BEST-STD 2.0: BALANCED AND EFFICIENT SPEECH TOKENIZER FOR SPOKEN TERM DETECTION

Anup Singh[⋆] *Kris Demuynck*[⋆‡] *Vipul Arora*[†‡]

[⋆] IDLab, Department of Electronics and Information Systems, imec - Ghent University, Belgium
[†] Department of Electrical Engineering, Indian Institute of Technology Kanpur, India

ABSTRACT

Fast and accurate spoken content retrieval is vital for applications such as voice search. Query-by-Example Spoken Term Detection (STD) involves retrieving matching segments from an audio database given a spoken query. Token-based STD systems, which use discrete speech representations, enable efficient search but struggle with robustness to noise and reverberation, and with inefficient token utilization. We address these challenges by proposing a noise and reverberation-augmented training strategy to improve tokenizer robustness. In addition, we introduce optimal transport-based regularization to ensure balanced token usage and enhance token efficiency. To further speed up retrieval, we adopt a TF-IDF-based search mechanism. Empirical evaluations demonstrate that the proposed method outperforms STD baselines across various distortion levels while maintaining high search efficiency.

***Index Terms*—** Bidirectional Mamba, Optimal Transport, Retrieval, Speech Tokenization, Spoken Term Detection.

1. INTRODUCTION

The rapid growth of spoken content across digital platforms has underscored the need for robust and efficient Query-by-Example Spoken Term Detection (QbE-STD) systems. However, existing systems often struggle in noisy and reverberant environments, limiting their effectiveness in real-world applications such as voice search [1].

The ASR-based approaches [2, 3, 4] represent speech using subword lattices to handle OOV terms but require highly accurate ASR models, which are difficult to train for short, low-context queries. Direct audio comparison systems [5, 6] circumvent transcription by performing direct audio-template matching using Segmental Dynamic Time Warping (SDTW) [7], but the high computational cost of SDTW hinders scalability. Other approaches [8, 9, 10] learn discriminative acoustic word embeddings, but these methods require precise word boundaries during training and inference, making them impractical for most real-world scenarios. Moreover, most existing STD systems are designed for clean acoustic environments and suffer severe degradation under noise and reverberation.

To alleviate these issues, BEST-STD [11] introduced a discrete tokenization strategy that converts speech into subword-like units, enabling fast retrieval with text-based search algorithms, eliminating boundary annotations at inference, and efficiently handling OOV terms. However, this approach still produces tokens that are sensitive to acoustic conditions and exhibit low entropy, limiting both robustness and discriminability. Existing speech tokenizers [12, 13, 14] that generate semantic tokens also remain entangled with speaker characteristics and other acoustic cues.

‡ Equal advising.

This paper introduces **BEST-STD 2.0**, a novel speech tokenizer designed to produce speaker-agnostic and noise-robust tokens. Our framework transforms input speech into a contextual frame-level embedding sequence using a bidirectional Mamba encoder [11], followed by tokenization through vector quantization [15]. To ensure consistent tokenization, we employ a self-supervised training objective that maps different utterances of the same term to the same token sequence, even under distortions like noise and reverberation. Furthermore, to address the common issue of token index collapse [16] in discrete representation learning, we reformulate token learning as a balanced clustering problem and introduce an optimal-transport regularization [17] that promotes uniform codebook usage and enhances discriminability. Empirical evaluations on the LibriSpeech and TIMIT datasets with added real-world noise scenarios show that our system outperforms existing STD baselines across a range of acoustic conditions, demonstrating improved robustness and search efficiency. By enabling text-like search capabilities over raw speech, our method provides a scalable solution for spoken content retrieval. Overall, the main contributions of our work are as follows:

- A noise-augmented training framework that produces noise-robust, speaker-agnostic speech tokens.
- A novel application of optimal transport to prevent codebook collapse and ensure balanced token utilization.
- A dedicated noise-robustness loss to further strengthen token stability under adverse conditions.
- A TF-IDF–based retrieval strategy enabling fast and efficient spoken-term retrieval.

