Topological Data Analysis for Speech Processing

Eduard Tulchinskii[1], Kristian Kuznetsov[1], Laida Kushnareva, Daniil Cherniavskii[3], Serguei Barannikov[1,2], Irina Piontkovskaya, Sergey Nikolenko[4], Evgeny Burnaev[1,3]

[1]Skolkovo Institute of Science and Technology (Russia), [2]CNRS, Université Paris Cité (France),
[3]AIRI, Artificial Intelligence Research Institute (Russia),
[4]St. Petersburg Department of the Steklov Institute of Mathematics (Russia)

INTERSPEECH 2023
20th-24th August 2023 - Dublin, Ireland

Introduction

We apply methods of Topological Data Analysis (TDA) to study the inner representation of transformer models to create efficient topologically-backend solutions for various tasks of Speech Processing and explore the inner-works of the attention.

Natural Language Processing was revolutionized by transformers – models without any recurrent parts, instead relying on the attention mechanism. They also have shown outstanding performance for many tasks in speech processing and other domains. However:

- For some tasks, usage of transformer internal representations yields significantly better performance than the traditional approach (via final embeddings).
- Attention allows model to focus on specific parts of data and learn complex dependencies. It is intended to emulate human cognitive attention, yet attention maps are hard to analyze.

Background: Topological Data Analysis

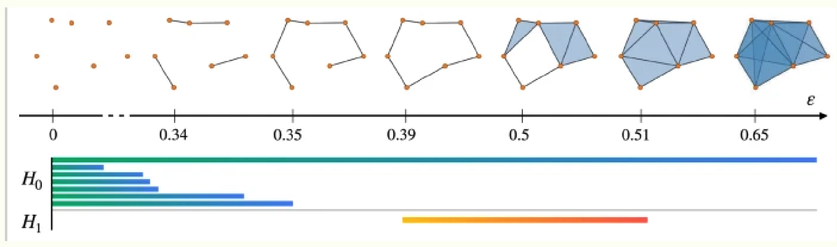

Fig. 1: Building the persistence barcode from the data.

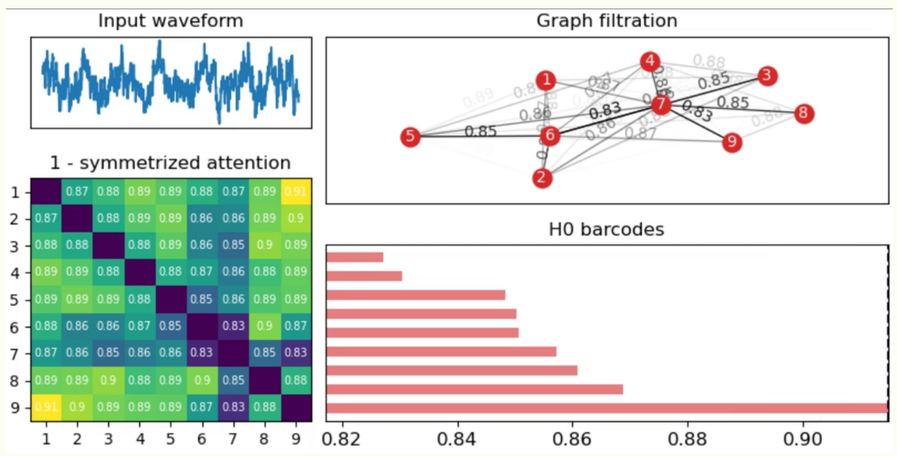

Fig. 2: Our pipeline: from waveform to attention maps via transformer model (e.g. HuBERT, Wav2Vec2), then graph filtration and barcode.

