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

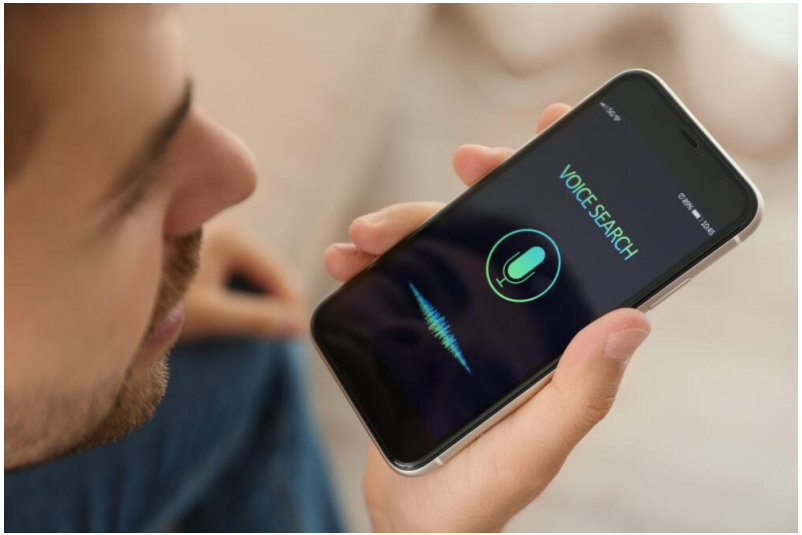
ICASSP 2025

Anup Singh[1,2], Kris Demuynck[1], Vipul Arora[2]

[1]Ghent University, [2]Indian Institute of Technology-Kanpur

VOICE SEARCH

- Speech-based technology that allows users to perform searches by speaking rather than typing.
- In general, converts spoken queries into text using ASR and retrieves relevant information.
- Common in virtual assistants (e.g., Siri, Google Assistant, Alexa) and hands-free search interfaces.



