`) with no offset (`pred_offset=0`), which we found provided the most reliable boundary detection performance. The similarity between predicted and actual frames was measured using cosine similarity with a scaling coefficient of 1.0.

For each positive pair, we constructed a negative pair using a random permutation strategy. While our implementation supported both within-utterance and cross-utterance negative samples through the `batch_shuffle` parameter, we found that using within-utterance negatives (setting `batch_shuffle=false`) produced more consistent results, as it forced the model to learn fine-grained distinctions within the same acoustic context.

c) Training Details: The detector was trained on a speech dataset, with the primary experiments conducted using the Liberspeech dataset. We used the Adam optimizer with a learning rate of 0.0002, a batch size of 80, and trained for up to 200 epochs. Early stopping was employed based on validation performance to prevent overfitting. All training was conducted on NVIDIA V100 GPUs.

d) Boundary Detection Inference: During inference, the feature detector outputs similarity scores that undergo several post-processing steps to identify boundaries. The raw scores from prediction steps are combined and normalized using min-max normalization. Boundary detection is then performed using a peak detection algorithm, which identifies local maxima in the processed similarity scores. The key parameters for peak detection include a prominence threshold of 0.01, along with optional width and distance constraints that were automatically tuned during training.

B. Details of Distinctive Codec

The Distinctive Codec builds upon the SEANet-based encoder-decoder architecture used in SpeechTokenizer, extending it with our distinctive feature detection and variable-length segment processing capabilities. Here, we provide implementation details not covered in the main text.

a) Model Architecture: The encoder consists of a SEANetEncoder with 64 initial filters and a feature dimension of 1024. For our primary configuration yielding 9.5 tokens per second, we use strides of [8,5,4,2] (resulting in a total downsampling ratio of 320). For our higher frame rate configuration (15.7 tokens per second), we employ strides of [8,5,2,2], which produces a lower downsampling ratio and consequently more tokens per second. The encoder includes 2 bidirectional LSTM layers and a residual network with kernel size 3 and 1 residual layer per block. After encoding, the high-dimensional features (1024) are projected to a lower dimension (72) to make the subsequent distinctive feature processing more efficient.

The Distinctive Feature Encoder (DFE), implemented as the PerSegmentAutoEncoder in our code, compresses variable-length segments into fixed-length representations. The encoder component uses two convolutional layers with kernel size 3 and stride 1, followed by an adaptive average pooling operation to compress the temporal dimension to a single token. This architecture efficiently captures the salient information within each distinctive segment while maintaining a consistent output shape regardless of input segment length.

For quantization, we implemented Group-wise Scalar Quantization (GSQ) through our RefinedProjectionFSQ module, which divides the feature vector into multiple groups. Each group undergoes independent projection-based quantization to improve stability and representation quality. This approach was crucial for maintaining performance when operating at lower token rates.

The Distinctive Feature Decoder (DFD), also implemented within the PerSegmentAutoEncoder module, reconstructs the variable-length segments from the quantized representations. It uses nearest-neighbor interpolation to expand the fixed-length representations to their original temporal dimensions, followed by two convolutional layers with kernel size 3 and stride 1 to refine the expanded features. The decoder output is then projected back to the original high dimension (1024) before being processed by the SEANetDecoder, which mirrors the encoder structure to generate the final waveform.

b) Training Methodology: The Distinctive Codec was trained using a combination of reconstruction and perceptual losses:

- **Time-Domain Reconstruction Loss**: We used L1 loss between the original and reconstructed waveforms, weighted by a factor of 500 to ensure accurate time-domain reconstruction.
- **Multi-resolution Mel-spectrogram Loss**: To capture perceptual qualities at different time scales, we employed a multi-resolution approach with four mel-spectrogram losses at different resolutions (using FFT sizes that vary by factors of 2). These losses combined L1 and L2 distances and were weighted at [45, 1, 1, 1] respectively, emphasizing the base resolution.
- **Adversarial Losses**: We employed multiple discriminators to improve the perceptual quality:
  - Multi-Period Discriminators with periods [2, 3, 5, 7, 11]
  - Multi-Scale Discriminators operating at different resolutions
  - Multi-Scale STFT Discriminators analyzing the spectral characteristics

c) Implementation Details: The model was trained for 20 epochs with a batch size of 9 using the Adam optimizer with learning rate 1e-4 and betas [0.9, 0.99]. Training was performed on LibriSpeech using 48000-sample segments (3 seconds at 16kHz) on 4 NVIDIA V100 GPUs. We used a cosine annealing learning rate schedule over the course of training. To ensure stable training, we found that initializing network weights with near-orthogonal initialization improved convergence. The model checkpoints were saved every 2,500 steps, with the final

TABLE II: Performance comparison of frame-based and distinctive feature-based speech tokenization methods. Frame: frame rate (Hz); TKR: tokens per second [62]; BPS: bits per second [58]. RVQ: Residual Vector Quantizer [63]; FSQ: Finite Scalar Quantization [31]; GSQ: Group-wise Scalar Quantization (ours). Lower values are better for MEL Error, STFT, and WER; higher values are better for PESQ [64] and STOI [65]. All models use identical SEANet-based encoder-decoder architecture trained on LibriSpeech.

