HAL open science

Topological Data Analysis for Speech Processing

Eduard Tulchinskii, Kristian Kuznetsov, Daniil Cherniavskii, Serguei Barannikov, Sergey Nikolenko, Evgeny Burnaev

To cite this version:

Eduard Tulchinskii, Kristian Kuznetsov, Daniil Cherniavskii, Serguei Barannikov, Sergey Nikolenko, et al.. Topological Data Analysis for Speech Processing. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2023, pp.311-315. 10.21437/Interspeech.2023-1861 . hal-04250733

HAL Id: hal-04250733

https://hal.science/hal-04250733v1

Submitted on 20 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

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]Skoltech, Russia; [2]CNRS, Université Paris Cité, France;
[3]Artificial Intelligence Research Institute (AIRI), Russia;
[4]St. Petersburg Department of the Steklov Institute of Mathematics, Russia

Eduard.Tulchinskiy@skoltech.ru

Abstract

We apply topological data analysis (TDA) to speech classification problems and to the introspection of a pretrained speech model, HuBERT. To this end, we introduce a number of topological and algebraic features derived from Transformer attention maps and embeddings. We show that a simple linear classifier built on top of such features outperforms a fine-tuned classification head. We achieve an improvement of about 9% accuracy and 5% ERR on two common datasets; on *CREMA-D*, the proposed feature set reaches a new state of the art performance with accuracy 80.155. We also show that topological features are able to reveal functional roles of speech Transformer heads; e.g., we find the heads capable to distinguish between pairs of sample sources (natural/synthetic) or voices without any downstream fine-tuning. Our results demonstrate that TDA is a promising new approach for speech analysis, especially for tasks that require structural prediction.

Index Terms: TDA, HuBERT, interpretability, emotion recognition

1. Introduction

The paradigm of learning universal large-scale Transformer-based models has been recently transferred to speech from natural language processing [1, 2, 3]. Due to the high complexity of speech data, the preferred way of downstream task adaptation for such models is to keep pretrained weights frozen and fine-tune small task-specific heads [4], so it is important to use the model's hidden states in the most efficient way. A common approach is to use the outputs of the last layer of the model, combining them via various pooling methods, although it has been shown that for some tasks, e.g. phoneme prediction in HuBERT or the analogy task in BERT, embeddings from lower and middle layers are more useful [5]. *Topological data analysis* (TDA) is a recently proposed way to obtain more efficient data representations from frozen Transformer weights [6, 7]. TDA features prove to be better suited for many downstream tasks, including artificial text detection and linguistic acceptability judgement. In particular, for ungrammatical sentence detection conventional sentence embeddings yield classification quality no better than random, while TDA has led to meaningful results. Inspired by these results, in this work we apply TDA to the HuBERT model to construct more powerful speech representations.

TDA has already been applied to signals of various nature. Previous attempts for speech used topological properties of the audio signal: persistent entropy for noise classification [8] and emotion recognition [9], detection of a periodic signal in noisy data [10] etc. However, TDA for Transformer-based models has so far been limited to natural language processing (NLP): TDA for attention maps for artificial text detection [6] and linguistic acceptability [7], TDA for word embeddings for dialogue term extraction [11] and constructing story trees [12]. The evolution of inner representations in neural networks has been studied with persistent Betti numbers [13] and recently introduced representation topology divergence (RTD) [14], which we apply in this work to intermediate embeddings and attention maps of a speech Transformer.

In this work we also apply TDA to interpreting pretrained Transformer models. It is well known in NLP that different heads are sensitive to different phenomena [5], and we demonstrate the same effect for speech Transformers, finding heads that are best for solving specific downstream tasks: separating a given pair of emotions, a given pair of speakers, detecting speech generated by a specific TTS model, or representing spectral features of sound samples (bit rate, LPCC etc.). Below, Section 2 introduces TDA, Section 3 shows our evaluation study, Section 4 applies TDA to interpreting attention maps, and Section 5 concludes the paper.

2. Methods

Below, we first introduce topological data analysis for weighted graphs and then define the features we extract from HuBERT.

