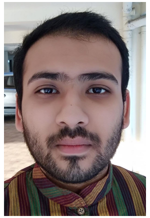
wav2tok: Deep Sequence Tokenizer for Audio Retrieval



Adhiraj Banerjee   Vipul Arora

Department of Electrical Engineering
Indian Institute of Technology Kanpur, India
{adhiraj, vipular}@iitk.ac.in

Introduction

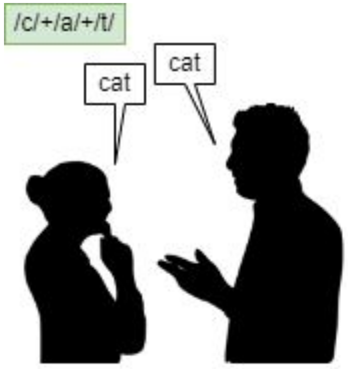
Sequence Retrieval aims at retrieving sequences similar to a query sequence, with the constraint that

an ordered alignment exists between the query and the target sequence.

Introduction

Sequence Retrieval aims at retrieving sequences similar to a query sequence, with the constraint that

an ordered alignment exists between the query and the target sequence.

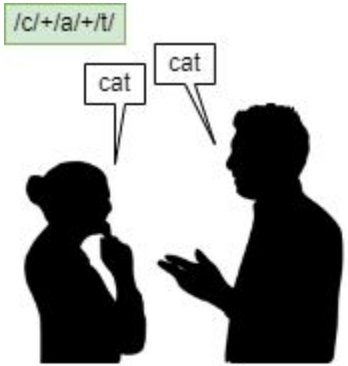


Speech Search

Introduction

Sequence Retrieval aims at retrieving sequences similar to a query sequence, with the constraint that

an ordered alignment exists between the query and the target sequence.



Speech Search

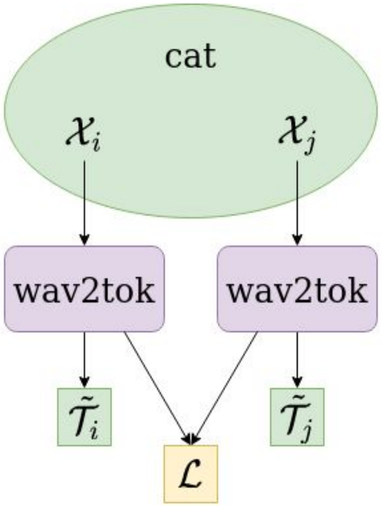
Music Search

Introduction

A model mapping audio
$\mathcal{X}$ to discrete tokens $\tilde{\mathcal{T}}$

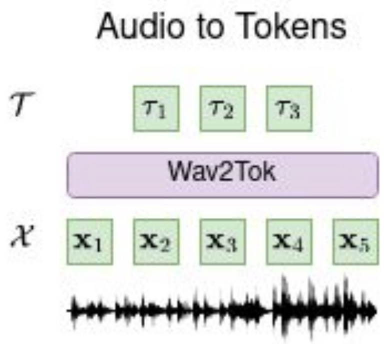
Introduction

A model mapping audio
$\mathcal{X}$ to discrete tokens $\tilde{\mathcal{T}}$



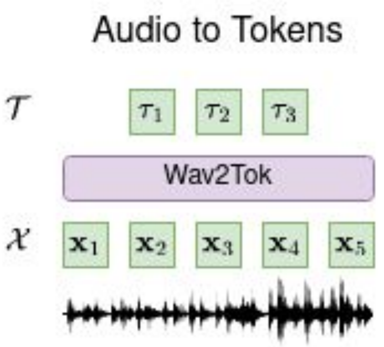
Model learns the tokens
un-supervised from pairs
of similar audio

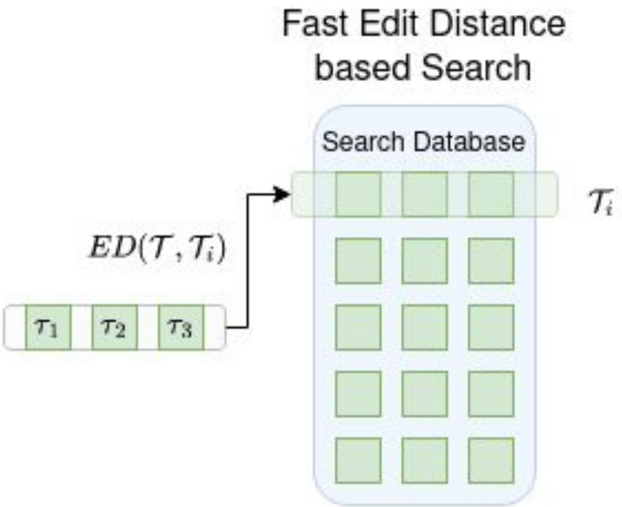
Motivation

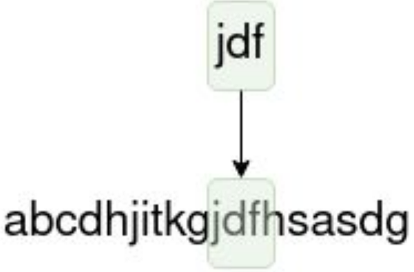


Audio to Tokens
$\mathcal{T}$
τ_1
τ_2
τ_3
Wav2Tok
$\mathcal{X}$
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$
$\mathbf{x}_4$
$\mathbf{x}_5$

Motivation







Fast approximate substring matching algorithms
for detecting substrings of tokens in a long string

Motivation

Preserving Languages
with no written form

Bo, (Tribal Language in Andaman Islands, extinct February 2010) [1]

Baghati (Solan, Himachal Pradesh) [1]

Khoisan, (Kalahari Desert, Tanzania) [2]

Sfyria (Greece) [2]

[1] https://www.tribuneindia.com/news/archive/features/when-a-language-faces-extinction-584691
[2] https://peakd.com/top10languages/@calmbrain/10-extraordinary-languages-that-do-not-involve-speaking

Motivation

Bird Language Acquisition

Distress Call ?

Hungry?