Segmentation	Quantization	Frame	TKR	BPS	Metrics				
					MEL Error↓	STFT↓	PESQ↑	STOI↑	WER↓
Frame-based	RVQ	10	10	100	0.4487	2.0183	1.2844	0.6695	0.9340
	RVQ	12.5	12.5	125	0.4290	1.9448	1.3245	0.6624	0.7701
	RVQ	20	20	200	0.3796	1.7401	1.4960	0.7225	0.6887
	RVQ	50	50	500	0.2342	0.5648	2.4496	0.8439	0.1698
	FSQ	20	20	320	0.1933	0.4699	2.5891	0.8615	0.1428
Distinctive Feature -based (Ours)	RVQ	9.5	9.5	95	0.4481	2.0040	1.3312	0.6930	0.8523
	FSQ	9.5	9.5	152	0.4033	1.9042	1.4649	0.7049	0.6794
	GSQ	9.5	9.5	152	0.2857	1.3213	1.9147	0.7675	0.4265
	GSQ	15.7	15.7	251	0.2468	0.9072	2.3092	0.8203	0.2637

model selected based on the lowest validation mel-spectrogram error. The average token rate of our model is 9.5 tokens per second, with the actual rate varying based on the acoustic complexity of the input speech.

C. Codec Evaluation Metrics

We evaluate our Distinctive Codec using metrics for reconstruction quality, intelligibility, and encoding efficiency following previous works [15], [14]. For reconstruction, we measure mel-spectral error, STFT distance, and PESQ (Perceptual Evaluation of Speech Quality) scores [64]. For intelligibility, we measure Word Error Rate (WER) using the Whisper en-medium model [66], following SpeechTokenizer [16], and STOI (Short-Time Objective Intelligibility) [65] to quantify how accurately the speech content is preserved. We also track encoding efficiency through Token Ratio (TKR) [62], representing tokens per second of 16 kHz audio, and Bits Per Second (BPS). The BPS calculation follows the approach in [58] which will consider the vocabulary size of quantization method.

D. Depression Detection Evaluation

Evaluating the downstream performance of speech tokenizers for clinical applications presents fundamental methodological challenges. Training full-scale clinical assessment models requires substantial computational resources and carefully curated clinical data, making comprehensive comparison across multiple tokenization approaches prohibitively expensive. Furthermore, traditional codec evaluation metrics that focus on reconstruction quality and speech recognition fail to capture the preservation of subtle acoustic and temporal characteristics essential for mental health assessment, creating a significant evaluation gap between codec performance and clinical utility.

To enable systematic comparison of how different tokenization strategies preserve clinically relevant temporal information,

we adopt the token projection evaluation framework proposed in [67]. This methodology isolates the effects of tokenization by maintaining identical downstream architectures and training procedures across all evaluated methods, ensuring that performance differences directly reflect the quality of information preserved in discrete representations rather than variations in model design. We apply this comparative framework to the binary depression classification task using the DAIC-WOZ dataset [61], which requires capturing subtle prosodic variations, pause patterns, and speech rate fluctuations—the temporal dynamics that our distinctive feature-based approach is designed to preserve.

a) Evaluation Protocol.: Depression detection is evaluated on the DAIC-WOZ dataset, containing 107 training participants and 35 development set participants from clinical interviews. Following standard protocols, we extract participant speech segments (excluding interviewer turns) from each session and tokenize them using different codec methods under comparison. Depression labels are derived from PHQ-8 scores, with scores ≥ 10 indicating clinical depression. For each tokenization method, we employ a Llama 3.1 8B model [2] with trainable projection module and classification components, trained using AdamW (lr=5×10^{-5}, batch size 16) for 50 epochs with bf16 mixed precision. Participant speech from each session is processed as variable-length segments projected to 128 tokens. We report F1-score, UAR, and accuracy as evaluation metrics, with all experiments conducted three times using different random seeds and results averaged across runs.

V. EXPERIMENTAL RESULTS

A. Codec Results

Table II presents a systematic comparison between frame-based and distinctive feature-based tokenization approaches under identical architectural and training conditions. Our distinctive feature-based codec operates at average token rates

TABLE III: Depression detection performance comparison on DAIC-WOZ development set. All models use identical downstream architecture (Llama 3.1 8B with projection module) and training protocol. Results are averaged over three random seeds.