Homologies of graphs and manifolds. Topology is the study of properties of manifolds in space other than their size, i.e., loosely speaking, topology considers an object as if it were made of an infinitely stretchable material that can stretch and compress as long as there are no cuts or gluings. *Topological invariants* are properties of manifolds that do not change under topological transformations. *Homology groups* are among the most important invariants, calculated separately for different dimensions. The homology group of dimension i consists of i-dimensional closed objects on the manifold that cannot be converted into each other by a topological transformation; e.g., the 0-th homology group consists of 0-dimensional objects (points) and equals in size to the number of connected components. In this work, we are not going beyond the 0-th dimension, but our methods can easily be extended to the H_1 group, which consists of the closed paths on the manifold, or cycles in a graph. For higher dimensions, the computations may become prohibitive.

Persistent homologies for weighted graphs and point clouds. We would like to apply TDA to sets of vectors in $\mathbb{R}^d$ (point cloud). This set can be viewed as a complete graph G with edges weighted by a distance-like metric between vectors. To estimate the topological properties of this graph, it looks natural to remove weak connections by thresholding, leaving only edges with weights lower than a given ϵ. However, it is not clear which threshold to choose. TDA can track the changes of topology across varying thresholds via *persistent homologies*.

10.21437/Interspeech.2023-1861

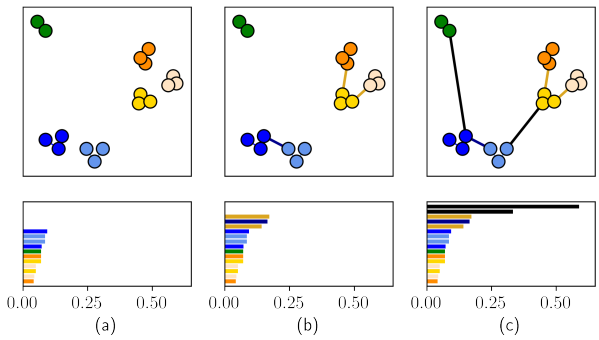

Figure 1: *Constructing an H_0 barcode for a dataset with a hierarchical structure: (a) first, small clusters are connected; (b) as the threshold grows, larger blue and yellow clusters are joined; (c) finally, all dots are joined into a single component. Each bar corresponds to an edge in the minimal spanning tree.*

Fig. 1 illustrates the persistence of 0-dimensional homologies H_0 (connected components) for a set of points. For ϵ below the minimal distance between vertices, we obtain a graph with no edges; as ϵ increases, new edges are added, ending in the complete graph. During this process, gradual changes of graph topology can be expressed in terms of the "birth" and "death" of basic features. We begin with $|V|$ connected components (all of them are "born"), and as ϵ increases, pairs of them are joined together (one component "dies"). "Birth" and "death" moments can be represented with a diagram called the *barcode* [15, 16], where the horizontal axis is a sequence of thresholds ϵ, and each horizontal bar corresponds to a single feature.

An important property of the H_0 barcode is that its bars correspond to the edges of the minimal spanning tree (MST) of the graph. Indeed, our process of adding edges as the threshold grows coincides with the classical Prim algorithm that finds the MST in a weighted graph; the length of the bar which is "dying" on every step equals the length of the corresponding MST edge. We denote by H_0^m the average length of bars in the H_0 barcode.

The H_0 barcode can capture information about the hierarchical structure of the dataset, as illustrated in Fig. 1. As we show in Section 4, our methods work well with the hierarchical information of speech signals, distinguishing the levels of frames, phonemes, words, and pauses.

Representation topology divergence (RTD) measures the topological dissimilarity between two data representations, formalized as two weighted graphs G^a and G^b with a common set of vertices. If RTD equals zero then the H_0 barcodes of the two graphs coincide. And vice versa, if the barcodes coincide and the resulting clusters of vertices are the same on all levels then RTD is zero. For details and the formal definition, see [14, 17].

Features. We use three groups of features: algebraic features of attention matrices, topological features of attention matrices, and topological features of embeddings; for comparison, we also pool embeddings from all layers. We consider HuBERT-base model [1] that consists of 12 layers with 12 attention heads each, with embedding dimension 768.