2. PROPOSED APPROACH

The overall architecture of the proposed approach is illustrated in Fig. 1.

2.1. Model Architecture

We adopt the design choices for the Mamba feature encoder as described in [11]. The encoder $f : X \rightarrow Z$ comprises L layers of bidirectional Mamba blocks, which map the input audio feature sequence $X = \{x_1, \ldots, x_t\}$ to the corresponding contextual speech representations $Z = \{z_1, \ldots, z_t\}$. The output sequence embeddings of the final encoder layer are projected into a d-dimensional space, followed by L_2 normalization to ensure a unit norm.

2.2. Self-Supervised Learning Framework

Robust frame-level embeddings. For a given spoken term w, we consider a pair of utterances $(u, \tilde{u})$ with durations t and $\tilde{t}$ (WLOG,

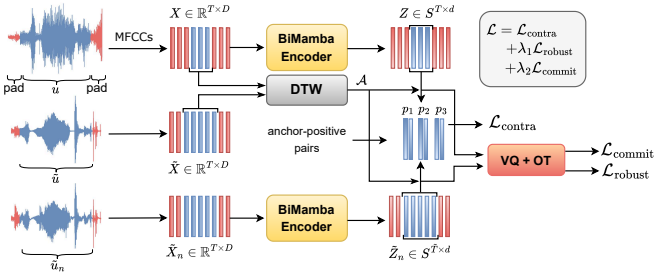

Fig. 1. Illustration of our self-supervised learning framework for robust speech tokenization.

$\tilde{t} > t$), spoken by different speakers. To simulate real-world scenarios, we add contextual padding to both utterances, ensuring they serve as fixed-length inputs to the encoder. Moreover, we stochastically add noise and reverberation to $\tilde{u}$ to generate its corresponding distorted version $\tilde{u}_n$. These padded utterances are then processed to extract their corresponding MFCC feature sequences X, $\tilde{X}$, and $\tilde{X}_n$.

We employ DTW to obtain the alignment $\mathcal{A}$ between X and $\tilde{X}$. Note that during the alignment process, frames in X and $\tilde{X}$ corresponding to the padding are excluded, yielding sequences of lengths T and $\tilde{T}$ for alignment:

$$\mathcal{A} = \{(t, \tilde{t}) \mid \tilde{t} = \arg\min_{t' \in S_t} d(X_t, \tilde{X}_{t'}), t \in [1, T], S_t \subseteq [1, \tilde{T}]\}, \tag{1}$$

where t denotes the frame index in X and S_t the set of indices in $\tilde{X}$ aligned with X_t, and $d(.)$ is the Euclidean distance function. Subsequently, X and $\tilde{X}_n$ are fed into the encoder to generate their respective embedding sequences Z and $\tilde{Z}_n$.

The DTW-based alignment $\mathcal{A}$ serves as self-supervision, enabling the creation of frame-level anchor-positive pairs:

$$p_t = (z_t, \tilde{z}_{n_{\tilde{t}}}), \text{where } (t, \tilde{t}) \in \mathcal{A} \tag{2}$$

Hence, for each training pair $(Z, \tilde{Z}_n)$, indexed by i, we define the contrastive loss function as:

$$\mathcal{L}^{(i)}_{\text{contrast}} = \frac{1}{T}\sum_{t=1}^{T} -\log\left(\frac{e^{(z_t \cdot \tilde{z}_{n_{\tilde{t}}}/\tau)}}{e^{(z_t \cdot \tilde{z}_{n_{\tilde{t}}}/\tau)} + \sum_{k=1}^{K} e^{(z_t \cdot z_k/\tau)}}\right), \tag{3}$$

where z_k are K negative embeddings randomly chosen from other training pairs in a batch, i.e., embeddings from terms $w' \neq w$.