Tokenization Method	F1-Score
Frame-based (10Hz RVQ)	0.471
Distinctive Feature-based (9.5Hz RVQ)	0.533
Distinctive Feature-based (9.5Hz GSQ)	**0.636**

of 9.5 Hz and 15.7 Hz, significantly lower than conventional fixed-rate processing while maintaining superior reconstruction quality.

The effectiveness of our distinctive feature-based approach is evident when comparing models at similar token rates. At 9.5 Hz, our method with RVQ quantization outperforms the frame-based baseline at 10 Hz across all metrics. Despite using a slightly lower token rate, our approach reduces MEL Error by 0.0006, decreases STFT distortion by 0.0143, improves perceptual quality (PESQ) by 0.0468, enhances intelligibility (STOI) by 0.0235, and reduces WER by 0.0817. These consistent improvements demonstrate that adaptive segmentation guided by acoustic boundaries preserves more acoustic information than arbitrary fixed-interval processing, even when using fewer tokens per second.

We observe further improvements when replacing RVQ with FSQ in our distinctive feature-based codec. Notably, frame-based approaches exhibit instability with FSQ at lower token rates (below 20 Hz), preventing direct comparison at our model's operating point of 9.5 Hz. This instability supports our hypothesis that fixed-rate segmentation at long intervals forces the quantizer to represent acoustically heterogeneous content within single frames, leading to training difficulties. The comparison between GSQ and FSQ at the same token rate (9.5 Hz) clearly demonstrates the effectiveness of our Group-wise Scalar Quantization approach. GSQ substantially outperforms FSQ across all metrics, reducing MEL error by 0.1176, decreasing STFT distortion by 0.5829, improving PESQ by 0.4498, increasing STOI by 0.0626, and lowering WER by 0.2529. This validates our hypothesis that decomposing high-dimensional quantization into multiple lower-dimensional sub-problems enhances stability and representation quality for variable-length segments generated by distinctive feature-based segmentation.

B. Depression Detection Results

Table III presents the depression detection performance comparison across tokenization approaches. Our distinctive feature-based codec with GSQ achieves an F1-score of 0.636, substantially outperforming both the frame-based baseline (F1=0.471, +35.0% relative improvement) and our method with standard RVQ quantization (F1=0.533, +19.3% relative improvement). These results provide direct evidence that timing-preserving tokenization captures clinically relevant temporal dynamics that are destroyed by fixed-rate processing.

The performance hierarchy reveals two complementary effects. First, comparing frame-based (F1=0.471) versus distinc-

tive feature-based with RVQ (F1=0.533) isolates the contribution of adaptive segmentation: aligning token boundaries with acoustic transitions preserves the pause patterns, speech rate variations, and prosodic contours that serve as established biomarkers for depression. Second, the further gain from GSQ (F1=0.636) demonstrates that stable quantization of variable-length segments is crucial for preserving fine-grained temporal information at low token rates. Together, these findings validate that our approach—adaptive segmentation plus robust quantization—effectively encodes the temporal dynamics essential for clinical speech analysis.

VI. ABLATION STUDY AND DISCUSSION

A. Analysis of Distinctive Feature Effectiveness

a) Theoretical Derivation Hypothesis: Consider an autoencoder that extracts a latent representation from input speech frames. For each segment of input frames $x \in \mathbb{R}^{D \times L}$, the encoder produces a latent vector $z = f(x) \in \mathbb{R}^d$. The autoencoder training objective and quantization distortion can be formulated as:

$$\min_{f,g} \ \mathbb{E}_x \left[\|x - g(f(x))\|^2 \right], \quad D = \mathbb{E} \left[\|z - Q(z)\|^2 \right] \quad (13)$$

where g is the decoder and Q is the quantizer that maps z to a finite codebook. High-resolution quantization theory [68] approximates this distortion as:

$$D \approx G(d) \left(\int p(z)^{\frac{d}{d+2}} dz \right)^{\frac{d+2}{d}} 2^{-\frac{2R}{d}} \quad (14)$$

where $G(d)$ is a dimension-dependent constant and R is the bit rate. The integral term is critical, as it depends on the distribution of latent vectors.