Better monitoring for wildlife conservation

Preserving Languages with no written form

Bo, (Tribal Language in Andaman Islands, extinct February 2010) [1]

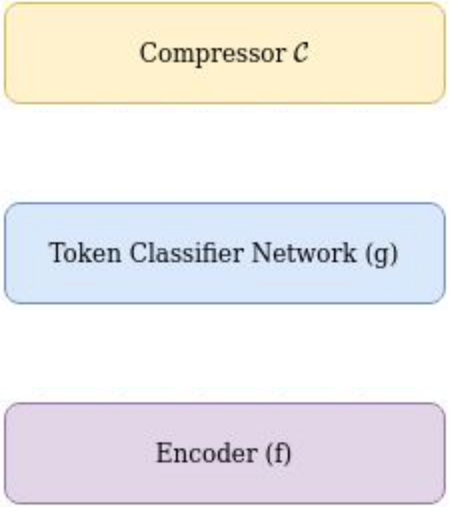
Baghati (Solan, Himachal Pradesh) [1]

Khoisan, (Kalahari Desert, Tanzania) [2]

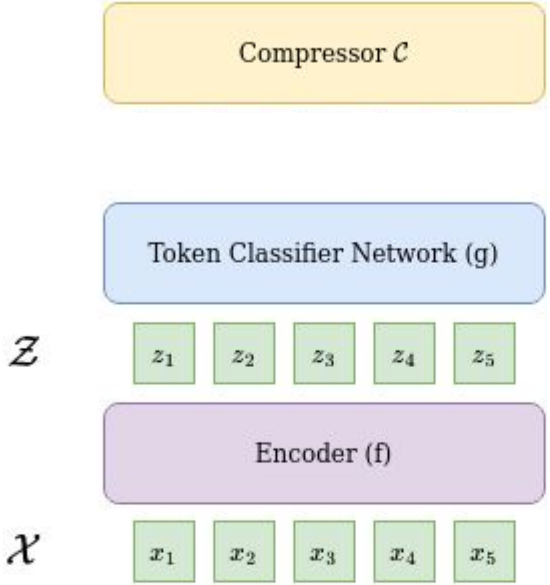
Sfyria (Greece) [2]

[1] https://www.tribuneindia.com/news/archive/features/when-a-language-faces-extinction-584691
[2] https://peakd.com/top10languages/@calmbrain/10-extraordinary-languages-that-do-not-involve-speaking

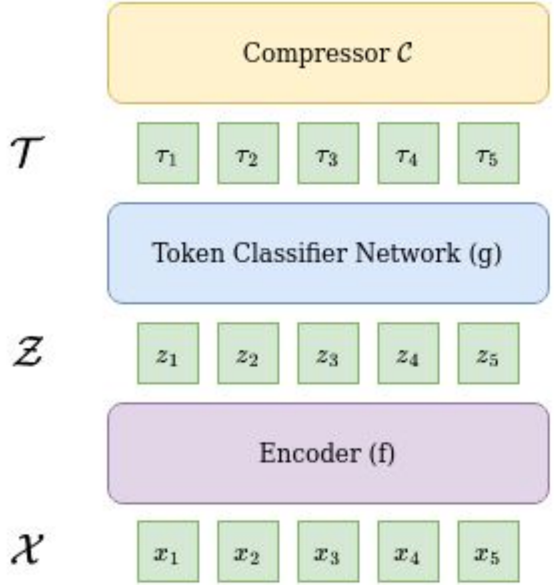
wav2tok: A Framework for learning discrete tokens



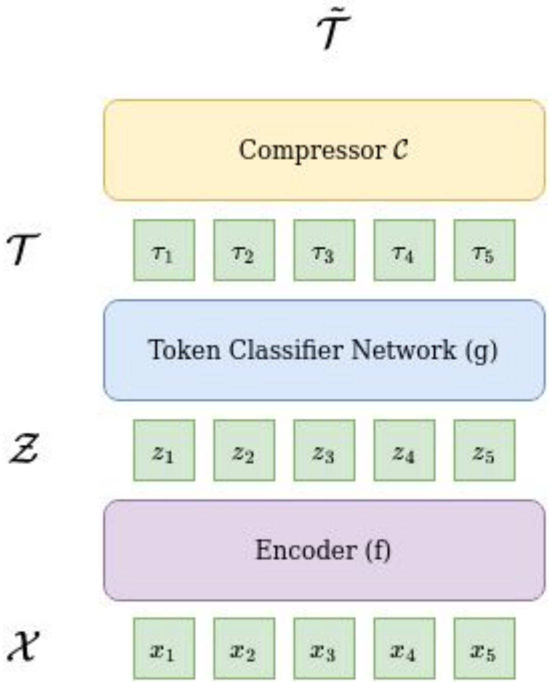
wav2tok: A Framework for learning discrete tokens



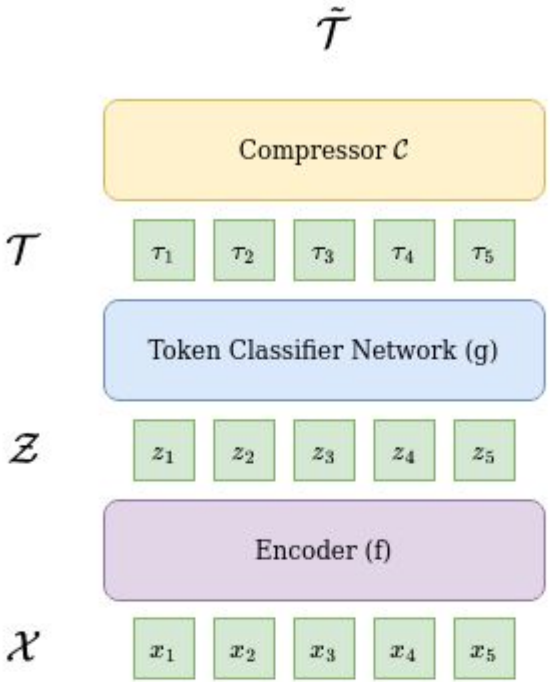
wav2tok: A Framework for learning discrete tokens