Tokenization. We further discretize Z and $\tilde{Z}_n$ using a Vector Quantizer (VQ), denoted as $q(\cdot)$, to obtain their corresponding discrete sequences $\hat{Z}$ and $\hat{Z}_n$, respectively. The VQ utilizes a trainable codebook C consisting of K d-dimensional discrete codewords, denoted as $C = \{c_1, c_2, \ldots, c_K\}$. Specifically, the function $q : Z \to \hat{Z}$ converts a sequence of continuous encoder representations $Z = \{z_t\}_{t=1}^{T}$ into $\hat{Z} = \{\hat{z}_t\}_{t=1}^{T}$ by mapping each z_t to its nearest codeword $c_k \in C$ in terms of cosine similarity:

$$\hat{z}_t = \frac{c_{k^*}}{\|c_{k^*}\|_2}, \text{where } k^* = \arg\max_{c_k \in C}\left(z_t \cdot \frac{c_k}{\|c_k\|_2}\right), \tag{4}$$

We also add a commitment loss to ensure the embeddings are closely aligned with their corresponding discrete representations, and we define the loss for the i^{th} pair as:

$$\mathcal{L}^{(i)}_{\text{commit}} = -\frac{1}{T}\sum_{t=1}^{T} z_t \cdot \hat{z}_t \tag{5}$$

Robust discrete tokens. To ensure robustness in tokenization, we aim to map frame-level embeddings from any given anchor-positive pair (z_t and $\tilde{z}_{n_{\tilde{t}}}$) to the same codeword consistently. To achieve this, we define a robust consistency loss for the i^{th} pair as follows:

$$\mathcal{L}^{(i)}_{\text{robust}} = \frac{1}{|\mathcal{A}|}\sum_{(t,\tilde{t}) \in \mathcal{A}} \mathcal{L}(z_t, \tilde{z}_{n_{\tilde{t}}}) + \mathcal{L}(\tilde{z}_{n_{\tilde{t}}}, z_t) \tag{6}$$

where $\mathcal{L}(z_t, \tilde{z}_{n_{\tilde{t}}})$ denotes the cross-entropy loss between two distributions associated with z_t and $\tilde{z}_{n_{\tilde{t}}}$. This loss is defined as:

$$\mathcal{L}(z_t, \tilde{z}_{n_{\tilde{t}}}) = -\sum_{k=1}^{K} p(z_t|c_k) \log\left(\frac{\exp((\tilde{z}_{n_{\tilde{t}}} \cdot c_k)/\tau')}{\sum_{k'} \exp((\tilde{z}_{n_{\tilde{t}}} \cdot c_{k'})/\tau')}\right), \tag{7}$$

where $p(z_t|c_k)$ represents the probability of z_t being assigned to codeword c_k, and τ' is a parameter that controls the sharpness of the distribution. We compute $\mathcal{L}(\tilde{z}_{n_{\tilde{t}}}, z_t)$ vice versa.

Balanced codebook. The trainable codebooks are observed to suffer from the index collapse problem, i.e. $p(z|c_k)$ in Eq. 7 becomes skewed distribution. To address this, we propose a novel balanced clustering objective to ensure that the frame-level embeddings z are uniformly distributed across all codewords c_k in each training batch. The objective is formulated as follows:

$$\max_{p} \mathbb{E}_z\left[\sum_{k=1}^{K} p(z|c_k) s_k(z)\right], \text{where } s_k(z) = z \cdot \frac{c_k}{\|c_k\|_2}$$
$$\text{subject to} \quad \mathbb{E}_z[p(z|c_k)] = \frac{1}{K} \forall k \tag{8}$$

The above formulation can be interpreted as a case of the optimal transport (OT) problem, which can be effectively computed in almost linear time with the Sinkhorn-Knopp algorithm [17]. In the OT framework, we define the cost of assigning the t^{th} sample z_t to c_k as -$s_k(z_t)$. The resulting solution provides $p(z_t|c_k)$, which we use in Eq. 7.