Algebraic features include the sum of the upper triangular part of the $n \times n$ attention matrix (normalized by n^2), which is used as a measure of asymmetry, and mean values of its 3 longest diagonals. This yields 4 features per attention map, 576 for the entire HuBERT model.

Topological features of attention matrices include the H_0^m feature for two graphs derived from each attention matrix A_{attn}. $H_0^{m,\text{sym}}$ is defined as H_0^m for the graph with adjacency matrix

$$A' = 1 - \max\left(A_{\text{attn}}, A_{\text{attn}}^\top\right), \tag{1}$$

that is, the symmetrization of A_{attn} (cf. [6]). $H_0^{m,\text{pc}}$ is defined as H_0^m for the rows of A_{attn} considered as a point cloud with L_1-distance. The intuition behind this choice of distance is that these points always lie on the L_1 sphere because the rows of attention matrix are normalized by *softmax*. Here we have two features per attention map, 288 in total.

Topological features of embeddings. Considering the i-th layer's embeddings $X^{(i)}$ as a point cloud with the L_2-distance, we obtain 3 features for each layer: $H_0^m(X^{(i)})$, RTD between $X^{(i)}$ and the last layer's embeddings $X^{(L)}$, and RTD between $X^{(i)}$ and initial embeddings $X^{(0)}$ (36 features in total). Besides, we add mean mel-frequency coefficients (MFCC); since we use 13 MFCCs, it gives 13 more features. Finally, we add two features: H_0^m of the MFCC features and of the initial embeddings. In total, we have 51 features for the entire model.

Embeddings from all layers. An alternative approach, inspired by [18], is to use pooled embeddings from each layer of the Transformer, not just the last. We explored two pooling strategies: averaging over the timescale (common for speech Transformers) and taking only the first embedding as suggested in [18]. This yields 9216 features in each case.

3. Evaluation

Datasets. We evaluate our method on tasks of emotion recogntion, antispoofing, and speaker verification. Our features are not applicable directly to ASR tasks because they characterize a speech sample as a whole; however, we additionally performed an evaluation on the word classification task to check that linguistic information can be captured. We have used standard datasets for these tasks. **IEMOCAP**, introduced in [19], contains $\approx$ 12 hours of audiovisual data, including video, speech, motion capture of faces, and text transcriptions; we use only speech samples from the "Anger", "Sadness", "Happiness", and "Neutral" classes (4490 samples), with 5-fold cross-validation, similar to [20, 21]. **CREMA-D** [22] has 7442 clips from 91 actors who spoke 12 different sentences in one of six basic emotions (anger, disgust, fear, happy, neutral, and sad); we perform multi-class classification with 6 classes and evaluate by averaging five splits with 70/15/15% train/development/evaluation subsets. **ASVSpoof** [23] was presented for the 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge; we used the standard split with 25380 training, 24844 development, and 71237 evaluation samples and performed classification into generated and bonafide labels, a standard task on ASVSpoof [24]. **VoxCeleb1** [25] contains over 100K utterances by 1251 celebrities extracted from *YouTube* videos; we performed binary classification between pairs of utterances from the same or different speakers, with 40000 pairs in the training set, 8000 in development, and 37720 in the test set; in this dataset, we clip each utterance to the first 5 seconds for all methods, due to the amount of data. **FSDD**[1] (Free Spoken Digit Dataset) is a simple dataset consisting of 3000 short recordings of spoken digits. Each recording was obtained from one of 6 speakers, and there are 50 recordings of each digit per speaker. The test set consists of 10% of the recordings that include all speakers and digits. We performed multi-class classification on this task.