wav2tok: A Framework for learning discrete tokens



wav2tok: A Framework for learning discrete tokens



Discrete Representations

$$E = \{e_1, \ldots, e_K\}$$

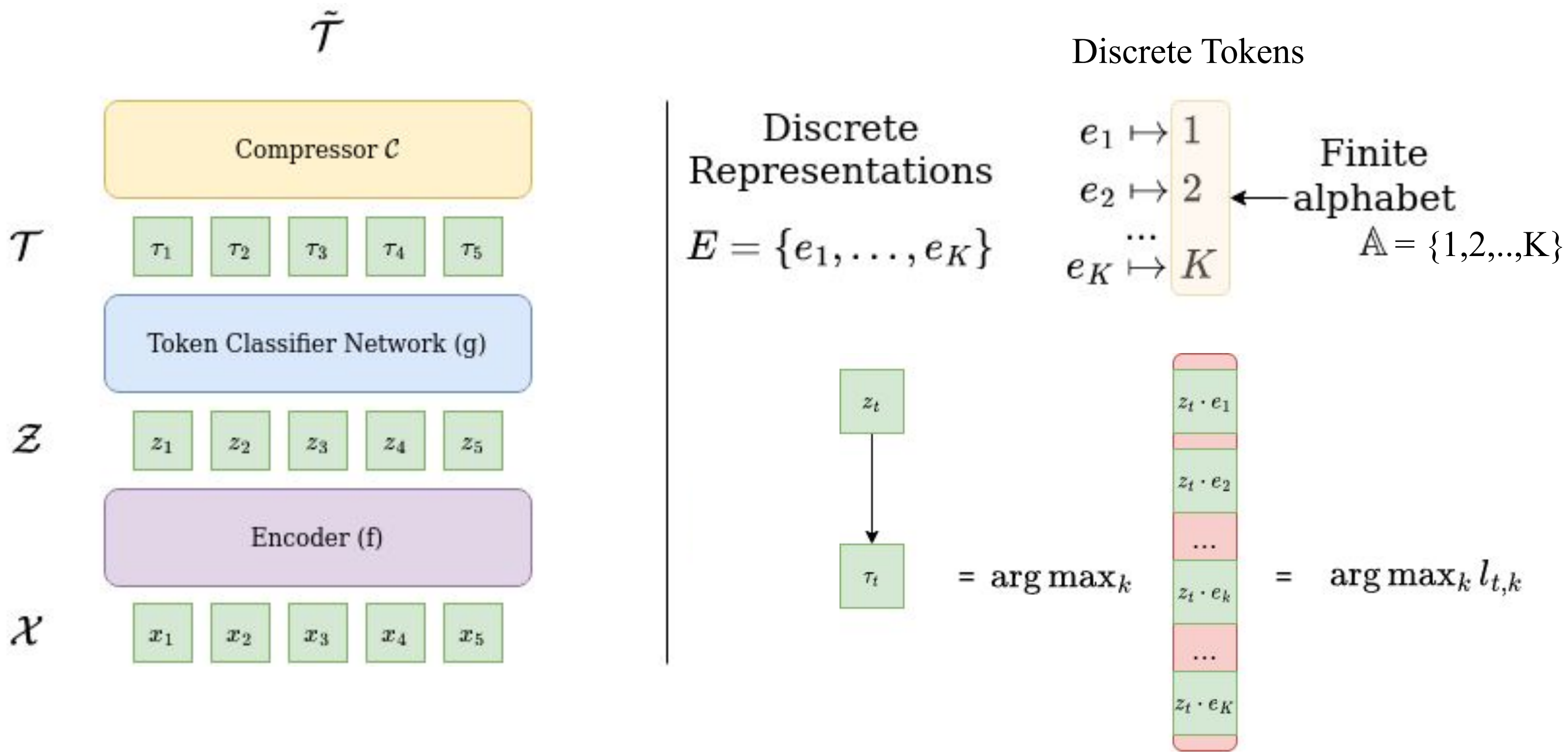
Discrete Tokens

$$e_1 \mapsto 1$$
$$e_2 \mapsto 2$$
$$\cdots$$
$$e_K \mapsto K$$

Finite alphabet

$\mathbb{A} = \{1,2,..,K\}$

wav2tok: A Framework for learning discrete tokens

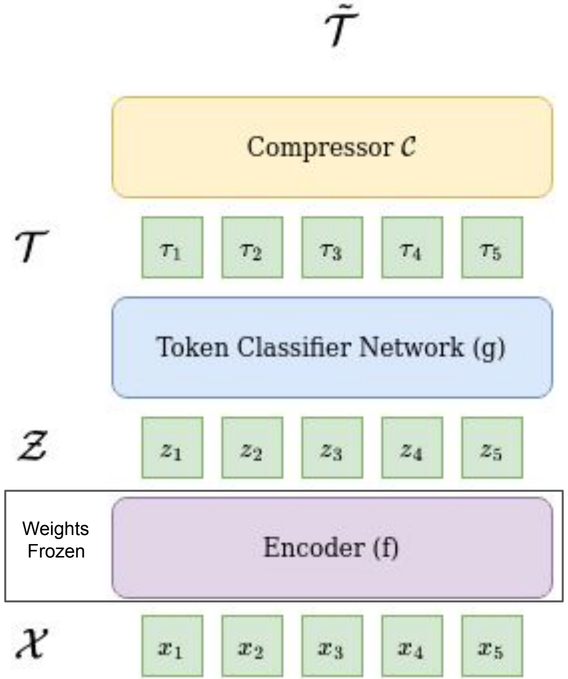


Training

Trained on pairs of similar sequences $(\mathcal{X}, \mathcal{X}')$

Expectation Maximization (EM) algorithm like training,

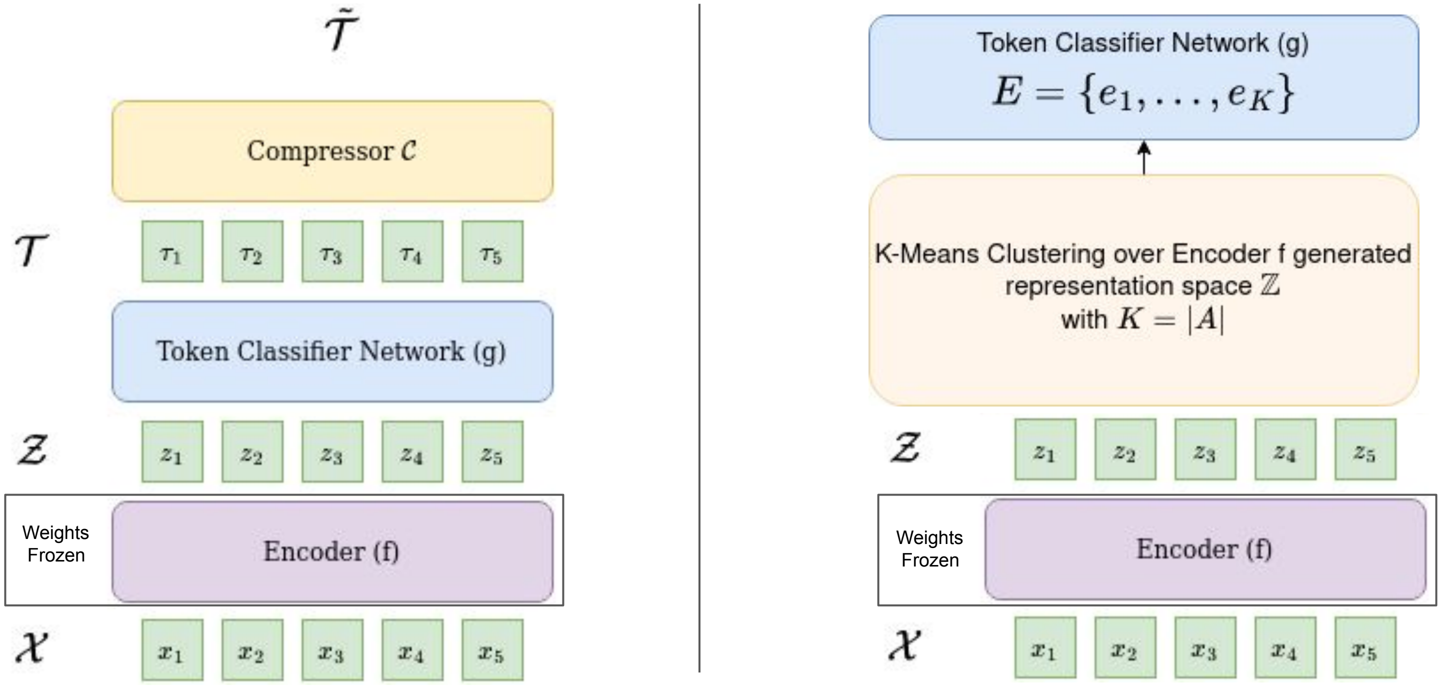
E-Step



Token Classifier Network (g)

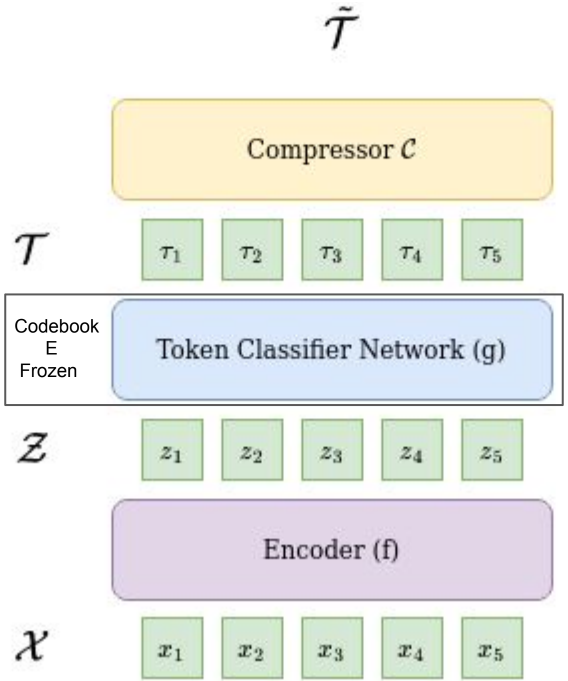
$$E = \{e_1, \ldots, e_K\}$$