Based on these formulations, we hypothesize that the effectiveness of distinctive features stems from their impact on the latent distribution $p(z)$. We posit that frame-based segmentation, which arbitrarily divides speech without regard to acoustic boundaries, may potentially force a single segment to capture multiple distinct speech states. This would cause the resulting latent vectors to represent a mixture of acoustic features, leading to a more diffuse distribution in the latent space. Under this hypothesis, our distinctive feature approach should yield more concentrated latent distributions by aligning segment boundaries with natural acoustic transitions. Specifically, when segments contain acoustically homogeneous content, the encoder can produce latent vectors that cluster more tightly around prototype representations of discrete speech units. Such a multimodal distribution would theoretically allow for more efficient quantization, as codebook entries could be optimally positioned to capture these distinct modes, thereby reducing the overall distortion D for a given rate R.

b) Codebook Analysis: Building on our theoretical analysis, we now examine the empirical evidence supporting our hypothesis through codebook utilization patterns. Since our theoretical framework suggests that distinctive features should allow for more efficient quantization by producing more concentrated latent distributions, analyzing codebook utilization provides a direct way to verify this effect in practice.
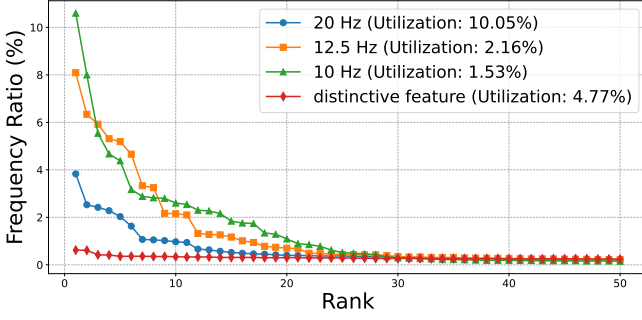
(a) Top 50 Code Frequency Ratios with Their Overall Utilization Rates Shown in the Legend.
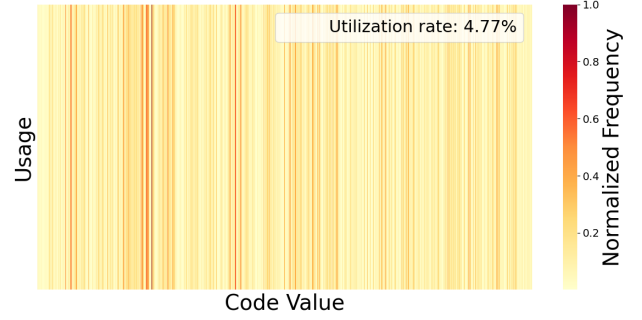
(b) Visualization of Codebook Usage for Codec Using Distinctive Feature

Fig. 4: Codebook utilization comparison between frame-based and distinctive feature-based processing. Our approach achieves more balanced and efficient codebook usage (4.77% utilization) compared to frame-based methods at similar frame rates (1.53% at 10Hz).

To investigate this hypothesis, we conducted experiments comparing codebook utilization across different processing approaches, with all models using FSQ for quantization. Figure 4 presents the utilization statistics for frame-based models operating at different frame rates (20Hz, 12.5Hz, and 10Hz) alongside our Distinctive Codec. The results reveal several important insights. As shown in Figure 4a, when frame rates decrease in conventional frame-based models, codebook utilization rates drop dramatically—from 10.05% at 20Hz to merely 1.53% at 10Hz. This declining utilization explains the instability we encountered when attempting to run Speech Tokenizer with FSQ at lower frame rates, as the quantizer struggles to effectively represent the diverse acoustic content when arbitrarily segmented at longer intervals.

In stark contrast, our Distinctive Codec achieves a substantially higher codebook utilization rate of 4.77% despite operating at a comparable frame rate (9.5Hz) to the 10Hz frame-based model. This represents over three times better utilization of the quantization space. Moreover, the frequency distribution in Figure 4a shows that our approach exhibits a more balanced utilization pattern across codebook entries, indicating a more effective mapping of acoustic features to the discrete representation space. Conventional frame-based approaches show highly skewed distributions with a few dominant codes and many rarely-used entries, whereas our distinctive feature-based segmentation leads to a more uniform distribution. The visualization of the actual codebook usage in Figure 4b further illustrates how our approach better leverages the available codebook capacity through perceptually-guided segmentation.

These empirical findings strongly support our theoretical hypothesis: by aligning segment boundaries with natural acoustic transitions, distinctive feature-based processing produces more coherent latent representations that can be more efficiently quantized. The improved codebook utilization directly translates to better reconstruction quality and speech intelligibility as demonstrated in our main experimental results, validating the fundamental advantage of our approach over uniform frame-based processing.

TABLE IV: Comparison between FSQ and our proposed GSQ (many to many) at different reconstruction levels. Both methods are evaluated on the same Distinctive Codec architecture operating at 9.5 Hz.