[1] https://doi.org/10.5281/zenodo.1342401

Table 1: *Experimental results; ⋆ — from SUPERB [4].*

Model	IEMO CAP Acc ↑	CREMA-D Acc ↑	ASV Spoof EER ↓	Vox Celeb1 EER ↓	FSDD Acc ↑
HuBERT (baseline)	64.92⋆	71.047 ±0.566	6.649	**7.45**	99.3 ±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**	8.46	97.7 ±0.5
Attention features	69.666 ±1.174	79.200 ±1.240	2.138	30.326	98.7 ±0.8
Attn. & non-attn. feat.	**69.955 ±0.972**	**80.155 ±0.680**	1.946	26.443	**99.6 ±0.4**

Models. We utilize the *HuBERT Base* model [1] pretrained for automatic speech recognition on 960 hours of audio from the *Librespeech* corpus [26]. We use HuBERT as a pretrained frozen instance, without fine-tuning or any other adjustment of the weights. We use a linear layer trained over the pooled output of the Transformer as the baseline, which is consistent with the SUPERB leaderboard [4]; we used the results for IEMOCAP and VoxCeleb1 published there and trained the baselines ourselves for the other datasets.

As our TDA-based approach, we train a logistic regression model with L_1-regularization over the set of features computed from HuBERT attention maps and/or embeddings. For automatic speaker verification (checking if two utterances are made by the same person) we compute the absolute value of elementwise differences between features of both utterances.

Experimental results. Table 1 shows experimental results on all five datasets; we report accuracy (Acc, in %) and equal error rate (EER, in %) for models trained on four different sets of features (see Section 2): (1) *Attention features* is a combination of algebraic and topological features calculated from attention maps; (2) *Attn. & non-attn. feat.* denotes a combination of algebraic and topological features of attention maps and topological features of embeddings; this combines all our TDA features for this task; (3) *All layer embs, 1st* is the concatenation of first embeddings from all HuBERT layers; (4) *All layer embs, mean* is the concatenation of all HuBERT embeddings with timescale averaging. Below we show our conclusions from Table 1.

First, on all datasets averaging the embeddings is a much better strategy than taking just the first. One reason might be that in HuBERT data representation there is no a fictional first token that represents the sample as a whole, similar to [CLS] in NLP Transformers. Second, adding non-attention features improves performance compared to just attention-based features, as expected since they contain more information, although for IEMOCAP the improvement is quite marginal. Third, topological features give better results than embeddings from all layers on multi-class emotion recognition datasets (IEMOCAP, CREMA-D). On emotion recognition datasets (IEMOCAP and CREMA-D) we achieve major improvements over the baseline (conventional usage of HuBERT). Our results on VoxCeleb1 are very close to the baseline (the second best values in every column of Table 1 are underlined), and on FSDD our method performs slightly better than the baseline and very close to perfect accuracy. For CREMA-D, we achieve a new state of the art result, improving the previous SOTA of 70.47% [27] by over 9%.

For emotion recognition tasks, topological features provide clear benefits over all considered baselines. For generated speech detection and speaker verification, attention features do not help with task solving in general, but in the next section we show that features computed for individual heads perform surprisingly well for restricted subtasks in the zero-shot setting.

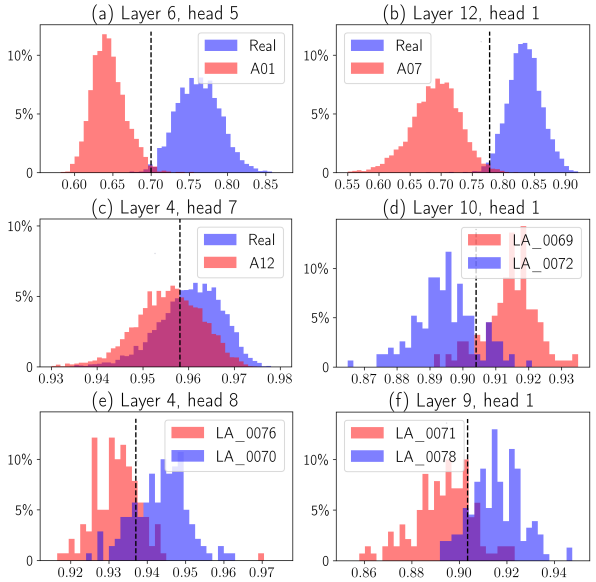

Figure 2: $H_0^{m,\mathrm{sym}}$ *for the best HuBERT heads for two tasks: (a-c) individual model separation between human (blue) and synthetic (red) speech; (d-e) speaker separation; LA_0069, LA_0072, LA_0078 – female speakers, LA_0070, LA_0076, LA_0071 – male speakers.*