Training objective. Finally, we train our model using the total loss $\mathcal{L}$ computed over training batch of size B as:

$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B} \mathcal{L}^{(i)}_{\text{contrast}} + \lambda_1 \mathcal{L}^{(i)}_{\text{robust}} + \lambda_2 \mathcal{L}^{(i)}_{\text{commit}}, \tag{9}$$

where λ_1 and λ_2 controls tradeoff between loss components.

2.3. Indexing and Retrieval

Given a set $\mathcal{D}$ of audio tracks, each track $a_i \in \mathcal{D}$ is divided into overlapping segments of length l s with a hop size of h s. For each segment, we compute its representation $Z = \{z_t\}_{t=1}^{T}$, followed by its corresponding tokenized representation $Z' = \{i_t\}_{t=1}^{T}$, where each i_t denotes the index of the nearest codeword for z_t in the codebook C, with $i_t \in \{1, 2, \ldots, K\}$. Subsequently, we construct a TF-IDF representation for each tokenized sequence Z' in the database and index them using IVF-PQ [18] for fast retrieval.

To further enhance retrieval efficiency and precision, we perform retrieval in multiple stages, progressively filtering candidates at each

stage to refine the results and improve precision. Given a query q, we first compute its token representation Z'_q and its corresponding TF-IDF representation. In the initial stage, we retrieve a set of candidate matches $\mathcal{P}_1$ using the index. In the second stage, we further refine $\mathcal{P}_1$ by computing the Jaccard similarity between Z'_q and $Z' \in \mathcal{P}_1$, resulting in a filtered set $\mathcal{P}_2$. Lastly, we apply an edit distance-based filtering on $\mathcal{P}_2$ to obtain the final set of best-matches $\mathcal{P}_3$. The edit distance is used to incorporate temporal information, which is absent in the Jaccard similarity computation performed in the previous stage. Our retrieval process is limited to this stage for efficiency reasons. However, to further improve precision, DTW can be applied as additional filters using the discrete representation Z' and the continuous representation Z in succession.

3. EXPERIMENTAL SETUP

3.1. Databases

We trained the model on the LibriSpeech *train-clean-360* subset and used the *test-clean* subset for validation. To construct the speech archive and queries, we selected the *train-clean-100* subset, which contains approximately 100 hours of spoken content. This setup ensures that the evaluation is always conducted on speakers unseen during training our model. To assess cross-dataset generalization, we also evaluated our model on the TIMIT dataset [19]. We employed Montreal Forced Aligner [20] to obtain word alignments for both datasets. To emulate realistic acoustic conditions, the training data were augmented with a diverse range of background noise types and reverberation effects by incorporating noise samples and room impulse responses (RIRs) from the MUSAN corpus [21]. For evaluation, we employed RIRs from the Aachen Impulse Response Database [22] and noise clips from the ETSI background noise database [23], both of which contain recordings of real-world acoustic environments such as offices, public spaces, and street settings.

3.2. Evaluation

We created two distinct query sets, each containing 100 unique terms

- In-Vocabulary – IV: This set includes terms whose text forms exist in the training data, but the query utterances are spoken by unseen speakers.
- Out-of-Vocabulary – OOV: This set contains terms whose text forms and speakers are absent from the training data.

We evaluated the spoken term detection baselines using the Mean Term Weighted Value (MTWV) [24] as the primary performance metric. Also, to assess the robustness of speech tokenizers against speaker variations and audio distortions, we measured the Jaccard similarity [25] between the tokenized representations of different utterances of the same spoken term.

3.3. Baselines

We evaluated ASR-based baselines for STD tasks, including HuBERT [12] and WavLM [13], by extracting posterior tokens from these models. Also, we evaluated speech tokenizers that generate semantic speech tokens, such as WavLM, SpeechTokenizer [14], and BEST-STD [11], for STD tasks. For WavLM, we used the pretrained KMeans model available in the SpeechBrain toolkit [26], which was trained on features extracted from the 23rd layer of the large variant of WavLM. For SpeechTokenizer [14], we extracted the semantic tokens generated by the first quantizer in its residual vector quantization (RVQ) stack.