Method	Level	MEL↓	STFT↓	PESQ↑	STOI↑	WER↓
FSQ	32	0.1768	0.5844	2.5467	0.8645	0.1301
FSQ	64	0.1407	0.4330	2.7909	0.8950	0.0732
GSQ	24	0.2056	0.6963	2.3578	0.8474	0.1794
GSQ	32	0.1872	0.6207	2.4732	0.8597	0.1433
GSQ	64	0.1456	0.4423	2.7848	0.8889	0.0749

B. Analysis of Group-wise Scalar Quantization Effectiveness
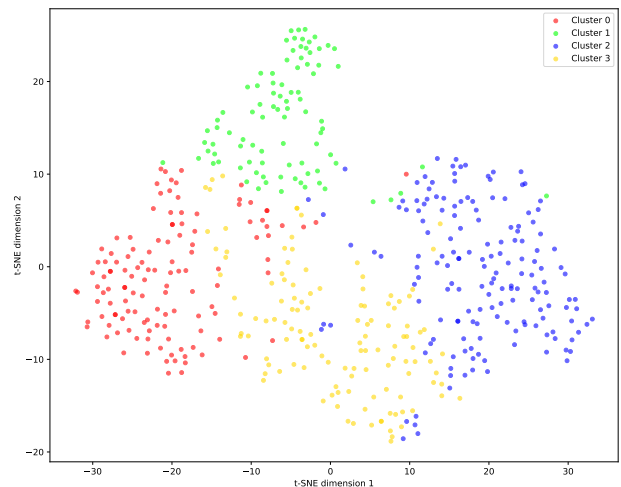
a) Comparative Analysis of Quantization Methods: Building on our theoretical analysis of distinctive features, we next investigate the effectiveness of our quantization strategy when applied to segments of varying complexity. Table IV compares our GSQ approach with standard FSQ across different reconstruction levels. While our primary results focused on compressing variable-length segments into single tokens, this analysis examines how effectively both methods preserve information at different dimensionalities. The reconstruction level represents the latent space dimension to which each segment is compressed before quantization. Our GSQ approach offers a significant advantage by dramatically reducing the effective dictionary size required for high-quality representations. Whereas direct application of FSQ to high-dimensional features would demand an extremely large dictionary, our group-wise strategy effectively partitions the feature space into multiple specialized quantization subproblems. This approach not only enhances computational efficiency but also improves representation stability for distinctive feature-based segments. Despite using a much smaller effective dictionary, GSQ achieves performance comparable to FSQ across all metrics, particularly at higher reconstruction levels.

b) Analysis of GSQ's Effectiveness Through Latent Space Geometry: We hypothesize that our GSQ approach outperforms standard quantization methods due to two key information-
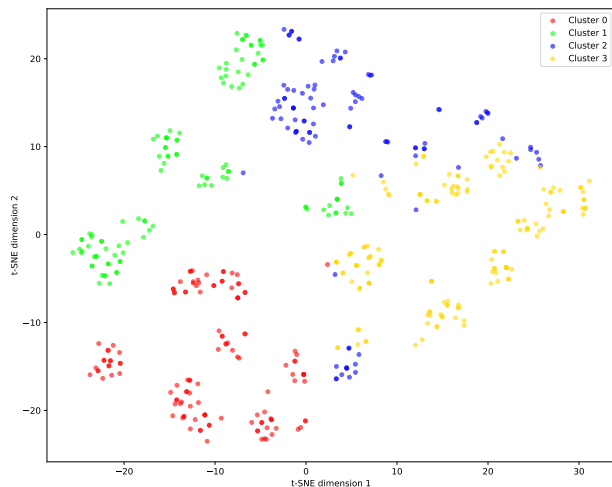
(a) FSQ latent space clustering   (b) GSQ latent space clustering

Fig. 5: t-SNE visualization [69] of latent space clusters. (a) Standard FSQ shows overlapping clusters. (b) GSQ produces more distinct clusters with clearer boundaries.

theoretic advantages. First, by projecting high-dimensional features onto specialized lower-dimensional subspaces before quantization, GSQ maximizes mutual information between input and quantized output. The learned projection matrices effectively discard redundant information while preserving essential acoustic patterns, resulting in a lower KL divergence between original and quantized distributions. Second, unlike standard FSQ which quantizes each dimension independently, GSQ's group-wise approach exploits statistical dependencies in the data, effectively aligning quantization axes with the signal's intrinsic structure. This results in quantization cells that better fit the data distribution, reducing overall distortion.

Our information-theoretic analysis suggests that if GSQ preserves more structural information, this should be reflected in the geometric organization of the latent space [70], [71]. Specifically, a quantization method that maintains higher mutual information and lower distortion should produce more coherent and well-separated clusters when visualized. To test this hypothesis, we employed t-SNE visualization [69], which maps high-dimensional data to two dimensions while preserving local neighborhood relationships, making it ideal for assessing how well structural information is maintained after quantization. Figure 5 shows striking differences between standard FSQ (Figure 5a) and our GSQ approach (Figure 5b). GSQ produces significantly more coherent clusters with clearer boundaries, indicating that GSQ preserves class structure and separability in the latent space. These results directly validate our hypothesis: GSQ's decomposition strategy prevents information mixing across feature groups, enabling it to preserves more meaningful information within severe token constraints. The improved geometric organization of GSQ's latent space explains its superior reconstruction quality across all our evaluation metrics.