E-Step



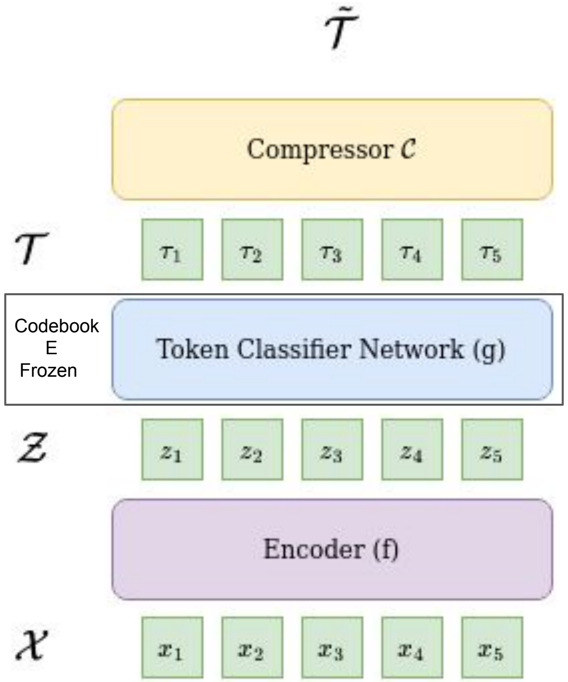
$\tilde{\mathcal{T}}$
Compressor $\mathcal{C}$
$\mathcal{T}$
τ_1 τ_2 τ_3 τ_4 τ_5
Token Classifier Network (g)
$\mathcal{Z}$
z_1 z_2 z_3 z_4 z_5
Weights Frozen
Encoder (f)
$\mathcal{X}$
x_1 x_2 x_3 x_4 x_5
Token Classifier Network (g)
$E = \{e_1, \ldots, e_K\}$
K-Means Clustering over Encoder f generated representation space $\mathbb{Z}$ with $K = |A|$
$\mathcal{Z}$
z_1 z_2 z_3 z_4 z_5
Weights Frozen
Encoder (f)
$\mathcal{X}$
x_1 x_2 x_3 x_4 x_5

M-Step



$\tilde{\mathcal{T}}$
Compressor $\mathcal{C}$
$\mathcal{T}$
τ_1 τ_2 τ_3 τ_4 τ_5
Codebook E Frozen
Token Classifier Network (g)
$\mathcal{Z}$
z_1 z_2 z_3 z_4 z_5
Encoder (f)
$\mathcal{X}$
x_1 x_2 x_3 x_4 x_5