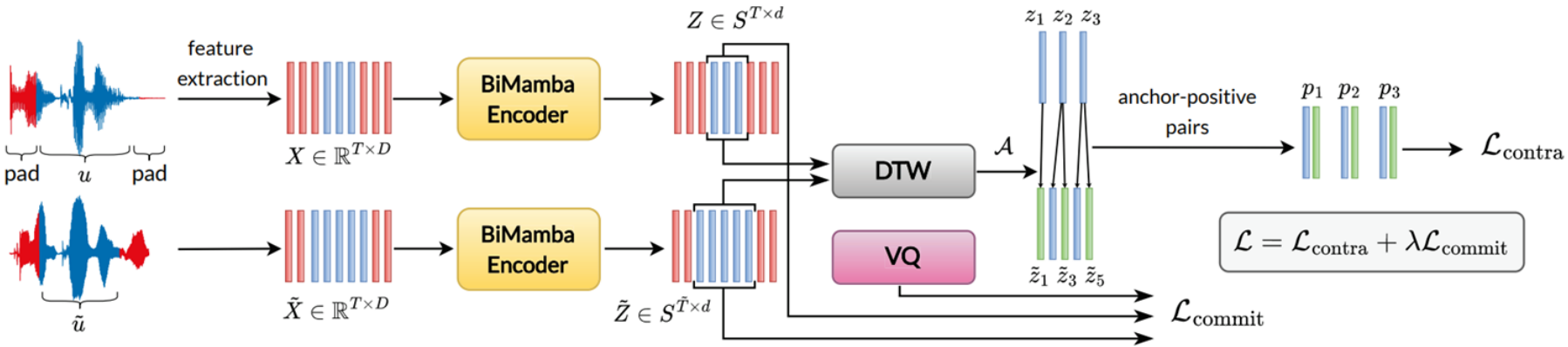
CHALLENGES

- Robustness
- High efficiency
- Less complexity

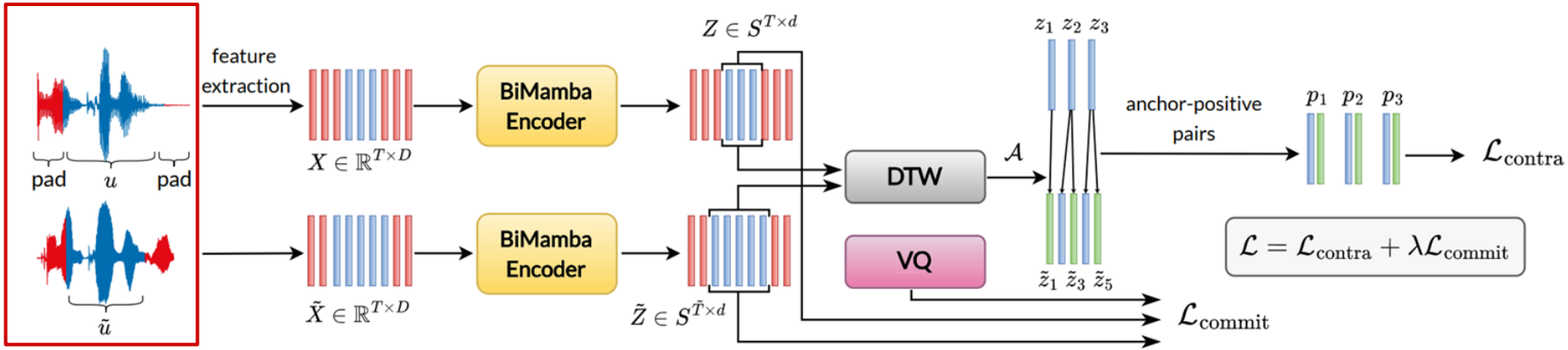
THIS WORK

- The task is to retrieve audio documents from a database that contain spoken query term.
- Eliminates the need for an ASR module.
- Proposes speech tokenization for voice search tasks.
- Introduces a novel method to generate speaker-agnostic speech tokens.
- Utilizes an inverted index for efficient and fast retrieval.

