Language-Agnostic Speech Tokenizer for Spoken Term Detection with Efficient Retrieval

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

TL;DR: We learn speaker and language-agnostic speech tokens for the voice search task.

Introduction

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

Motivation

- Code-switched and multilingual audio content is increasingly common.
- Existing systems are monolingual and rely on transcription.
- ASR-based methods need large labeled datasets.
- We need fast, scalable, transcription-free speech retrieval methods.

What do we do?

✓ A novel method to generate speaker-agnostic and language-agnostic speech tokens.
✓ An optimal transport-based novel method to tackle codebook collapse.
✓ TF-IDF and multistage search for fast and efficient retrieval.

Objective

Method

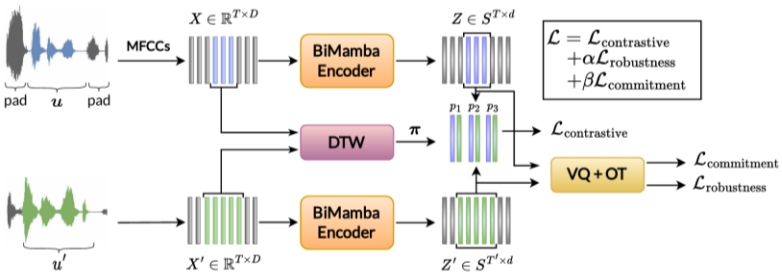
Token Consistency: Our model is trained to generate consistent token sequences z_q and z'_q for utterance pairs (u, u') of the same spoken term w.

Contrastive Learning: We learn discriminative frame-level embeddings.

Self Supervision: DTW-based alignment $\boldsymbol{\pi}$ is leveraged to generate anchor-positive pairs.

Vector Quantization: Optimal-transport-based clustering discretizes the frame-level embeddings. Allows learning high entropy codebook.

Efficient Retrieval: TF-IDF → Jaccard Similarity → Edit Distance.



Self-supervision

$$\boldsymbol{\pi} = \{(t, t') \mid t' = \arg\min_{\tilde{t} \in \mathcal{I}_t} d(X_t, X'_{\tilde{t}}), t \in [1, T], \mathcal{I}_t \subseteq [1, T']\}$$

$$p_t = (z_t, z'_{t'}), \text{where } (t, t') \in \boldsymbol{\pi}$$

Training Objectives

$$\mathcal{L}^{(i)}_{\text{contrastive}} = \frac{1}{T}\sum_{t=1}^{T} -\log\left(\frac{e^{(z_t \cdot z'_{t'}/\tau)}}{e^{(z_t \cdot z'_{t'}/\tau)} + \sum_{n=1}^{N} e^{(z_t \cdot z_n/\tau)}}\right)$$

$$\mathcal{L}^{(i)}_{\text{commitment}} = -\frac{1}{T}\sum_{t=1}^{T} z_t \cdot \hat{z}_t + z'_t \cdot \hat{z}'_t$$

$$\mathcal{L}^{(i)}_{\text{robustness}} = \frac{1}{|\boldsymbol{\pi}|}\sum_{(t,t') \in \boldsymbol{\pi}} \mathcal{L}(z_t, z'_{t'}) + \mathcal{L}(z'_{t'}, z_t)$$

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

Research Questions:

- ❑ How to learn speech tokens that are speaker-agnostic and language-agnostic?
- ❑ How to ensure distinct and discriminative token representations for different spoken terms?
- ❑ How to maintain high-entropy codebook?

Results

Model	Hindi		Marathi		Punjabi		Gujarati		Tamil		Telugu		Sanskrit	
	MTWV	Recall	MTWV	Recall	MTWV	Recall	MTWV	Recall	MTWV	Recall	MTWV	Recall	MTWV	Recall
HuBERT- Posteriors:														
Hindi	0.40	0.42	0.35	0.36	0.35	0.36	0.34	0.35	0.35	0.37	0.34	0.36	0.37	0.38
Marathi	0.34	0.35	0.34	0.35	0.33	0.34	0.41	0.43	0.33	0.34	0.32	0.33	0.39	0.41
Punjabi	0.29	0.30	0.26	0.27	0.27	0.28	0.24	0.25	0.27	0.28	0.27	0.28	0.29	0.30
Gujarati	0.46	0.48	0.39	0.41	0.37	0.39	0.43	0.44	0.38	0.40	0.37	0.38	0.40	0.41
Tamil	0.30	0.33	0.34	0.37	0.32	0.36	0.26	0.30	0.37	0.41	0.37	0.40	0.34	0.38
Telugu	0.43	0.45	0.44	0.46	0.42	0.43	0.38	0.40	0.44	0.46	0.46	0.48	0.46	0.47
Sanskrit	0.43	0.45	0.42	0.45	0.41	0.44	0.39	0.41	0.44	0.47	0.45	0.47	0.46	0.48
HuBERT-KMeans: train lang same as query lang	0.50	0.51	0.31	0.32	0.34	0.35	0.44	0.46	0.43	0.46	0.41	0.42	0.46	0.47
Ours-Transformer	0.64	0.67	**0.66**	0.68	**0.70**	0.71	0.66	0.67	0.56	0.60	**0.60**	**0.63**	0.67	0.69
Ours-Bimamba	**0.67**	**0.71**	0.65	**0.69**	0.68	**0.72**	**0.69**	**0.71**	**0.61**	**0.64**	0.59	0.62	**0.70**	**0.73**

Findings:

- ❑ Our system works well even on **unseen languages.**
- ❑ Our tokens are **language and speaker agnostic**.
- ❑ Our tokens represent **subword-level** units.
- ❑ Our system **outperforms baselines** in monolingual and multilingual settings.

- Dataset: **Kathbath -** consists of 11 Indian languages. We release the word-level alignments for the corpus.
- Evaluation performed on ~200 hours of data.

Contact: anup.singh@ugent.be

CODE DATASET