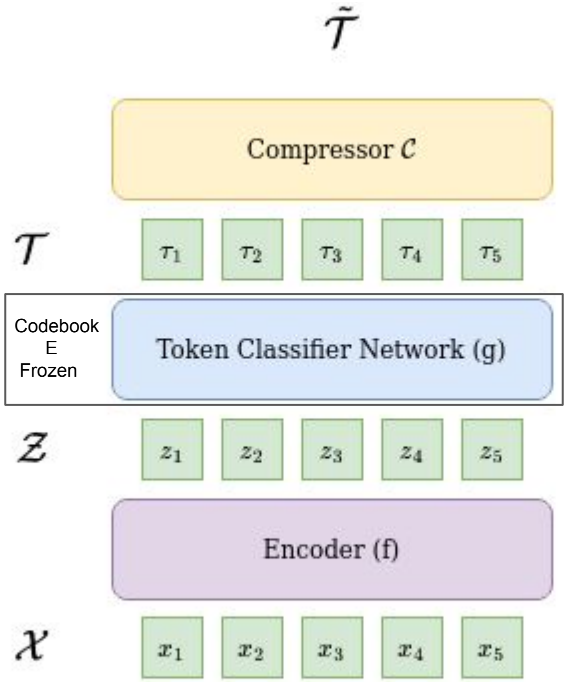
M-Step



$$\mathcal{X} \mapsto \tilde{\mathcal{T}}, \mathcal{X}' \mapsto \tilde{\mathcal{T}}'$$

$$\mathcal{L} = \mathcal{L}_m(\mathcal{X}, \mathcal{X}') + \alpha \mathcal{L}_{ctc}(\mathcal{X}, \tilde{\mathcal{T}}') + \beta \mathcal{L}_{ctc}(\mathcal{X}', \tilde{\mathcal{T}})$$

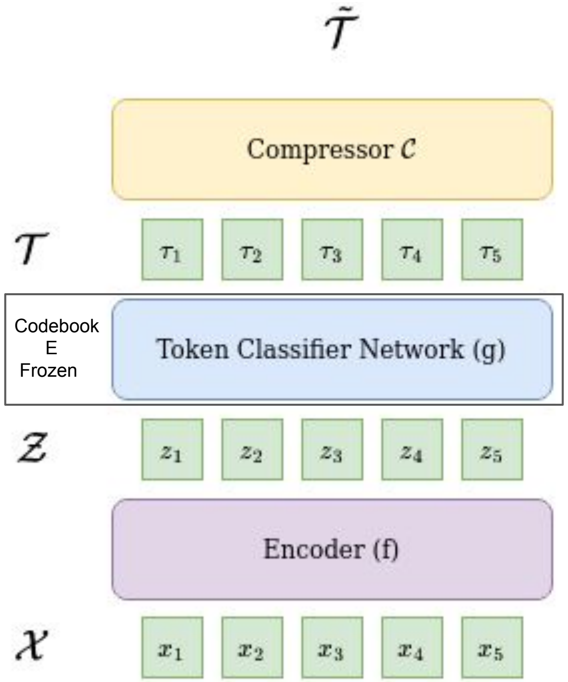
M-Step



$$\mathcal{X} \mapsto \tilde{\mathcal{T}}, \mathcal{X}' \mapsto \tilde{\mathcal{T}}'$$

$$\mathcal{L} = \mathcal{L}_m(\mathcal{X}, \mathcal{X}') + \underbrace{\alpha \mathcal{L}_{ctc}(\mathcal{X}, \tilde{\mathcal{T}}') + \beta \mathcal{L}_{ctc}(\mathcal{X}', \tilde{\mathcal{T}})}_{\text{Likelihood Loss}}$$

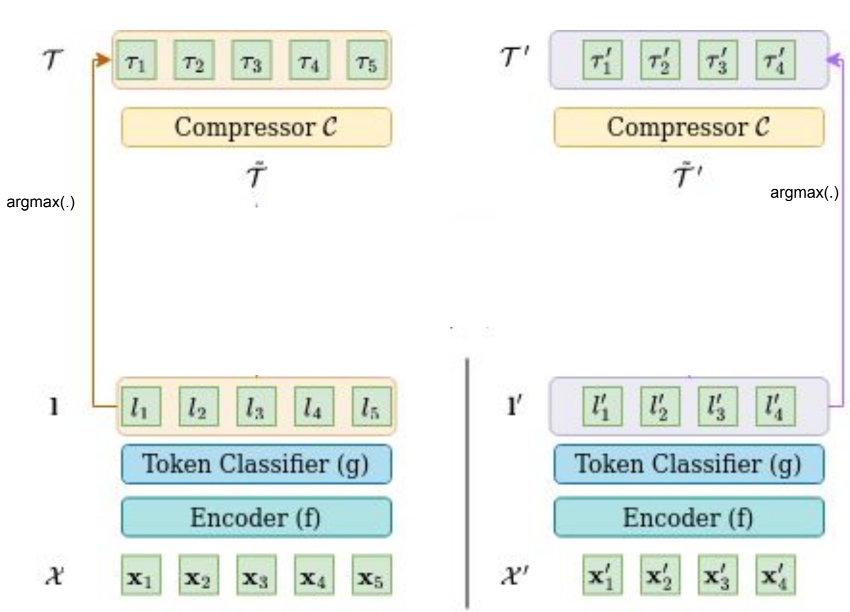
M-Step

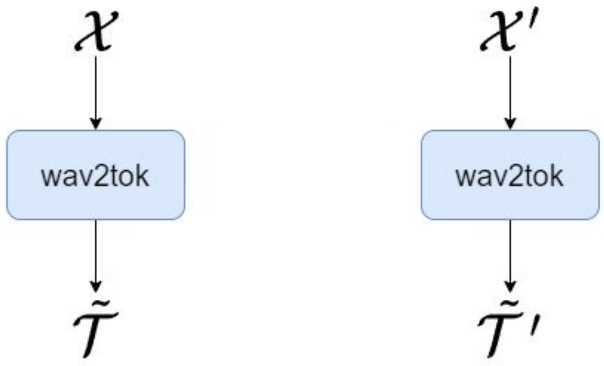


$$\mathcal{X} \mapsto \tilde{\mathcal{T}}, \mathcal{X}' \mapsto \tilde{\mathcal{T}}'$$