BEST-STD

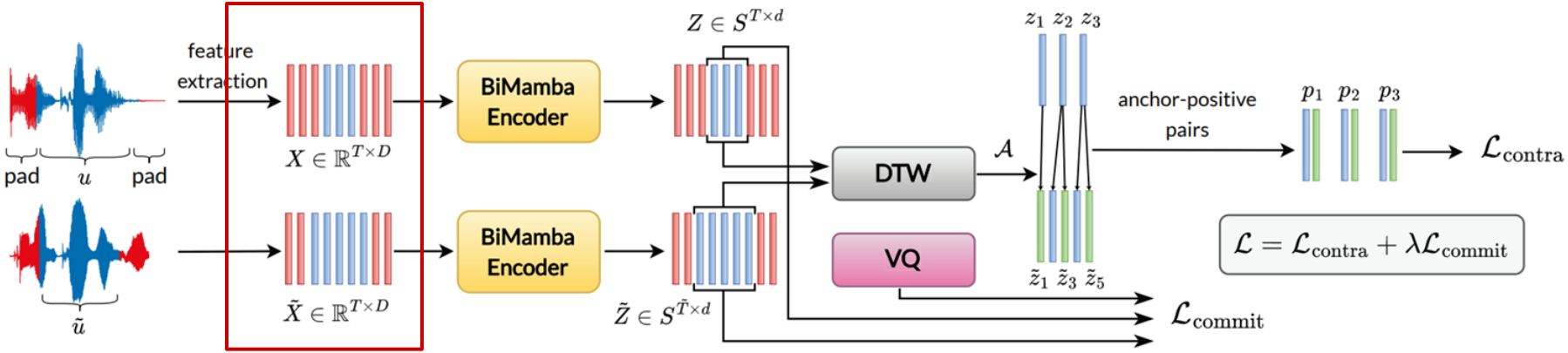

feature extraction
pad
u
pad
$X \in \mathbb{R}^{T \times D}$
BiMamba Encoder
$Z \in S^{T \times d}$
$\tilde{u}$
$\tilde{X} \in \mathbb{R}^{\tilde{T} \times D}$
BiMamba Encoder
$\tilde{Z} \in S^{\tilde{T} \times d}$
DTW
VQ
$\mathcal{A}$
z_1 z_2 z_3
$\tilde{z}_1$ $\tilde{z}_3$ $\tilde{z}_5$
anchor-positive pairs
p_1 p_2 p_3
$\mathcal{L}_{\text{contra}}$
$\mathcal{L}_{\text{commit}}$
$\mathcal{L} = \mathcal{L}_{\text{contra}} + \lambda \mathcal{L}_{\text{commit}}$

BEST-STD



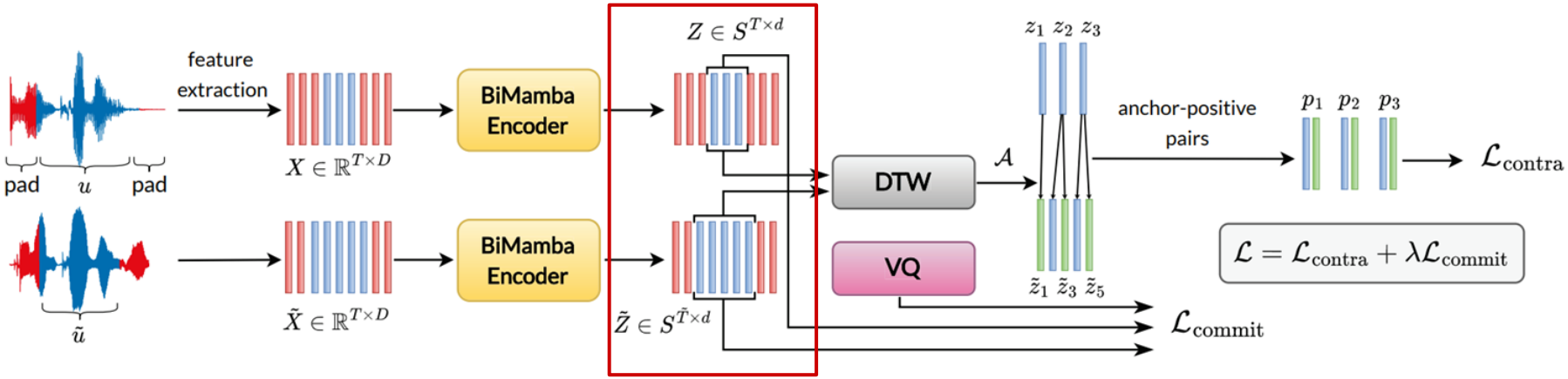
Spoken term uttered by different speakers