C. Generalization to Out-of-Domain and Code-Switched Speech

To evaluate the generalization capability of our model beyond English and the LibriSpeech training domain, we conduct a zero-shot inference experiment on the SEAME

dataset [72], a 200-hour spontaneous Mandarin-English code-switching corpus widely used in multilingual speech research. SEAME poses three primary challenges: (i) Conversational spontaneity: frequent disfluencies (hesitations, overlaps, false starts) and colloquial prosody; (ii) Telephony distortions and environmental noise: low-bitrate channel artifacts and background sounds typical of mobile or landline settings; (iii) Cross-lingual code-switching: rapid alternation between Mandarin and English within utterances.

Without any fine-tuning, we directly apply our Distinctive Codec model trained on LibriSpeech to a subset of 50 utterances randomly selected from SEAME. The evaluation focuses on PESQ, which serves as the primary perceptual quality metric and correlates strongly with both intelligibility and signal fidelity.

Under a low token rate setting (9.5 Hz), our model achieves a PESQ score of 1.4214, significantly outperforming SpeechTokenizer's PESQ score of 1.0695 under the same configuration. This substantial improvement demonstrates that the proposed distinctive feature-based tokenization not only preserves critical acoustic cues in the English domain but also exhibits robust transferability to unseen languages and mixed-lingual acoustic conditions.

These results provide empirical evidence that Distinctive Codec retains perceptually significant information in cross-lingual and code-switched scenarios, supporting its potential as a universal speech tokenizer across diverse linguistic domains.

D. Comparison with WaveTokenizer

To further validate the effectiveness of our distinctive feature-based approach, we conducted additional experiments comparing our method with WaveTokenizer [73], another state-of-the-art neural speech codec. While our main results (Table II) demonstrate competitive performance against SpeechTokenizer, this additional comparison provides broader context for our approach within the current landscape of speech tokenization methods.

TABLE V: Performance comparison between WaveTokenizer and our Distinctive Codec at similar token rates. Lower values are better for MEL Error, STFT, and WER; higher values are better for PESQ and STOI.

Model	Token Rate	Metrics				
		MEL↓	STFT↓	PESQ↑	STOI↑	WER↓
WaveTokenizer	10 Hz	0.3139	1.0459	1.8333	0.7449	0.5535
Distinctive Codec (GSQ)	9.5 Hz	0.2006	0.6021	2.1901	0.8114	0.3061

For a fair comparison, we trained WaveTokenizer on the LibriSpeech dataset with a token rate of 10 Hz, matching the operating conditions of our Distinctive Codec (9.5 Hz). The results are presented in Table V.

Despite operating at a slightly lower token rate, our Distinctive Codec with GSQ significantly outperforms WaveTokenizer across all evaluation metrics. Specifically, our approach reduces MEL Error by 36.1%, STFT distortion by 42.4%, and WER by 44.7%, while improving PESQ by 19.5% and STOI by 8.9%. These substantial improvements further validate the effectiveness of our distinctive feature-based approach and Group-wise Scalar Quantization method.

E. Investigate the impact of semantic distillation

TABLE VI: Impact of Semantic Distillation (SD) on speech tokenization models. SD refers to the process of guiding the first RVQ layer with HuBERT representations. Lower values are better for MEL Error, STFT, and WER; higher values are better for PESQ and STOI.

Model	Frame Rate	Metrics				
		MEL↓	STFT↓	PESQ↑	STOI↑	WER↓
Distinctive Codec	9.5	0.286	1.321	1.915	0.768	0.427
Distinctive Codec + SD	9.5	0.337	1.660	1.721	0.758	**0.359**
Speech Tokenizer	10	0.314	1.046	1.833	0.745	0.554
Speech Tokenizer + SD	10	0.561	2.242	1.070	0.606	0.999
Speech Tokenizer	50	0.234	0.565	2.450	0.844	0.170
Speech Tokenizer + SD	50	0.295	0.750	2.220	0.811	**0.110**

To investigate the impact of semantic distillation on speech tokenization models, we conducted experiments comparing performance with and without distillation across different frame rates. Table VI presents these results, revealing several important insights not previously reported in the original SpeechTokenizer work.

Our findings demonstrate a consistent pattern: when semantic distillation is applied, reconstruction quality metrics (MEL Error, STFT, PESQ, STOI) tend to decrease, while speech content preservation measured by WER improves. This trade-off is evident in both our Distinctive Codec at 9.5 Hz and SpeechTokenizer at 50 Hz, where the addition of semantic distillation increases reconstruction errors but significantly reduces word error rates.