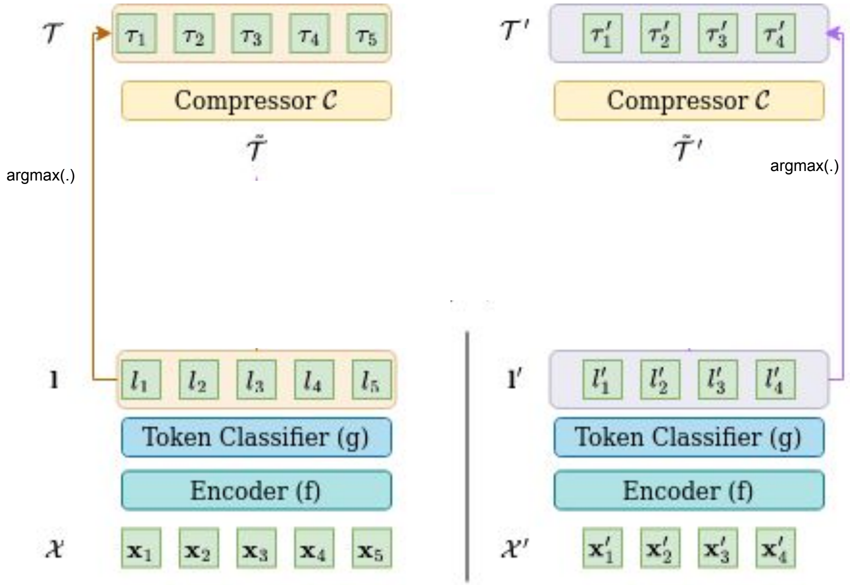
$$\mathcal{L} = \underbrace{\mathcal{L}_m(\mathcal{X}, \mathcal{X}')}_{\text{Contrastive Loss}} + \underbrace{\alpha\mathcal{L}_{ctc}(\mathcal{X}, \tilde{\mathcal{T}}') + \beta\mathcal{L}_{ctc}(\mathcal{X}', \tilde{\mathcal{T}})}_{\text{Likelihood Loss}}$$

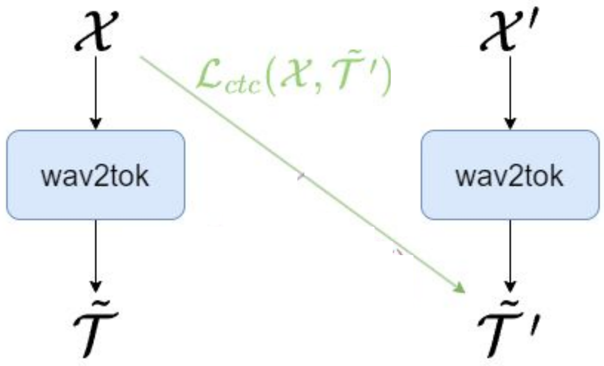
Likelihood Loss



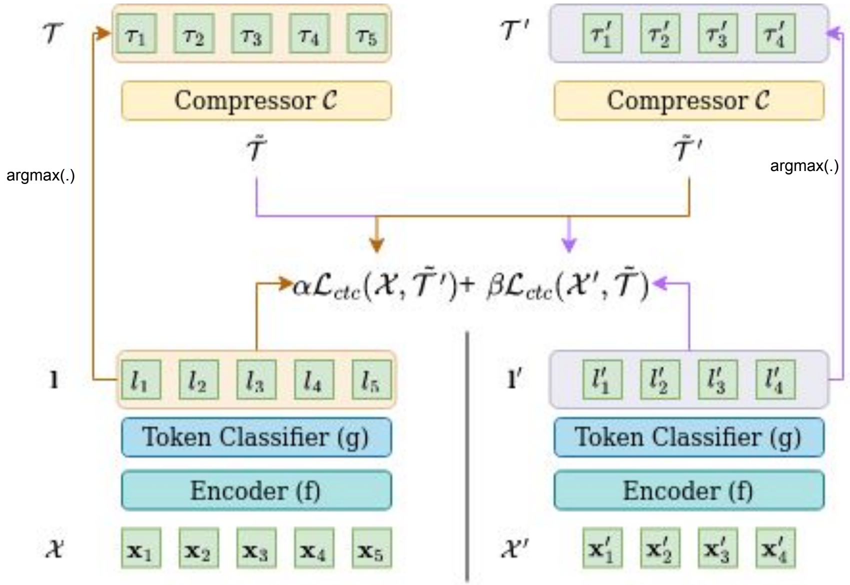


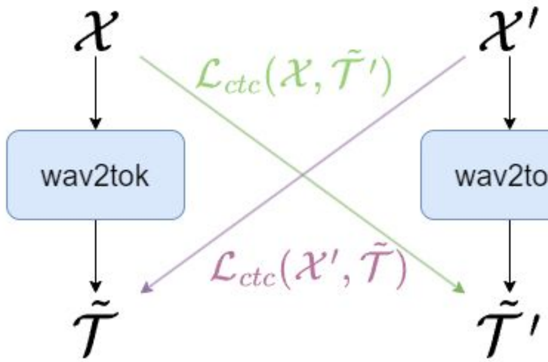
Likelihood Loss





Likelihood Loss



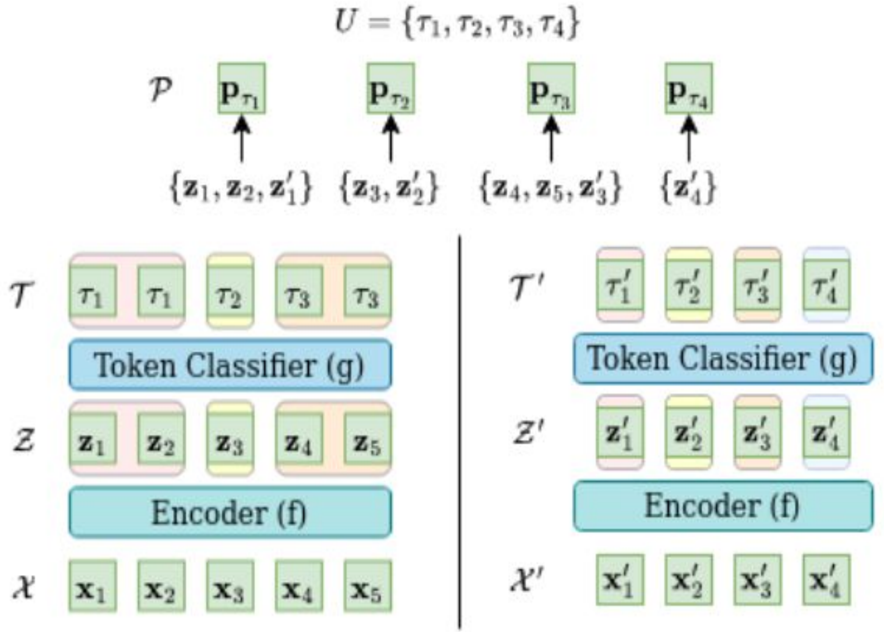


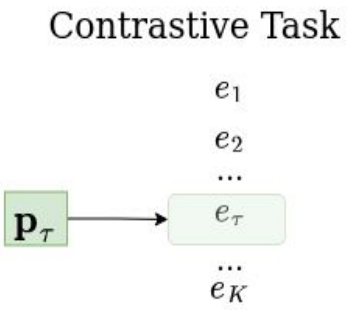
$$\alpha\mathcal{L}_{ctc}(\mathcal{X},\tilde{\mathcal{T}}')+\beta\mathcal{L}_{ctc}(\mathcal{X}',\tilde{\mathcal{T}})$$

