Can we S👀 the sound of human speech?

In 1966, Mark Kac asked the famous question:

Can you hear the shape of a drum?

To hear the shape of a drum is to infer information about the shape of the drumhead from the sound it makes.

In this poster, we mirror the question across senses and address instead:

Can we see the sound of human speech?

Topology-enhanced Machine Learning for Consonant Recognition

Zeyang Ding, Pingyao Feng, Yuhe Qin, Qingrui Qu, Siheng Yi, Zhiwang Yu, Haiyu Zhang, Yifei Zhu

Southern University of Science and Technology, Shenzhen, China

Motivations

This research drew inspiration from Carlsson and his collaborators' discovery of the Klein-bottle distribution of high-contrast, local patches of natural images [1], as well as their subsequent work on topological convolutional neural networks (CNN) for learning image and video data [2, 3].

By analogy, we aim to understand a topological distribution space for speech data (cf. Fig. 1), even a directed graph structure on it modeling the complex network of speech-signal sequences for practical purposes such as speaker diarisation, and how these topological inputs may enable smarter learning.

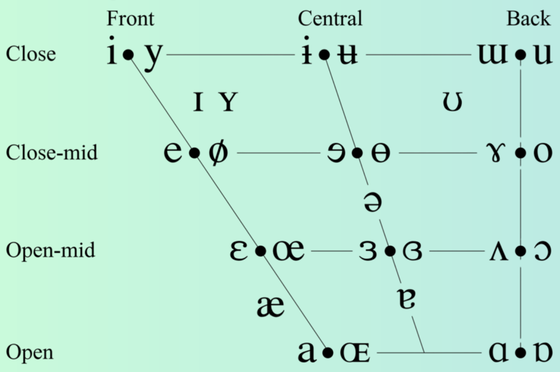


Figure 1. A charted distribution space of vowels created by linguists (from the International Phonetic Association).

Topological approaches to analysing time series data

With applications to speech recognition as one of the essential components of artificial intelligence, we established two conceptually novel approaches to address the challenges and difficulties in analysing nonlinear time series data, not limited to speech signals.

An apparently paradoxical parameter selection scheme: high ambient vs. low intrinsic dimensions

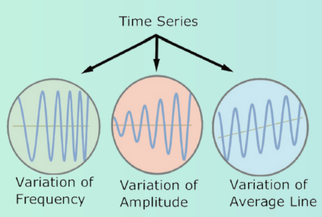
As a specific type of time series, we endow phonetic data with a point-cloud structure in a high-dimensional Euclidean space via time-delay embedding (TDE). At this preprocessing stage for TDA, selection of the dimension becomes a critical issue. Despite the curse of dimensionality, when the data are embedded in a higher-dimensional space, the computation will be a little faster, the point cloud will appear smoother and more regular, and most importantly, more salient topological features can be spotted. Fig. 2 illustrates the complexity involved in selecting multiple parameters for TDE.

dimension = 10, desired delay = 40			dimension = 50, desired delay = 8			dimension = 100, desired delay = 4		
delay	skip	MP	delay	skip	MP	delay	skip	MP
1	1	0.0610	1	1	0.2834	1	1	0.4270
10	1	0.1299	3	1	0.3021	2	1	0.4337
20	1	0.1312	4	1	0.3054	2	5	0.4146
30	1	0.1281	5	1	0.3058	3	1	0.4357
39	1	0.1229	6	1	0.3042	3	5	0.4120
39	5	0.1134	7	1	0.3052	4	1	0.4381
40	1	0.1290	7	5	0.2886	4	5	0.4139
40	5	0.1195	8	1	0.3093	5	1	0.4375
41	1	0.1200	8	5	0.2928	5	5	0.4105
41	5	0.1153	9	1	0.3091	6	1	0.4347
45	1	0.0940	9	5	0.2913	6	5	0.4114
50	1	0.1226	10	1	0.3069	7	1	0.4380
60	1	0.1315	15	1	0.3070	8	1	0.4378
94	1	empty	18	1	empty	9	1	empty

Figure 2. The computed topological desciptor of maximal persistence (MP) per choices of dimension, delay, and skip in TDE. Empty in MP means the delay is too large to obtain point-cloud data. Desired delay is theoretically deduced according to work of Perea and Harer.

Beyond periodicity: detecting the "three fundamental variations"

After TDE from above, 1-dimensional persistence diagrams (PDs) are then computed using persistent homology. In many cases, points distributed near the birth–death diagonal line in a PD are regarded as descriptors of noise and are often disregarded. However, using simulated data, we demonstrated that by noting patterns in these regions, PD can distinguish three kinds of "fundamental variations" as finer structures inherent in time series data: namely, variabilities of frequency, of amplitude, and of average line. Moreover, it can detect how significant they are (Fig. 3).