Notably, the comparison between SpeechTokenizer at 10 Hz and Distinctive Codec at 9.5 Hz highlights the superior stability of our approach at lower frame rates. While both models experience changes when semantic distillation is applied, SpeechTokenizer at 10 Hz shows dramatic degradation across all metrics. In contrast, our Distinctive Codec maintains relatively stable performance with more moderate reconstruction quality decreases and substantial WER improvements.

VII. CONCLUSION

In this paper, we investigated the effectiveness of distinctive features for speech representation in depression detection, demonstrating that adaptive segmentation aligned with acoustic boundaries preserves critical temporal dynamics. Our experiments address the fundamental limitation of fixed-rate processing that destroys timing information essential for clinical applications. Results show that distinctive feature-based tokenization produces more coherent latent representations with improved codebook utilization, while our Group-wise Scalar Quantization strategy enables stable quantization at low token rates. The substantial performance gains in depression detection (+35.0% relative improvement) validate that preserving temporal structure is crucial for capturing clinically relevant biomarkers such as pause patterns and speech rate variations. This work establishes distinctive features as a promising direction for neural speech codecs in temporally sensitive applications, with potential benefits for clinical assessment systems and other domains requiring faithful preservation of timing dynamics.

REFERENCES

[1] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[2] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[3] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[4] 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.

[5] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma *et al.*, "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024.

[6] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang *et al.*, "Cosyvoice 2: Scalable streaming speech synthesis with large language models," *arXiv preprint arXiv:2412.10117*, 2024.

[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 transactions on audio, speech, and language processing*, vol. 31, pp. 2523–2533, 2023.

[8] T. A. Nguyen, E. Kharitonov, J. Copet, Y. Adi, W.-N. Hsu, A. Elkahky, P. Tomasello, R. Algayres, B. Sagot, A. Mohamed *et al.*, "Generative spoken dialogue language modeling," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 250–266, 2023.

[9] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[10] 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 *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 15 757–15 773.

[11] L. Yu, D. Simig, C. Flaherty, A. Aghajanyan, L. Zettlemoyer, and M. Lewis, "Megabyte: modeling million-byte sequences with multiscale transformers," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023, pp. 78 808–78 823.

[12] L. The, L. Barrault, P.-A. Duquenne, M. Elbayad, A. Kozhevnikov, B. Alastruey, P. Andrews, M. Coria, G. Couairon, M. R. Costa-jussà *et al.*, "Large concept models: Language modeling in a sentence representation space," *arXiv preprint arXiv:2412.08821*, 2024.

[13] A. Pagnoni, R. Pasunuru, P. Rodriguez, J. Nguyen, B. Muller, M. Li, C. Zhou, L. Yu, J. Weston, L. Zettlemoyer *et al.*, "Byte latent transformer: Patches scale better than tokens," *arXiv preprint arXiv:2412.09871*, 2024.

[14] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2022.

[15] 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.

[16] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speechtokenizer: Unified speech tokenizer for speech language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[18] 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 transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[19] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1872–1891, 2002.

[20] Z. Huang, J. Epps, and D. Joachim, "Investigation of speech landmark patterns for depression detection," *IEEE transactions on affective computing*, vol. 13, no. 2, pp. 666–679, 2019.

[21] X. Zhang, H. Liu, Q. Zhang, B. Ahmed, and J. Epps, "Speecht-rag: Reliable depression detection in llms with retrieval-augmented generation using speech timing information," *arXiv preprint arXiv:2502.10950*, 2025.

[22] R. JAKOBSON, "Preliminaries to speech analysis: The distinctive features and their correlates," *Tech, Rep, No. 13, Acoustics Laboratory, MIT*, 1952.

[23] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3417–3430, 1996.

[24] X. Zhang, D. Liu, T. Xiao, C. Xiao, T. Szalay, M. Shahin, B. Ahmed, and J. Epps, "Auto-landmark: Acoustic landmark dataset and open-source toolkit for landmark extraction," *arXiv preprint arXiv:2409.07969*, 2024.

[25] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[27] N. Li and P. C. Loizou, "The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise," *the Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. 3947–3958, 2008.

[28] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[29] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[30] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, "Seanet: A multi-modal speech enhancement network," in *Proc. Interspeech 2020*, 2020, pp. 1126–1130.

[31] F. Mentzer, D. Minnen, E. Agustsson, and M. Tschannen, "Finite scalar quantization: Vq-vae made simple," in *The Twelfth International Conference on Learning Representations*, 2023.

[32] K.-F. Lee, *Automatic speech recognition: the development of the SPHINX system*. Springer Science & Business Media, 1988, vol. 62.