- We use the CTC forward backward algorithm [3]

[3] Graves, Alex, and Alex Graves. "Connectionist temporal classification." *Supervised sequence labelling with recurrent neural networks* (2012): 61-93., https://www.cs.toronto.edu/~graves/icml_2006.pdf

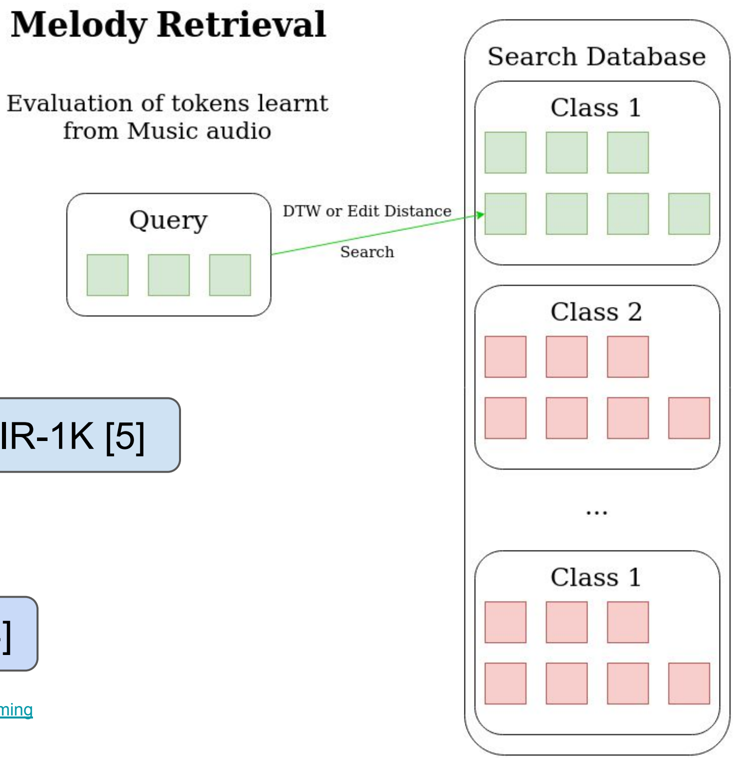
Contrastive Loss





Experiments

Experiments



Training:

MIR-QbSH [4]

MIR-1K [5]

Testing:

MIR-QbSH [4]

[4] https://www.music-ir.org/mirex/wiki/2020:Query_by_Singing/Humming
[5] http://mirlab.org/dataset/public/

Results (Music tokens)

Table 1: Quality of Tokenization

Model	V (MRR)	TS (MRR)	PS (MRR)	Search Time (s)	Infer (s)
	Vanilla query	Time Stretched query	Pitch Shifted query		
MIDI ED	0.75	0.64	0.72	3.84	0.62
Relative Note DTW	0.84	0.74	0.8	0.02	0.62
wav2vec2-O ED	0.72	0.72	0.71	0.01	0.43
wav2vec2-Multi ED	0.82	0.82	0.82		1.2
wav2tok ED	**0.84**	**0.84**	**0.84**	0.04	0.14

Results (Music tokens)

Audio to MIDI with SOTA melody extraction algorithm [6]

Table 1: Quality of Tokenization

Model	V (MRR)	TS (MRR)	PS (MRR)	Search Time (s)	Infer (s)
	Vanilla query	Time Stretched query	Pitch Shifted query		
MIDI ED	0.75	0.64	0.72	3.84	0.62
Relative Note DTW	0.84	0.74	0.8	0.02	0.62
wav2vec2-O ED	0.72	0.72	0.71	0.01	0.43
wav2vec2-Multi ED	0.82	0.82	0.82		1.2
wav2tok ED	**0.84**	**0.84**	**0.84**	0.04	0.14

[6] Justin Salamon and Emilia Gomez. Melody extraction from polyphonic music signals using pitch ´ contour characteristics. IEEE transactions on audio, speech, and language processing, 20(6): 1759–1770, 2012.

Results (Music tokens)

Audio to MIDI with SOTA melody extraction algorithm [6]

MIDI [6] to Relative Note sequence [7]

MIDI : 55,..,55,56,...,78
to
Notes: (55,0.1s), (56,0.3s), …
to
Relative notes: (0,0.1s),(+3, 0.3s),...

Table 1: Quality of Tokenization

Model	V (MRR) Vanilla query	TS (MRR) Time Stretched query	PS (MRR) Pitch Shifted query	Search Time (s)	Infer (s)
MIDI ED	0.75	0.64	0.72	3.84	0.62
Relative Note DTW	0.84	0.74	0.8	0.02	0.62
wav2vec2-O ED	0.72	0.72	0.71	0.01	0.43
wav2vec2-Multi ED	0.82	0.82	0.82		1.2
wav2tok ED	**0.84**	**0.84**	**0.84**	0.04	0.14

[6] Justin Salamon and Emilia Gomez. Melody extraction from polyphonic music signals using pitch ´ contour characteristics. IEEE transactions on audio, speech, and language processing, 20(6): 1759–1770, 2012.
[7] Naziba Mostafa and Pascale Fung. A note based query by humming system using convolutional neural network. In INTERSPEECH, pp. 3102–3106, 2017.