3.4. Implementation details

The input audio segments were set to 1 s (l), covering approximately 93% of spoken terms in the database to ensure sufficient term coverage during training. We extracted 16 MFFCs along with their first and second derivatives, computed over 25ms windows with a 10ms frameshift. The encoder consisted of 8 bidirectional Mamba layers, followed by a projection layer mapping the output to a 128-dimensional embedding space, totaling 8.1M trainable parameters. Hyperparameters τ and τ' were fixed at 0.1, while λ_1 and λ_2 were set to 1 and 10, respectively. The model was trained for 740k steps using the Adam optimizer with a learning rate of 5×10^{-4} and a batch size of 96. During training, clean speech was mixed with noise at SNRs uniformly sampled from 0–10 dB. Evaluation was performed at fixed SNRs in range -5 to 20dB, including values outside the training range to assess generalisation to unseen noise levels.

4. RESULTS

4.1. Analysis of Speech Tokens

We assess token consistency using 5k cross-speaker spoken-term pairs and compute their Jaccard similarity across all acoustic conditions. Table 1 shows that our tokenizer produces the most consistent tokens across all conditions, outperforming every baseline, including K-Means tokens, ASR posterior tokens from WavLM, and the earlier BEST-STD system. Even under unseen low-SNR ($\leq$ 5 dB) and highly reverberant conditions, our tokens maintain high Jaccard similarity, whereas all baselines degrade sharply. Notably, in clean conditions, our tokens still improve similarity over BEST-STD by 10%, indicating that the gains are not limited to noise robustness but also reflect stronger speaker invariance and more efficient tokenization. These gains stem from our noise-augmented training and optimal-transport objective, which together mitigate codebook collapse and yield well-separated, discriminative token representations. This analysis provide direct evidence that our tokens retain speaker-invariant behavior in addition to noise robustness.

We further compare tokens extracted from our Transformer-based encoder, trained within the same framework, to those from the BiMamba encoder. Despite both models having similar sizes, the Transformer-based tokens exhibit lower efficiency. We attribute this performance gap to the rotary positional encoding used in the Transformer, which may capture unnecessary temporal variations, whereas the BiMamba model provides more effective temporal modeling.

4.2. Codebook analysis

We further evaluate the effectiveness of our proposed balanced clustering objective (Eq. 8) in ensuring uniform token utilization. To assess token balance, we compute the normalized entropy of the codebook C, defined as:

$$\mathcal{H}(C) = -\frac{1}{\log K} \sum_{k=1}^{K} p_k \log(p_k) \tag{10}$$

As shown in Table 3, our approach consistently achieves a normalized entropy close to 1, indicating near-perfect balance across different codebook sizes, ranging from 1024 to 4096. In contrast,

Table 1. The average Jaccard similarity (↑) between the tokenized representations of utterance pairs across various distortion conditions.

Model	Tokens	Clean	Noise					Noise+Reverb ($t_{60} = 0.7s$)				
			-5dB	0dB	5dB	10dB	15dB	-5dB	0dB	5dB	10dB	15dB
ASR Posteriors:												
HuBERT-Large [12]	32	0.73	0.46	0.59	0.67	0.71	0.72	0.24	0.37	0.49	0.58	0.64
WavLM-Large [13]	32	0.72	0.62	0.67	0.70	0.71	0.71	0.52	0.60	0.65	0.68	0.70
Speech Tokens:												
SpeechTokenizer [14]	1024	0.45	0.09	0.12	0.15	0.18	0.19	0.03	0.04	0.05	0.07	0.08
WavLM-Large [13]	1000	0.40	0.18	0.19	0.21	0.22	0.23	0.16	0.17	0.18	0.18	0.20
BEST-STD [11]	1024	0.72	0.21	0.29	0.42	0.60	0.65	0.19	0.22	0.38	0.55	0.62
Ours - Transformer	1024	0.78	0.67	0.73	0.75	0.77	0.77	0.57	0.64	0.68	0.72	0.73
BEST-STD 2.0	1024	**0.86**	**0.72**	**0.78**	**0.81**	**0.83**	**0.84**	**0.61**	**0.69**	**0.74**	**0.77**	**0.79**