4. TDA for interpreting attention maps

Restricted tasks. Inspired by recent works on attention interpretability in natural language processing [5], we analyze the roles of individual attention heads in HuBERT. Since their "areas of expertise" are much more narrow than general problems considered above, we use two restricted tasks: separation of individual models (one synthetic model vs real speech) and separation (binary classification) of two speakers.

For individual model separation, we collected all samples produced by a given synthetic model and an equal amount of bonafide samples (real speech) randomly selected from train and validation sets. For each HuBERT attention head, we calculate the distributions of the $H_0^{m,\mathrm{sym}}$ feature for synthetic and real samples and rank the heads by separation quality defined as

$$\mathrm{SQ}_{1,2} = \frac{|m_1 - m_2|}{\max(\sigma_1, \sigma_2)}, \tag{2}$$

where m_i are the means and σ_i are the variances of classes. Fig. 2 shows sample separations; it turns out that for every voice model there are several heads that separate them well ($\mathrm{SQ} > 1$) and for most models there are heads that separate them very well ($\mathrm{SQ} > 3$). The best heads are different and are usually situated in the middle-to-top layers of HuBERT. The worst results were obtained on the A19 model that was specially fine-tuned on evaluation data (best $\mathrm{SQ} = 1.45$). Separation quality of the threshold classifier for the best head varies from EER 0.03% for the A14 model ($\mathrm{SQ} = 3.5$) to EER 36.5% for the A19 model. For individual speaker separation, a similar approach

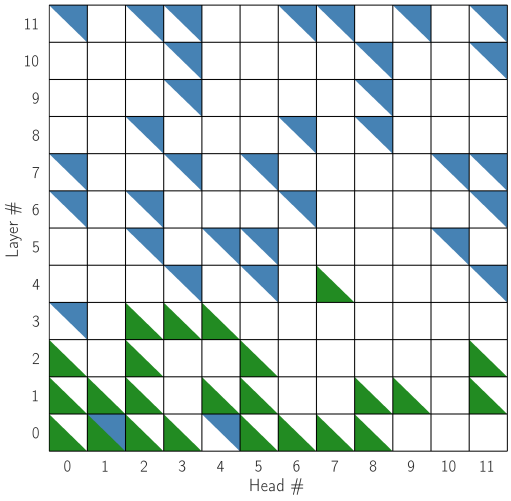

Figure 3: *Distribution of heads with high correlation (Pearson correlation coefficient* ≥ 0.5*) of* $H_0^{m,\mathrm{pc}}$ *and standard acoustic features: blue, MFCC; green, PLP.*

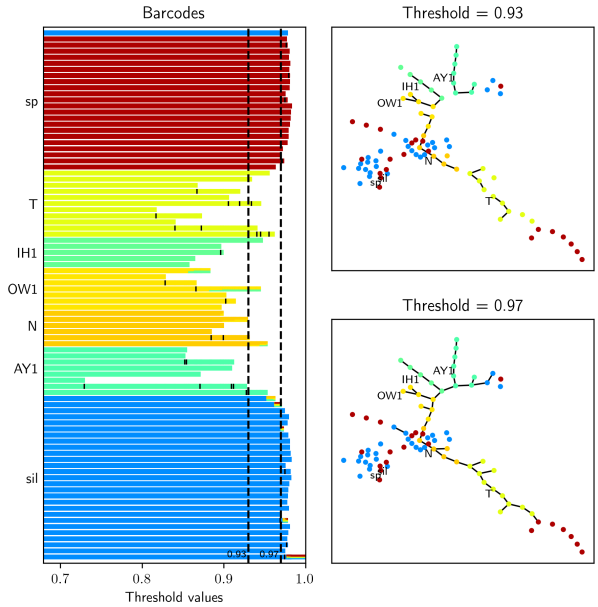

Figure 4: *Sample barcode and MST on different levels for head (2, 4) for one speech sample; sample text: "I know it", sample phones: "sil AY1 N OW1 IH1 T sp". Nodes and bars are colored with respect to the phonemes they represent. Black dashed lines show barcode levels corresponding to the trees on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