Results (Music tokens)

Audio to MIDI with SOTA melody extraction algorithm [6]

MIDI [6] to Relative Note sequence [7]

MIDI : 55,..,55,56,...,78
to
Notes: (55,0.1s), (56,0.3s), …
to
Relative notes: (0,0.1s),(+3, 0.3s),...

wav2vec 2.0 base model pretrained on 960h LibriSpeech and Finetuned on TIMIT [8]

Table 1: Quality of Tokenization

	Vanilla query	Time Stretched query	Pitch Shifted query		
Model	**V** (MRR)	**TS** (MRR)	**PS** (MRR)	**Search Time** (s)	**Infer** (s)
MIDI ED	0.75	0.64	0.72	3.84	0.62
Relative Note DTW	0.84	0.74	0.8	0.02	0.62
wav2vec2-O ED	0.72	0.72	0.71	0.01	0.43
wav2vec2-Multi ED	0.82	0.82	0.82		1.2
wav2tok ED	**0.84**	**0.84**	**0.84**	0.04	0.14

[6] Justin Salamon and Emilia Gomez. Melody extraction from polyphonic music signals using pitch ´ contour characteristics. IEEE transactions on audio, speech, and language processing, 20(6): 1759–1770, 2012.
[7] Naziba Mostafa and Pascale Fung. A note based query by humming system using convolutional neural network. In INTERSPEECH, pp. 3102–3106, 2017.
[8] wav2vec 2.0: A Framework for Self-Supervised ... - NIPS papers

Results (Music tokens)

Audio to MIDI with SOTA melody extraction algorithm [6]

MIDI [6] to Relative Note sequence [7]

MIDI : 55,..,55,56,...,78
to
Notes: (55,0.1s), (56,0.3s), …
to
Relative notes: (0,0.1s),(+3, 0.3s),...

wav2vec 2.0 base model pretrained on 960h LibriSpeech and Finetuned on TIMIT [8]

Multi-Lingually (53 languages) pretrained wav2vec 2.0 large model and Finetuned on Common Voice [9]

Table 1: Quality of Tokenization

	Vanilla query	Time Stretched query	Pitch Shifted query		
Model	**V** (MRR)	**TS** (MRR)	**PS** (MRR)	**Search Time** (s)	**Infer** (s)
MIDI ED	0.75	0.64	0.72	3.84	0.62
Relative Note DTW	0.84	0.74	0.8	0.02	0.62
wav2vec2-O ED	0.72	0.72	0.71	0.01	0.43
wav2vec2-Multi ED	0.82	0.82	0.82		1.2
wav2tok ED	**0.84**	**0.84**	**0.84**	0.04	0.14

[6] Justin Salamon and Emilia Gomez. Melody extraction from polyphonic music signals using pitch ´ contour characteristics. IEEE transactions on audio, speech, and language processing, 20(6): 1759–1770, 2012.
[7] Naziba Mostafa and Pascale Fung. A note based query by humming system using convolutional neural network. In INTERSPEECH, pp. 3102–3106, 2017.
[8] wav2vec 2.0: A Framework for Self-Supervised ... - NIPS papers
[9] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. CoRR, abs/2006.13979, 2020. URL https://arxiv.org/abs/2006.13979

Results (Music tokens)

Table 2: Some Variations

Model	Vanilla query **V** (MRR)	Time Stretched query **TS** (MRR)	Pitch Shifted query **PS** (MRR)
log-mel DTW	0.72	0.7	0.67
vq-**log-mel** ED	0.71	0.6	0.62
wav2tok+MIR1K ED	0.72	0.64	0.67
wav2tok ED	**0.84**	**0.84**	**0.84**

Results (Music tokens)

K-Means over log-mel features

Table 2: Some Variations

Model	Vanilla query **V** (MRR)	Time Stretched query **TS** (MRR)	Pitch Shifted query **PS** (MRR)
log-mel DTW	0.72	0.7	0.67
vq-**log-mel** ED	0.71	0.6	0.62
wav2tok+MIR1K ED	0.72	0.64	0.67
wav2tok ED	**0.84**	**0.84**	**0.84**

Results (Music tokens)

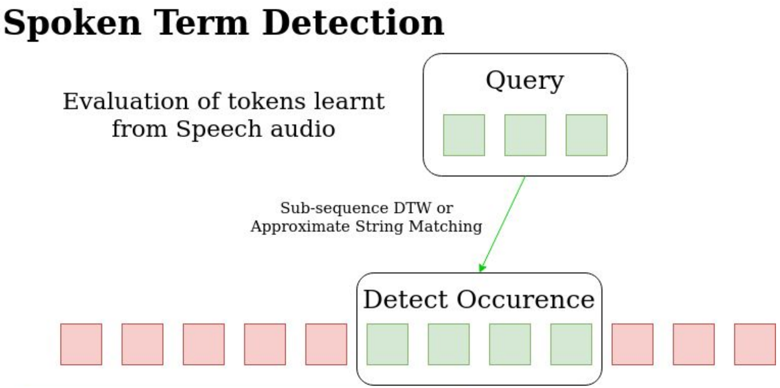
Table 2: Some Variations

Model	Vanilla query V (MRR)	Time Stretched query TS (MRR)	Pitch Shifted query PS (MRR)
log-mel DTW	0.72	0.7	0.67
vq-**log-mel** ED	0.71	0.6	0.62
wav2tok+MIR1K ED	0.72	0.64	0.67
wav2tok ED	**0.84**	**0.84**	**0.84**

K-Means over log-mel features

wav2tok trained on MIR-1K polyphonic music audio

Experiments



Training:

TIMIT [9]

LibriSpeech [10]

Testing:

TIMIT [9]

[9] Garofolo, John S., et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium, 1993.
[10] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.

Results (Speech tokens)

Table 4: Spoken Term Detection

Model	**ED** (F1)	**Search Time** (s)	**DTW** (F1)	**Search Time** (s)
log-mel DTW	-	-	0.41	0.003
wav2vec2-O	0.61	0.29	0.43	0.23
wav2vec2-Multi	0.63	0.72	0.48	0.66
wav2tok	**0.65**	**0.064**	**0.52**	**0.09**
wav2tok+Libri	0.63		0.44	0.1

[11] Xavier Anguera and Miquel Ferrarons. Memory efficient subsequence dtw for query-by-example spoken term detection. In 2013 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6, 2013. doi: 10.1109/ICME.2013.6607546.

Results (Speech tokens)

Spoken Term Detection system based on [11]

S-DTW over Posteriorograms generated by SOTA ASRs

Table 4: Spoken Term Detection

Model	ED (F1)	Search Time (s)	DTW (F1)	Search Time (s)
log-mel DTW	-	-	0.41	0.003
wav2vec2-O	0.61	0.29	0.43	0.23
wav2vec2-Multi	0.63	0.72	0.48	0.66
wav2tok	**0.65**	**0.064**	**0.52**	**0.09**
wav2tok+Libri	0.63		0.44	0.1

[11] Xavier Anguera and Miquel Ferrarons. Memory efficient subsequence dtw for query-by-example spoken term detection. In 2013 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6, 2013. doi: 10.1109/ICME.2013.6607546.

Thank You!