BEST-STD



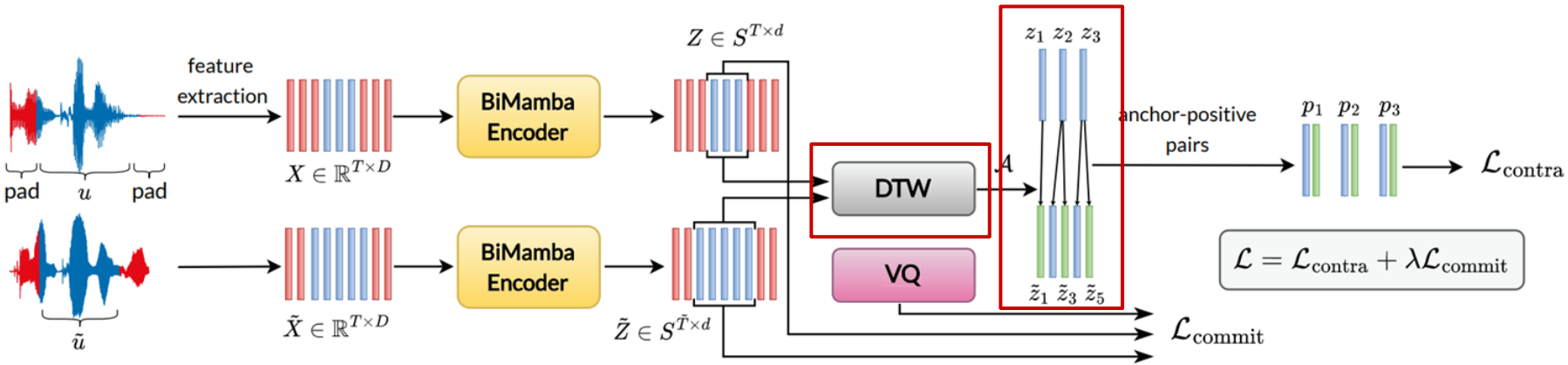
MFCCs feature extraction

BEST-STD



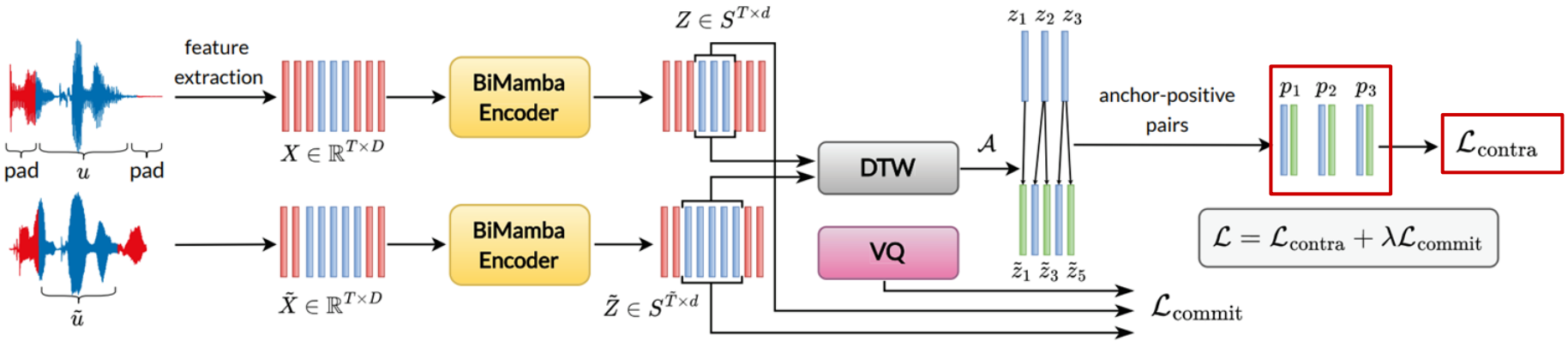
Contextual frame-level embeddings

BEST-STD



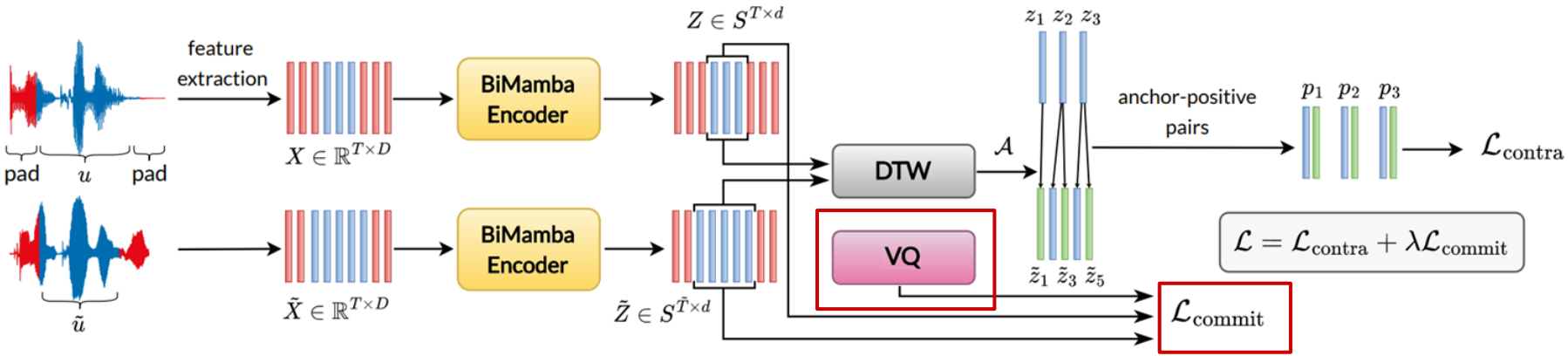
Self-supervision using DTW-based alignment