Table 2. Spoken Term Detection MTWV (↑) under various distortion conditions on LibriSpeech (left) and TIMIT (right).

Model	LibriSpeech												TIMIT											
	IV						OOV						IV						OOV					
	-5dB	0dB	5dB	10dB	15dB	20dB	-5dB	0dB	5dB	10dB	15dB	20dB	-5dB	0dB	5dB	10dB	15dB	20dB	-5dB	0dB	5dB	10dB	15dB	20dB
Noise																								
ASR Posteriors:																								
HuBERT-Large [12]	0.13	0.21	0.30	0.40	0.47	0.47	0.16	0.27	0.34	0.40	0.41	0.43	0.14	0.22	0.31	0.43	0.49	0.51	0.16	0.28	0.37	0.43	0.44	0.46
WavLM-Large [13]	0.31	0.36	0.43	0.52	0.55	0.58	0.29	0.37	0.41	0.42	0.43	0.45	0.33	0.35	0.44	0.52	0.55	0.61	0.33	0.41	0.46	0.47	0.49	0.50
SpeechTokens:																								
SpeechTokenizer [14]	0.14	0.27	0.39	0.49	0.52	0.53	0.13	0.21	0.30	0.42	0.48	0.49	0.15	0.28	0.42	0.53	0.56	0.57	0.15	0.26	0.34	0.43	0.48	0.52
WavLM-Large [13]	0.17	0.34	0.40	0.53	0.55	0.55	0.17	0.25	0.35	0.43	0.47	0.49	0.19	0.38	0.44	0.57	0.59	0.61	0.19	0.29	0.35	0.46	0.47	0.51
BEST-STD [11]	0.27	0.35	0.43	0.50	0.57	0.62	0.22	0.29	0.37	0.44	0.49	0.54	0.29	0.38	0.47	0.54	0.62	0.66	0.25	0.33	0.40	0.49	0.50	0.56
Ours-Transformer	0.51	0.58	0.61	0.65	0.67	0.67	0.50	0.56	0.60	0.62	0.64	0.65	0.55	0.62	0.66	0.73	0.74	0.75	0.52	0.60	0.64	0.66	0.68	0.69
BEST-STD 2.0	**0.58**	**0.64**	**0.72**	**0.75**	**0.77**	**0.77**	**0.51**	**0.62**	**0.65**	**0.67**	**0.68**	**0.68**	**0.60**	**0.67**	**0.78**	**0.80**	**0.81**	**0.82**	**0.53**	**0.63**	**0.67**	**0.69**	**0.70**	**0.71**
Noise + Reverberation ($t_{60} = 0.7s$)																								
ASR Posteriors:																								
HuBERT-Large [12]	0.02	0.06	0.09	0.13	0.21	0.24	0.02	0.07	0.12	0.20	0.26	0.29	0.03	0.08	0.12	0.23	0.25	0.27	0.08	0.15	0.24	0.26	0.28	0.30
WavLM-Large [13]	0.11	0.18	0.24	0.30	0.32	0.36	0.15	0.22	0.29	0.31	0.35	0.37	0.12	0.21	0.23	0.35	0.37	0.39	0.18	0.24	0.31	0.35	0.39	0.41
SpeechTokens:																								
SpeechTokenizer [14]	0.03	0.05	0.11	0.14	0.18	0.20	0.02	0.04	0.06	0.11	0.13	0.16	0.05	0.12	0.18	0.19	0.23	0.23	0.07	0.11	0.14	0.18	0.21	0.23
WavLM-Large [13]	0.06	0.12	0.19	0.25	0.34	0.39	0.04	0.07	0.14	0.21	0.27	0.31	0.08	0.16	0.23	0.26	0.30	0.36	0.10	0.17	0.23	0.25	0.29	0.30
BEST-STD [11]	0.18	0.26	0.34	0.40	0.46	0.51	0.13	0.20	0.27	0.33	0.39	0.43	0.20	0.28	0.36	0.44	0.49	0.54	0.17	0.26	0.33	0.34	0.42	0.48
Ours-Transformer	0.41	0.50	0.55	0.58	0.58	0.60	0.40	0.46	0.52	0.55	0.57	0.57	0.43	0.52	0.56	0.62	0.63	0.64	0.41	0.51	0.55	0.56	0.58	0.59
BEST-STD 2.0	**0.45**	**0.53**	**0.61**	**0.67**	**0.68**	**0.68**	**0.40**	**0.50**	**0.56**	**0.58**	**0.61**	**0.62**	**0.47**	**0.54**	**0.63**	**0.67**	**0.70**	**0.71**	**0.43**	**0.54**	**0.60**	**0.61**	**0.65**	**0.66**