[33] D. He, X. Yang, B. P. Lim, Y. Liang, M. Hasegawa-Johnson, and D. Chen, "When ctc training meets acoustic landmarks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5996–6000.

[34] K. Ishikawa, J. MacAuslan, and S. Boyce, "Toward clinical application of landmark-based speech analysis: Landmark expression in normal adult speech," *The Journal of the Acoustical Society of America*, vol. 142, no. 5, pp. EL441–EL447, 2017.

[35] K. Ishikawa, M. Pietrowicz, S. Charney, and D. Orbelo, "Landmark-based analysis of speech differentiates conversational from clear speech in speakers with muscle tension dysphonia," *JASA Express Letters*, vol. 3, no. 5, 2023.

[36] X. Zhang, H. Liu, K. Xu, Q. Zhang, D. Liu, B. Ahmed, and J. Epps, "When llms meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 146–158.

[37] X. Zhang, B. Ahmed, and J. Epps, "Why pre-trained models fail: Feature entanglement in multi-modal depression detection," *arXiv preprint arXiv:2503.06620*, 2025.

[38] M. S. Vinton and L. E. Atlas, "Scalable and progressive audio codec," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 5. IEEE, 2001, pp. 3277–3280.

[39] J. Epps, "Wideband extension of narrowband speech for enhancement and coding," Ph.D. dissertation, UNSW Sydney, 2000.

[40] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *The Eleventh International Conference on Learning Representations*, 2023.

[41] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, "Matcha-tts: A fast tts architecture with conditional flow matching," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 341–11 345.

[42] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[43] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *arXiv preprint arXiv:2303.03926*, 2023.

[44] C. J. Cho, N. Lee, A. Gupta, D. Agarwal, E. Chen, A. W. Black, and G. K. Anumanchipalli, "Sylber: Syllabic embedding representation of speech from raw audio," *arXiv preprint arXiv:2410.07168*, 2024.

[45] F. Shen, Y. Guo, C. Du, X. Chen, and K. Yu, "Acoustic bpe for speech generation with discrete tokens," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 746–11 750.

[46] A. Baade, P. Peng, and D. Harwath, "Syllablelm: Learning coarse semantic units for speech language models," *arXiv preprint arXiv:2410.04029*, 2024.

[47] S. Cuervo and R. Marxer, "Scaling properties of speech language models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024, pp. 351–361.

[48] H. Wang, H. Wang, Y. Guo, Z. Li, C. Du, X. Chen, and K. Yu, "Why do speech language models fail to generate semantically coherent outputs? a modality evolving perspective," *arXiv preprint arXiv:2412.17048*, 2024.

[49] C. Sobin, "Psychomotor symptoms of depression," *American Journal of Psychiatry*, 1997.

[50] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal acoustic biomarkers of depression severity and treatment response," *Biological psychiatry*, vol. 72, no. 7, pp. 580–587, 2012.

[51] T. F. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity." in *Interspeech*, vol. 2, 2012, pp. 1059–1062.

[52] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[53] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[54] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," in *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*. Association for Computational Linguistics (ACL), 2021, pp. 6894–6910.

[55] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.

[56] F. Kreuk, J. Keshet, and Y. Adi, "Self-supervised contrastive learning for unsupervised phoneme segmentation," in *Proc. Interspeech 2020*, 2020, pp. 3700–3704.

[57] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, A. Gupta, X. Gu, A. G. Hauptmann *et al.*, "Language model beats diffusion-tokenizer is key to visual generation," in *ICLR*, 2024.

[58] 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," *arXiv preprint arXiv:2411.19842*, 2024.

[59] 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, pp. 27 980–27 993, 2023.

[60] 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.

[61] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014, pp. 1061–1068.

[62] Z. Du, S. Zhang, K. Hu, and S. Zheng, "Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 591–595.

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

[64] R. I.-T. P. ITU, "862.2: Wideband extension to recommendation P. 862 for the assessment of wideband telephone networks and speech codecs. ITU-Telecommunication standardization sector, 2007."

[65] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[66] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[67] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An audio language model for audio tasks," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 090–18 108, 2023.

[68] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE transactions on information theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[69] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[70] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, "On the information bottleneck theory of deep learning," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 12, p. 124020, 2019.

[71] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 19–38, 2020.

[72] D.-C. Lyu, T.-P. Tan, E. S. Chng, and H. Li, "Seame: a mandarin-english code-switching speech corpus in south-east asia," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[73] S. Ji, Z. Jiang, W. Wang, Y. Chen, M. Fang, J. Zuo, Q. Yang, X. Cheng, Z. Wang, R. Li *et al.*, "Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling," *arXiv preprint arXiv:2408.16532*, 2024.