BEST-STD



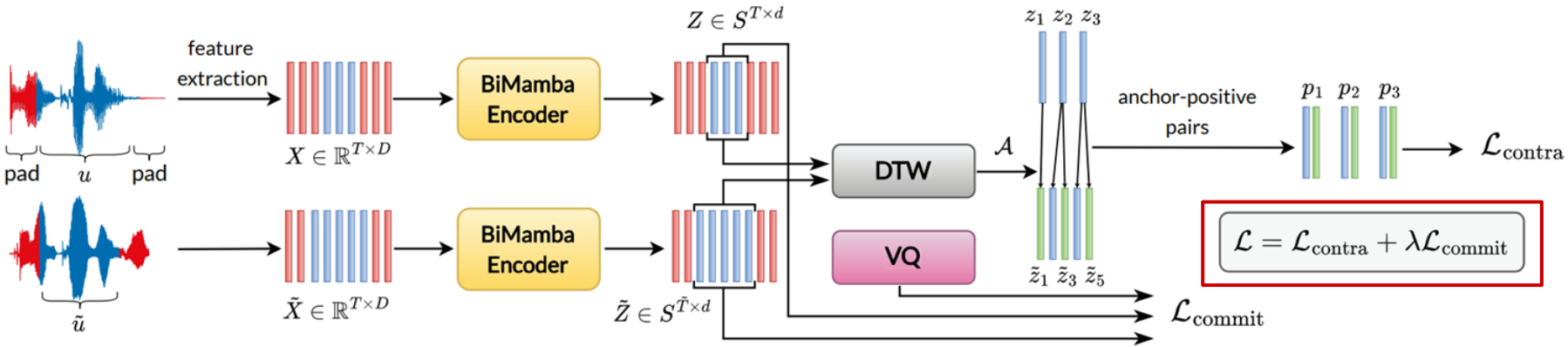
Frame-level anchor-positive pairs. Uses contrastive learning to learn discriminative frame-level embeddings.

BEST-STD



Vector Quantization to discretize frame-level embeddings. ➜ **TOKENIZATION**

BEST-STD



Final Objective: Obtain discriminative frame-level embeddings + discretization

RESULTS

TABLE I

THE AVERAGE JACCARD SIMILARITY BETWEEN DISCRETE REPRESENTATIONS OF PAIRS OF SPOKEN TERM UTTERANCES.

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

Low Jaccard similarity → **Low speaker invariant tokens**

RESULTS

TABLE I
THE AVERAGE JACCARD SIMILARITY BETWEEN DISCRETE REPRESENTATIONS OF PAIRS OF SPOKEN TERM UTTERANCES.

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

High Jaccard similarity → **High speaker invariant tokens**

RESULTS

TABLE II
SPOKEN CONTENT RETRIEVAL RESULTS (HIGHER THE BETTER) ON LIBRISPEECH TRAIN-CLEAN-100 SUBSET AND TIMIT DATASETS.

		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**

FOR MORE INFORMATION

- Join us at our poster presentation
  - Date/Time: April 08, 2025, 05:00 PM (IST)
  - Session name: Spoken and Written Document Retrieval
- Or contact me at:
  - Email: anup.singh@ugent.be
  - X: @15_anup