Supplementary Material for

2716: Noise-robust and Balanced Speech Tokenization for Spoken Term Detection

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 for fast retrieval.

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.

Ablation

We evaluate the contribution of each proposed component—balanced codebook constraint and data augmentation—on the MTWV scores for the spoken content retrieval task. The results, presented in Table 1, are obtained under the same experimental setup as described in the main paper, and all evaluations are conducted using in-vocabulary (IV) queries.

The baseline model uses a parametric codebook trained end-to-end, without any balanced codebook constraint or augmentation. Introducing augmentation into the baseline model leads to consistent improvements across all SNR levels, demonstrating its effectiveness in enhancing robustness under noisy conditions (e.g., an improvement from 0.07 to 0.13 at -5 dB). On the other hand, incorporating the balanced codebook constraint yields a substantial gain in retrieval performance—even without augmentation—boosting the MTWV score from 0.42 to 0.74 in clean conditions (a 76% relative improvement) and achieving up to 4× improvement in noisy conditions (e.g., 0.07 to 0.29 at -5 dB). This significant gain can be attributed to the reduction in false alarms enabled by a more discriminative and stable tokenized representation of spoken terms. effectively addressing the codebook collapse problem. The balanced constraint mitigates the codebook collapse problem and

helps the model distinguish acoustically similar terms, including homophones, thereby enhancing retrieval precision. Finally, combining both components results in the best performance across all noise levels, confirming that our complete system is robust and effective under varying acoustic conditions.

Table 1: Impact of each proposed component on MTWV score in noisy conditions.

	-5dB	**0dB**	**5dB**	**10dB**	**clean**
Baseline	0.07	0.09	0.13	0.21	0.42
Baseline + Augmentation	0.13	0.18	0.26	0.39	0.51
Baseline + Balanced Codebook	0.29	0.44	0.58	0.64	0.74
Baseline + Balanced Codebook + Augmentation (Ours)	**0.58**	**0.64**	**0.72**	**0.75**	**0.76**

Additional Experimental Results

In Table 2, we provide results under various noise and reverberation conditions for the TIMIT database.

Table 2: Spoken Term Detection MTWV results (↑) under various distortion conditions.

Model	**IV**						**OOV**					
	-5dB	**0dB**	**5dB**	**10dB**	**15dB**	**20dB**	**-5dB**	**0dB**	**5dB**	**10dB**	**15dB**	**20dB**
Noise												
ASR Posteriors:												
HuBERT-Large	0.13	0.20	0.29	0.40	0.46	0.46	0.15	0.26	0.35	0.40	0.42	0.43
WavLM-Large	0.30	0.32	0.34	0.41	0.52	0.56	0.30	0.38	0.42	0.43	0.42	0.46
Speech Tokens:												
SpeechTokenizer	0.14	0.26	0.39	0.49	0.53	0.52	0.14	0.24	0.31	0.39	0.41	0.48
WavLM-Large	0.18	0.35	0.39	0.48	0.52	0.56	0.18	0.26	0.31	0.42	0.43	0.46
BEST-STD	0.26	0.34	0.42	0.50	0.56	0.61	0.23	0.30	0.36	0.45	0.46	0.48
Ours - Transformer	0.48	0.55	0.57	0.71	0.71	0.72	0.49	0.59	0.62	0.63	0.64	0.65
Ours - Bimamba	**0.55**	**0.61**	**0.68**	**0.71**	**0.71**	**0.72**	**0.49**	**0.59**	**0.62**	**0.64**	**0.64**	**0.65**
Noise + Reverberation ($t_{60} = 0.7s$)												
ASR Posteriors:												
HuBERT-Large	0.02	0.07	0.11	0.21	0.23	0.25	0.07	0.13	0.21	0.23	0.25	0.27
WavLM-Large	0.10	0.19	0.21	0.33	0.35	0.36	0.16	0.21	0.28	0.32	0.36	0.38
Speech Tokens:												
SpeechTokenizer	0.05	0.11	0.16	0.17	0.21	0.21	0.06	0.12	0.16	0.21	0.21	0.23
WavLM-Large	0.07	0.14	0.21	0.25	0.26	0.29	0.09	0.15	0.21	0.22	0.26	0.27
BEST-STD	0.19	0.21	0.30	0.34	0.36	0.40	0.15	0.23	0.29	0.31	0.34	0.35
Ours - Transformer	0.39	0.48	0.52	0.57	0.59	0.60	0.37	0.47	0.50	0.52	0.53	0.55
Ours - Bimamba	**0.43**	**0.50**	**0.58**	**0.62**	**0.65**	**0.66**	**0.40**	**0.50**	**0.55**	**0.56**	**0.61**	**0.62**

Qualitative Analysis

To investigate the nature of the learned tokens, we conducted a qualitative analysis by examining what each token represents. Specifically, we concatenated short audio segments (each 25ms long) that were assigned the same token. This allowed us to listen to the acoustic patterns captured by individual tokens.

Our observations reveal that these tokens are semantic in nature, consistently corresponding to syllable-like units, irrespective of speaker variation. This suggests that the learned representation captures meaningful subword structures that are more aligned with syllables than phonemes, contrary to some claims in prior work on semantic speech tokenizers.

We include example audio samples corresponding to selected tokens in the supplemental material to illustrate this behavior. These samples can be found at `https://anonymous.4open.science/r/PaperID_2716_rebuttal/`