KMeans-based tokenization [11] yields substantially lower entropy that drops further as the codebook size increases. Furthermore, we encountered codebook collapse when employing a trainable codebook regularized with KL divergence [27], despite extensive optimization efforts. These results demonstrate that our approach effectively overcomes the codebook collapse problem, enabling stable and scalable learning of large speech-token codebooks.

Table 3. Normalized entropy of the codebook for different codebook sizes.

Models	Tokens type	1024	2048	4096
KL Divergence [27]	Learnable	0.63	0.50	0.38
BEST-STD [11]	Kmeans	0.76	0.65	0.43
BEST-STD 2.0	Learnable	**0.98**	**0.97**	**0.96**

4.3. Retrieval

Table 2 presents the retrieval performance of various methods across different noisy environments. Our approach consistently outperforms every baseline even under severe noise and reverberation, demonstrating strong robustness. ASR-based models exhibit significant retrieval failures due to transcription errors, particularly when limited context leads to misrecognitions. Notably, our method surpasses WavLM-based approaches, including both KMeans-derived tokens and ASR posteriors, despite WavLM being explicitly trained on large-scale datasets for noise-robustness. Similar trends are observed on the TIMIT database across different acoustic conditions.

While our Transformer-based model, trained within the same framework, achieves competitive results, it falls short of the Bidirectional Mamba model in noisy and reverberant settings. This performance gap can be attributed to the Transformer's less effective temporal modeling compared to Bidirectional Mamba, which processes temporal dependencies in a linear fashion, making it more suited for our context. BEST-STD also underperforms due to tokenization inefficiencies, where the lack of token diversity leads to a higher false-positive rate. In contrast, our method ensures well-separated and distinctive tokens, resulting in improved retrieval accuracy. Importantly, our approach demonstrates strong performance on OOV terms, highlighting the compositional nature of the generated tokens. This suggests that our method can generalize effectively beyond seen vocabulary, reinforcing its applicability in real-world spoken term detection scenarios.

We benchmark retrieval latency using an in-memory search on an Intel Xeon Platinum 8268 CPU. Our system retrieves the top-10 matches for a spoken query in $\sim$1.2s on average, compared to $\sim$3.4s for BEST-STD, which relies on an inverted index. This shows a roughly $3\times$ speedup. We attribute this gain to the proposed TF-IDF-based retrieval strategy, which reduces search overhead while maintaining retrieval accuracy.

Due to space constraints, additional results including the ablation study and qualitative analysis are available on our project webpage [1].

