BEST-STD: Bidirectional Mamba-Enhanced Speech Tokenization for Spoken Term Detection

Anup Singh[1,2], Kris Demuynck[1], Vipul Arora[2]
[1]Ghent University, [2]Indian Institute of Technology-Kanpur

Introduction

What is Spoken Term Detection?
The process of locating instances of a specific spoken term or phrase within an audio recording.

What are the applications?
Voice Search, Surveilance, Multimedia Retrieval, Wake Word Detection etc.

What did existing methods do?
They can be categorized into ASR-based, phonetic-based, and acoustic-based methods. Recent methods employ self-supervised learning to generate speaker-agnostic word embeddings.

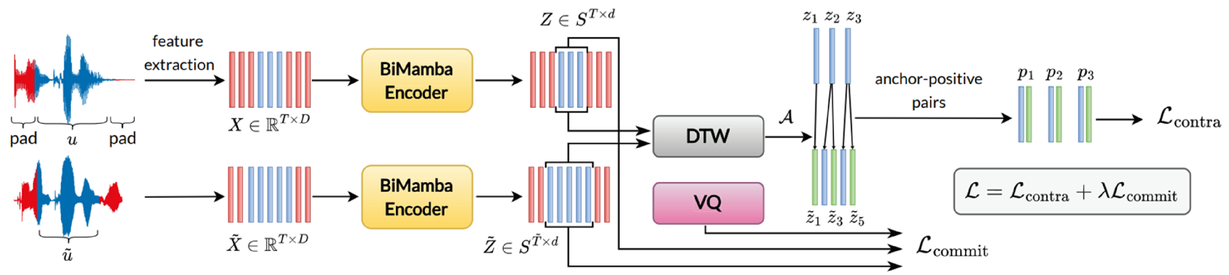
What makes it challenging?
✗ Ensuring robustness
✗ Handling pronunciation variations
✗ Handling Out-of-Vocabulary terms
✗ Efficient search in large databases
✗ Practical deployment

What do we do?
✓ Speech Tokenization for voice search.
✓ Eliminates the need for an ASR module.
✓ A novel method to generate speaker-agnostic speech tokens.
✓ Inverted index for fast and efficient retrieval.

Advantages
✓ Enables use of efficient text-based search algorithms.
✓ Eliminates need for explicit word segmentation during inference.
✓ Efficient storage of audio database.

Research Questions:
- How to learn speech tokens that are speaker-agnostic?
- How to ensure distinct and discriminative token representations for different spoken terms?

Method



Training Objectives

$$\mathcal{L}_{\text{contrast}}^{(i)} = \frac{1}{T}\sum_{t=1}^{T} -\log\left(\frac{e^{(z_t \cdot \tilde{z}_{t^*}/\tau)}}{e^{(z_t \cdot \tilde{z}_{t^*}/\tau)} + \sum_{n=1}^{N} e^{(z_t \cdot z_n/\tau)}}\right)$$

$$\mathcal{L}_{\text{commit}}^{(i)} = \frac{1}{T}\sum_{t=1}^{T} \|z_t - z_{q_t}\|_2^2 + \frac{1}{\tilde{T}}\sum_{t=1}^{T} \|\tilde{z}_t - \tilde{z}_{q_t}\|_2^2$$

$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B} \mathcal{L}_{\text{contrast}}^{(i)} + \lambda \mathcal{L}_{\text{commit}}^{(i)}$$

Self-Supervision

$$\mathcal{A} = \{(t, S_t) \mid t \in [1, T], S_t \subseteq [1, \tilde{T}]\}$$

$$p_t = (z_t, \tilde{z}_{t^*}), \text{ where } t^* = \arg\max_{j \in S_t} \cos(z_i \cdot \tilde{z}_j)$$

- **Token Consistency:** Our model is trained to generate consistent token sequences z_q and $\tilde{z}_q$ for utterance pairs $(u, \tilde{u})$ of the same spoken term w. For example, "*hello*" may tokenize as {1,1,3,3,8} and {1,1,1,3,3,3,3,8,8}.
- **Contrastive Learning:** We learn discriminative frame-level embeddings.
- **Self Supervision:** DTW-based alignment A is leveraged to generate anchor-positive pairs.
- **Vector Quantization:** K-Means-based clustering discretizes the frame-level embeddings.
- **Efficient Retrieval:** Overlapping audio chunks are tokenized and indexed using inverted index.

