BEST-STD2.0: Balanced and Efficient Speech Tokenizer for Spoken Term Detection

Anup Singh[1], Vipul Arora[2], Kris Demuynck[1]
[1]Ghent University, [2]KU Leuven

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, Keyword Spotting, Voice Analytics

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.

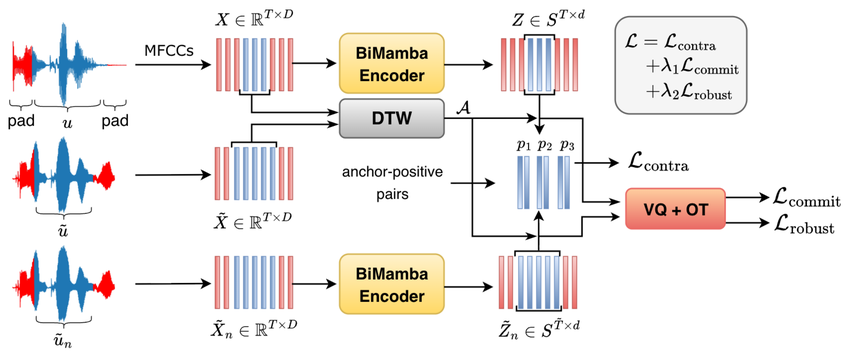
What makes it challenging?
- ✗ Ensuring robustness
- ✗ Handling pronunciation variations
- ✗ Handling Out-of-Vocabulary terms
- ✗ Efficient search in large databases
- ✗ Practical deployment

What do we do?
- ✅ Speech Tokenization for voice search.
- ✅ Eliminate the need for an ASR module.
- ✅ A novel method to generate speaker-agnostic speech tokens.
- ✅ Optimal Transport framework to learn high-entropy codebook.
- ✅ TF-IDF based search for fast and efficient retrieval.

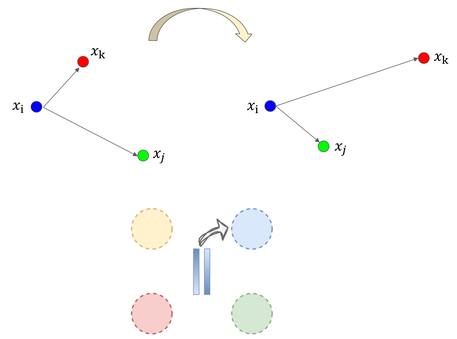
Advantages
- ✅ Enables use of efficient text-based search algorithms.
- ✅ Eliminates need for explicit word segmentation during inference.
- ✅ Efficient storage of audio database.

Research Questions:
- How to learn speech tokens that are speaker-agnostic and robust to noise and reverberation?
- How to ensure distinct and discriminative token representations for different spoken terms?
- How to maintain high-entropy codebook?

Method

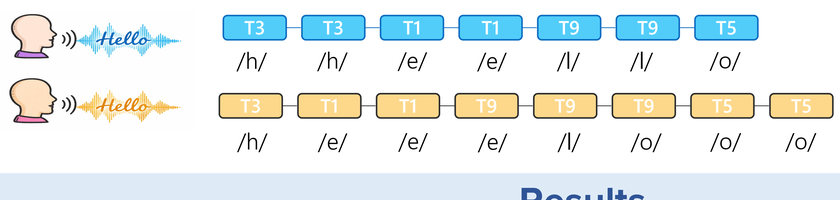


$$\mathcal{L}_{\text{contrast}}^{(i)} = \frac{1}{T}\sum_{t=1}^{T} -\log\left(\frac{e^{(z_t\cdot\tilde{z}_{n_{\tilde{t}}}/\tau)}}{e^{(z_t\cdot\tilde{z}_{n_{\tilde{t}}}/\tau)} + \sum_{k=1}^{K} e^{(z_t\cdot z_k/\tau)}}\right)$$



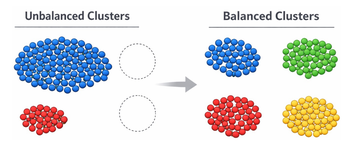
$$\mathcal{L}_{\text{commit}}^{(i)} = -\frac{1}{T}\sum_{t=1}^{T} z_t \cdot \hat{z}_t$$

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



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

Using Optimal Transport



Results

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 [11]	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 [12]	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 [13]	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 [12]	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 [10]	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 [11]	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 [12]	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 [13]	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 [12]	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 [10]	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 [11]	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 [12]	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 [13]	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 [12]	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 [10]	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**

Findings

- Our tokens are **highly robust to speaker variations and noise-robust.**
- Our tokens represents **subword units**.
- Our tokens **preserves compositionality** and effectively handles Out-of-Vocabulary terms.
- **Optimal Transport** enables learning large, **high-entropy codebooks**.

References

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

Code Paper

Contact:
anup.singh@ugent.be