5. CONCLUSION

In this paper, we present a robust speech tokenization method that performs effectively under challenging conditions such as noise and reverberation. Furthermore, by incorporating a balanced token learning strategy, our method offers a scalable solution to the codebook collapse problem, ensuring efficient and stable training even with large codebooks. Our approach underscores a paradigm shift toward token-based speech representations that enable text-like indexing and search, thereby bridging the methodological gap between spoken content retrieval and natural language processing.

6. REFERENCES

[1] Ye-Yi Wang, Dong Yu, Yun-Cheng Ju, and Alex Acero, "An introduction to voice search," *IEEE Signal Processing Maga-*

[1] https://github.com/anupsingh15/BEST-STD2.0

zine, vol. 25, no. 3, pp. 28–38, 2008.

[2] Jonathan Mamou, Bhuvana Ramabhadran, and Olivier Siohan, "Vocabulary independent spoken term detection," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 615–622.

[3] David RH Miller, Michael Kleber, Chia-Lin Kao, Owen Kimball, Thomas Colthurst, Stephen A Lowe, Richard M Schwartz, and Herbert Gish, "Rapid and accurate spoken term detection.," in *Interspeech*, 2007, vol. 7, pp. 314–317.

[4] Dong Wang, Joe Frankel, Javier Tejedor, and Simon King, "A comparison of phone and grapheme-based spoken term detection," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4969–4972.

[5] Dhananjay Ram, Lesly Miculicich, and Hervé Bourlard, "Cnn based query by example spoken term detection.," in *Interspeech*, 2018, pp. 92–96.

[6] Dhananjay Ram, Lesly Miculicich, and Hervé Bourlard, "Neural network based end-to-end query by example spoken term detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1416–1427, 2020.

[7] TJ Tsai, "Segmental dtw: A parallelizable alternative to dynamic time warping," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 106–110.

[8] Wanjia He, Weiran Wang, and Karen Livescu, "Multi-view recurrent neural acoustic word embeddings," *arXiv preprint arXiv:1611.04496*, 2016.

[9] Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, and Lin-Shan Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," *arXiv preprint arXiv:1603.00982*, 2016.

[10] Herman Kamper, Weiran Wang, and Karen Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4950–4954.

[11] Anup Singh, Kris Demuynck, and Vipul Arora, "Best-std: Bidirectional mamba-enhanced speech tokenization for spoken term detection," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[12] 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*, vol. 29, pp. 3451–3460, 2021.

[13] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[14] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu, "Speechtokenizer: Unified speech tokenizer for speech large language models," *arXiv preprint arXiv:2308.16692*, 2023.

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

[16] Minyoung Huh, Brian Cheung, Pulkit Agrawal, and Phillip Isola, "Straightening out the straight-through estimator: Overcoming optimization challenges in vector quantized networks," in *International Conference on Machine Learning*. PMLR, 2023, pp. 14096–14113.

[17] Marco Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, 2013.

[18] Herve Jegou, Matthijs Douze, and Cordelia Schmid, "Product quantization for nearest neighbor search," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 1, pp. 117–128, 2010.

[19] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, pp. 27403, 1993.

[20] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Interspeech*, 2017, vol. 2017, pp. 498–502.

[21] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[22] Marco Jeub, Magnus Schafer, and Peter Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *2009 16th International Conference on Digital Signal Processing*. IEEE, 2009, pp. 1–5.

[23] Speech Processing Transmission, "Speech processing, transmission and quality aspects (stq); speech quality performance in the presence of background noise; part 1: Background noise simulation technique and background noise database," *Part 1: Background Noise Simulation Technique and Background Noise Database*, 1994.

[24] Sanket Shah, Satarupa Guha, Simran Khanuja, and Sunayana Sitaram, "Cross-lingual and multilingual spoken term detection for low-resource indian languages," *arXiv preprint arXiv:2011.06226*, 2020.

[25] Tie-Yan Liu et al., "Learning to rank for information retrieval," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.

[26] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al., "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[27] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.