on ASVSpoof also shows that every pair of speakers has heads with good separation (SQ > 1), but this time there are no heads with SQ > 3. Fig. 2(d-e) shows sample separations. On the other hand, more general tasks are not handled well by individual heads; e.g. the best achieved separation of male vs female speakers has SQ $= 0.72$.

For emotion pair separation our approach yields results that are quite similar with those for model and speaker separation. They, and additional results for other experiments from this section, are presented at the website[2].

Attention maps and classical acoustic features. HuBERT model works with raw signals, but its attention mechanisms can extract a lot of varied information, so we searched for potential similarities between TDA features and common acoustic features extracted to process speech: mel-frequency cepstral coefficients (MFCC) and perceptual linear predictive features (PLP). Following [28], we extract the acoustic features for samples from the ASVSpoof dataset (real human speech only) and compute Pearson correlations between them and attention features. Fig. 3 shows the results for $H_0^{m,\mathrm{pc}}$; heads that have a strong correlation between $H_0^{m,\mathrm{pc}}$ and mean MFCC and/or mean PLP are marked in Fig. 3. Note that PLP features are detected mainly on lower layers, and MFCC on the middle and top layers of the model.

Interpretation of 0-dimensional barcodes. Fig. 4 illustrates how the H_0 barcode can capture hierarchical information in a speech sample. We take a sample from *LibriSpeech* with phoneme alignment [29] and consider the graph built from the adjacency matrix by formula (1).

Each bar is colored with respect to the proportion of each phoneme in the corresponding connected component. When edges are absent, each bar corresponds to the component of one vertex, so the entire component belongs to a single phoneme. For larger thresholds, mixed components will appear, which would change the bar's color. However, on the left of Fig. 4 we can see that all bars do not change color until very high threshold values. Moreover, there is a clear difference in length between bars corresponding to phonemes (green, yellow, orange) and bars in non-speech parts labeled as *sp* (pause, red) or *sil* (silence, blue). In general, the barcode reflects the hierarchy: there is a large number of short single-phoneme bars, a few longer bars connect phonemes to each other, and finally the longest bars in Fig. 4 correspond to non-speech parts of the sample.

These structure levels can be observed in the right-hand part of Fig. 4. The top right graph shows each phoneme in a separate connected component, while the bottom right graph shows the whole phrase covered by the tree, but silence and pause nodes stay apart. It means that as the threshold increases, the vertices first form phoneme clusters, then connect the whole sentence, and only after that silence is attached. This tendency can be observed in many examples on specific heads.

5. Conclusion

In this work, we have applied topological data analysis to solving downstream tasks based on a pretrained HuBERT model and analysis and interpretation of individual attention heads. We have shown that TDA yields compact feature sets that give excellent results for tasks such as emotion recognition and speaker classification, including a new state of the art result on CREMA-D. Besides, we have shown how TDA can help interpret individual heads in a Transformer-based architecture, and shown that the structure defined by TDA features corresponds well to the semantic structure of a speech sample. We believe that topological analysis is an important and currently underexplored venue of research for large machine learning models such as Transformers, and propose TDA as a potentially fruitful direction of study.

6. Acknowledgements

The work of E.T., S.B. and E.B. was partially supported by the Analytical center under the RF Government (subsidy agreement 000000D730321P5Q0002, Grant No. 70-2021-00145 02.11.2021).

[2] topohubert.github.io/speech-topology-webpages

7. References

[1] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[3] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 244–250.

[4] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T. hsien Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. rahman Mohamed, and H. yi Lee, "Superb: Speech processing universal performance benchmark," in *Interspeech*, 2021.

[5] I. Vulić, E. M. Ponti, R. Litschko, G. Glavaš, and A. Korhonen, "Probing pretrained language models for lexical semantics," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7222–7240.