Results

		LibriSpeech						TIMIT					
		In-Vocabulary			Out-of-Vocabulary			In-Vocabulary			Out-of-Vocabulary		
Methods	**Tokens**	**MAP**	**MRR**	**MTWV**	**MAP**	**MRR**	**MTWV**	**MAP**	**MRR**	**MTWV**	**MAP**	**MRR**	**MTWV**
MFCC	-	0.32	0.37	0.48	0.41	0.46	0.45	0.31	0.39	0.50	0.45	0.46	0.44
Phone Posteriors	-	0.44	0.46	0.53	0.49	0.53	0.51	0.43	0.46	0.55	0.43	0.45	0.49
BNF	-	0.16	0.26	0.18	0.17	0.20	0.12	0.20	0.28	0.20	0.22	0.25	0.24
HuBERT-Base	512	0.29	0.32	0.52	0.29	0.30	0.66	0.26	0.26	0.42	0.33	0.34	0.40
HuBERT-Base	1000	0.23	0.26	0.42	0.28	0.27	0.40	0.24	0.22	0.30	0.28	0.28	0.21
WavLM-Base	512	0.44	0.49	0.57	0.44	0.45	0.60	0.38	0.38	0.48	0.40	0.41	0.44
WavLM-Base	1000	0.38	0.39	0.49	0.40	0.37	0.47	0.33	0.34	0.38	0.37	0.37	0.39
Encodec	1024	0.20	0.21	0.31	0.21	0.21	0.35	0.10	0.11	0.17	0.04	0.03	0.21
SpeechTokenizer	1024	0.57	0.62	0.56	0.53	0.53	0.65	0.46	0.48	0.46	0.45	0.46	0.45
Ours:													
Transformer	512	0.80	0.84	0.63	0.76	0.77	0.56	0.69	0.74	0.76	0.66	0.73	0.69
Transformer	1024	0.77	0.82	0.68	0.73	0.74	0.61	0.66	0.73	0.70	0.65	0.69	0.64
BEST-STD	256	**0.86**	0.90	0.62	**0.83**	**0.84**	0.55	**0.75**	**0.78**	0.69	**0.70**	**0.75**	0.63
BEST-STD	512	0.86	**0.91**	0.66	0.82	0.83	0.60	0.72	0.78	0.74	0.69	0.75	0.65
BEST-STD	1024	0.78	0.84	**0.73**	0.77	0.78	**0.65**	0.68	0.75	**0.75**	0.66	0.71	**0.70**

Results

Tokenizer	Tokens	Unigram	Bigram
HuBERT-Base	512	0.31	0.11
HuBERT-Base	1000	0.26	0.14
WavLM-Base	512	0.40	0.21
WavLM-Base	1000	0.33	0.19
Encodec	1024	0.16	0.08
SpeechTokenizer	1024	0.51	0.31
Ours:			
Transformer	512	0.74	0.64
Transformer	1024	0.71	0.60
BEST-STD	256	**0.84**	**0.77**
BEST-STD	512	0.80	0.72
BEST-STD	1024	0.78	0.69

Findings

- Our tokens are **highly robust to speaker variations.**
- Our tokens represents **subword units**.
- Our tokens **preserves compositionality** and effectively handles Out-of-Vocabulary terms.
- **Existing speech tokenizers lack robustness** against speaker variations.
- Our approach **achieves higher efficiency and efficacy** than baselines.

References

1. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R. and Mohamed, A., 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing,
2. .Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X. and Wu, J., 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing
3. Zhang, X., Zhang, D., Li, S., Zhou, Y. and Qiu, X., 2023. Speechtokenizer: Unified speech tokenizer for speech large language models.
4. Defossez, A., Copet, J., Synnaeve, G. and Adi, Y., 2022. High fidelity neural audio compression

Contact:
anup.singh@ugent.be

Code Paper