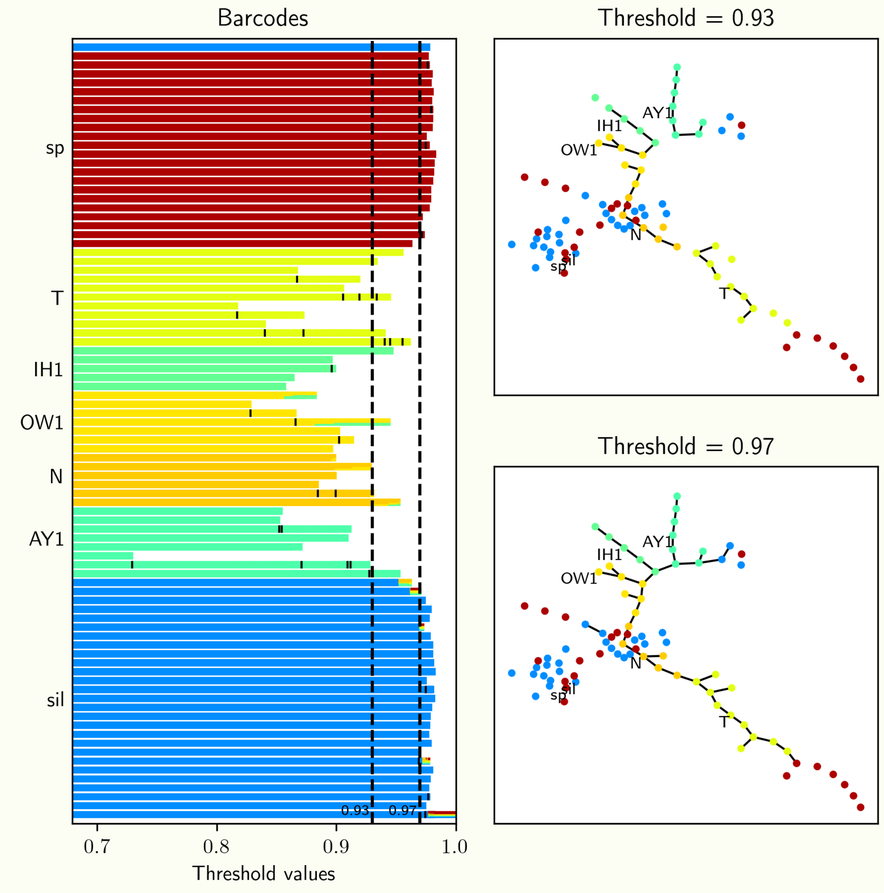

Fig. 3: H_0-barcode reflects the hierarchial structure of data. Sample H_0-barcode and MST on different levels for head (2, 4); speech sample text: "I know it", sample phones: "sil AY1 N OW1 IH1 T sp". Nodes and bars are colored with respect to the phonemes they represent. Black dashed lines show barcode levels corresponding to the trees on the right. Separators inside the bars show levels where nodes from the same phoneme are joined to the bar's component.

Our approach

- TDA methods are known to capture well **surface and structural patterns** in different types of data; they are also known to be more robust to noise in data.
- The attention maps generated by the Transformer model can be represented as **weighted graphs** and further efficiently investigated with TDA.
- We take a pretrained model, extract topological features from its attention maps and embeddings, then use those features to build a linear classifier
- Three groups of features: algebraic features of attention matrices (e.g. means of 3 largest diagonals, measure of asymmetry), topological features of attention matrices, and topological features of embedding.

Main results

- Linear classifier built on top of topological features outperforms a fine-tuned classification head
- New state-of-the-art performance result for emotion recognition on CREMA-D
- Topological features are able to reveal functional roles of transformer heads

Model	IEMOCAP Acc ↑	CREMA–D Acc ↑	ASV Spoof EER ↓	Vox Celeb1 EER ↓	FSDD EER ↑
HuBERT (baseline)	64.92*	71.047 ±0.566	6.649	**7.45**	<u>99.3</u> ±1.3
All layer embs, 1st	65.612 ±1.050	71.320 ±0.479	2.706	46.240	96.0 ±0.7
All layer embs, mean	69.355 ±1.801	76.260 ±1.148	**1.519**	<u>8.46</u>	97.7 ±0.5
Attention features	<u>69.666</u> ±1.174	<u>79.200</u> ±1.240	2.138	30.326	98.7 ±0.8
Attn. & non-attn. feat.	**69.955 ±0.972**	**80.155 ±0.680**	<u>1.946</u>	26.443	**99.6 ±0.4**

Tab. 1: Experimental results. Accuracy (Acc) and equal-error-rate (EER) are reported

Studying the individual attention heads

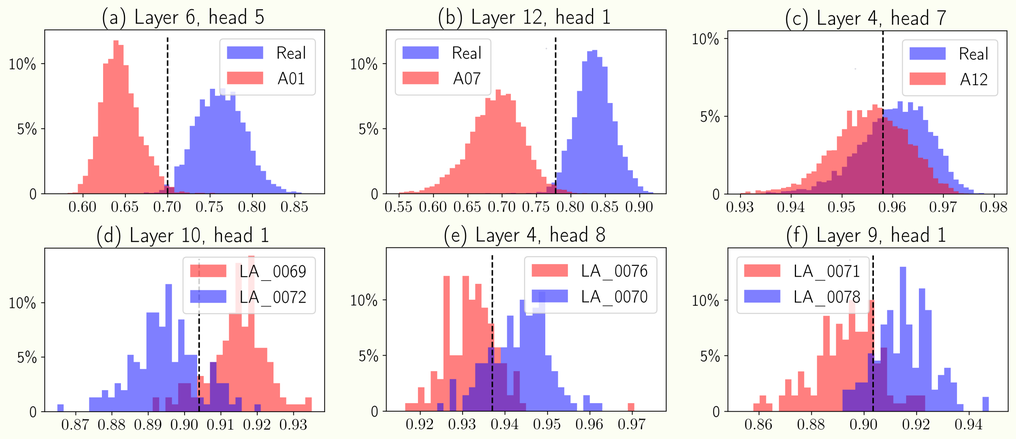

Fig. 4: Average length of bar in H_0 barcode for the best HuBERT attention heads for two tasks: (a-c) individual model separation between human (blue) and synthetic (red) speech; (d-e) speaker separation; LA_0069/72/78 – female speakers, LA_0070/76/71 – male speakers.

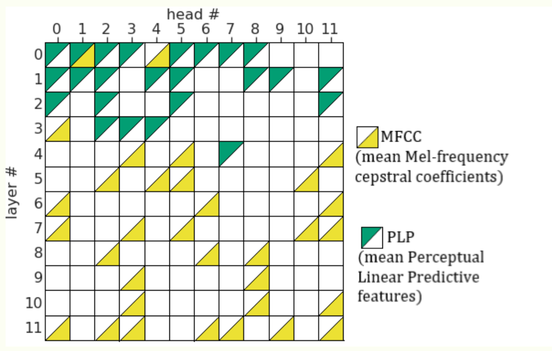

Fig. 5: Heads with high correlation between $H_0^{m,pc}$ and MFCC/PLP

More pictures...

... and some additional information can be found on the project's website
https://topohubert.github.io/speech-topology-webpages/