[6] L. Kushnareva, D. Cherniavskii, V. Mikhailov, E. Artemova, S. Barannikov, A. Bernstein, I. Piontkovskaya, D. Piontkovski, and E. Burnaev, "Artificial text detection via examining the topology of attention maps," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 635–649. [Online]. Available: https://aclanthology.org/2021.emnlp-main.50

[7] D. Cherniavskii, E. Tulchinskii, V. Mikhailov, I. Proskurina, L. Kushnareva, E. Artemova, S. Barannikov, I. Piontkovskaya, D. Piontkovski, and E. Burnaev, "Acceptability judgements via examining the topology of attention maps," in *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, dec 2022, pp. 88–107. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.7

[8] M. Rucco, R. Gonzalez-Diaz, M.-J. Jimenez, N. Atienza, C. Cristalli, E. Concettoni, A. Ferrante, and E. Merelli, "A new topological entropy-based approach for measuring similarities among piecewise linear functions," *Signal Processing*, vol. 134, pp. 130–138, 2017.

[9] R. Gonzalez-Diaz, E. Paluzo-Hidalgo, and J. Quesada, "Towards emotion recognition: A persistent entropy application," in *Computational Topology in Image Context*. Springer International Publishing, 2019, pp. 96–109.

[10] T. Fireaizen, S. Ron, and O. Bobrowski, "Alarm sound detection using topological signal processing," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 211–215.

[11] R. Vukovic, M. Heck, B. Ruppik, C. van Niekerk, M. Zibrowius, and M. Gasic, "Dialogue term extraction using transfer learning and topological data analysis," in *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Edinburgh, UK: Association for Computational Linguistics, Sep. 2022, pp. 564–581.

[12] P. Haghighatkhah, A. Fokkens, P. Sommerauer, B. Speckmann, and K. Verbeek, "Story trees: Representing documents using topological persistence," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 2413–2429.

[13] G. Naitzat and A. Zhitnikov, "Topology of deep neural networks," *Journal of Machine Learning Research*, vol. 21, pp. 1–40, 01 2020.

[14] S. Barannikov, I. Trofimov, N. Balabin, and E. Burnaev, "Representation Topology Divergence: A Method for Comparing Neural Network Representations," in *International Conference on Machine Learning*. PMLR 162, 2022.

[15] S. Barannikov, "Canonical Forms = Persistence Diagrams. Tutorial," in *European Workshop on Computational Geometry (EuroCG 2021)*, 2021.

[16] F. Chazal and B. Michel, "An introduction to topological data analysis: Fundamental and practical aspects for data scientists," *Frontiers in Artificial Intelligence*, vol. 4, 2021.

[17] S. Barannikov, "The framed Morse complex and its invariants," *Advances in Soviet Mathematics*, vol. 21, pp. 93–115, 1994.

[18] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[19] C. Busso, M. Bulut, C.-C. Lee, E. A. Kazemzadeh, E. M. Provost, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[20] G. Ramet, P. N. Garner, M. Baeriswyl, and A. Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 126–131.

[21] I. Gat, H. Aronowitz, W. Zhu, E. Morais, and R. Hoory, "Speaker normalization for self-supervised speech emotion recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7342–7346.

[22] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[23] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "Asvspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[24] H. Yu, Z. Tan, Z. Ma, R. Martin, and J. Guo, "Spoofing detection in automatic speaker verification systems using dnn classifiers and dynamic acoustic features," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, pp. 4633–4644, 2018.

[25] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[27] N. C. Ristea, R. T. Ionescu, and F. S. Khan, "SepTr: Separable Transformer for Audio Spectrogram Processing," in *Proc. Interspeech 2022*, 2022, pp. 4103–4107.

[28] S. Borzì, O. Giudice, F. Stanco, and D. Allegra, "Is synthetic voice detection going into the right direction?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 71–80.

[29] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech Model Pre-Training for End-to-End Spoken Language Understanding," in *Proc. Interspeech 2019*, 2019, pp. 814–818. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2396