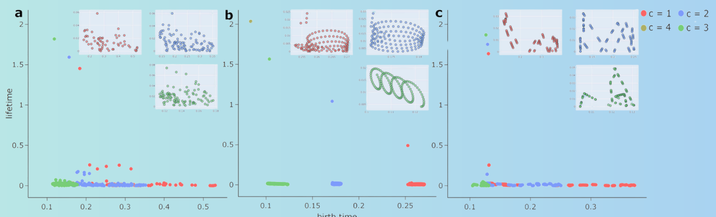


Figure 3. 1-dimensional PD reveals three fundamental variations. a, Detecting variation of frequency. b, Detecting variation of amplitude. c, Detecting variation of average line.

Based on the approaches above, we provide a transparent and broadly applicable methodology, TopCap, to capture topological features inherent in time series for machine learning.

Our method: TopCap

In view of the capability of topological methods to discern vibration patterns in time series, we applied them to classifying consonant signals into voiced and voiceless categories.

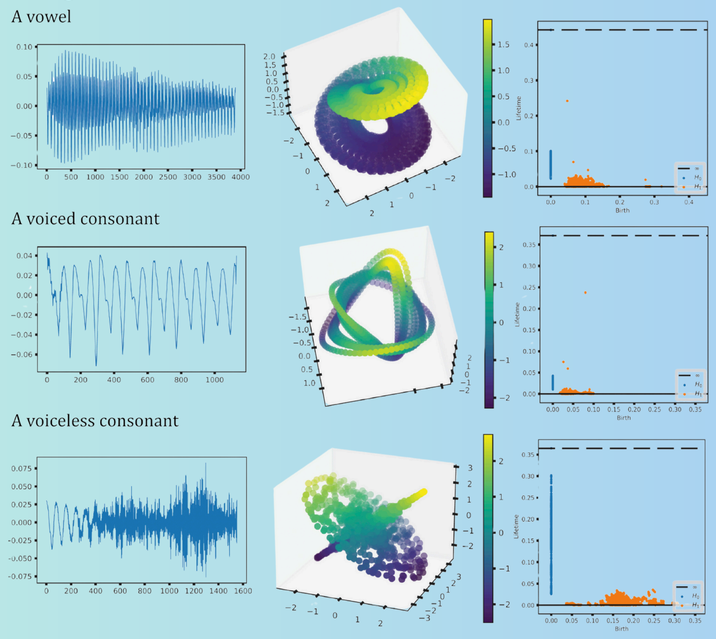


Figure 4. The varied shapes of vowels, voiced/voiceless consonants.

Rooted in high-dimensional ambient spaces, TopCap is capable of capturing features rarely detected in datasets with low intrinsic dimensionality. Compared to prior approaches, we obtain descriptors which probe finer information such as the vibration of a time series. This information is then vectorised and fed to multiple machine learning algorithms to do classification. The detailed process of TopCap goes as follows.

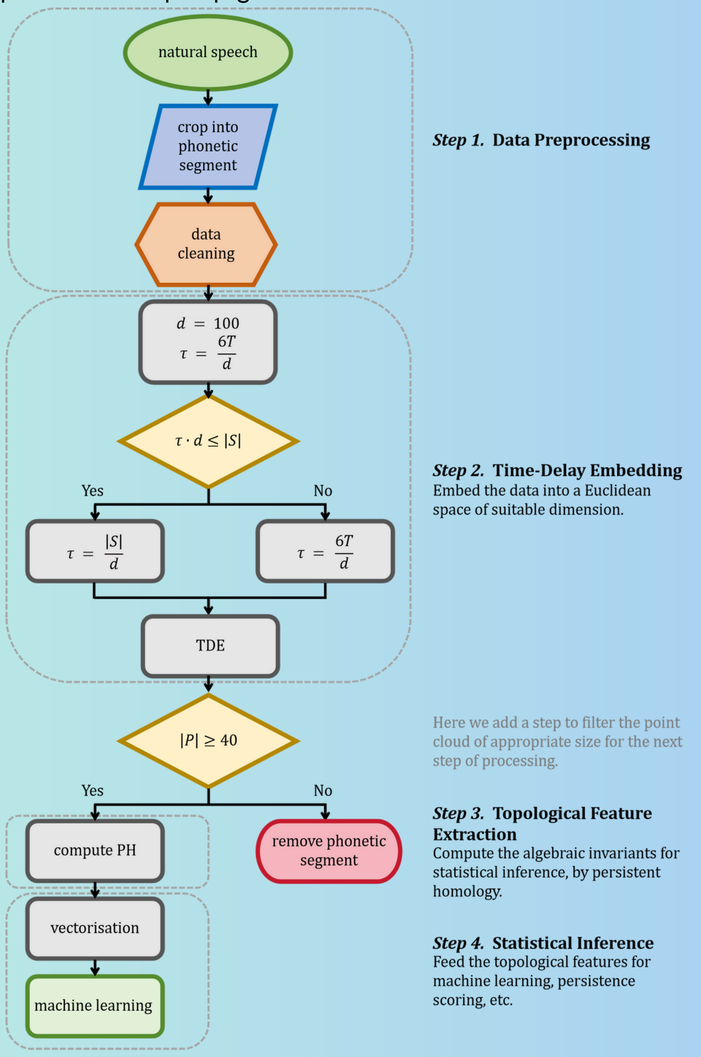


Figure 5. A pipeline for TopCap. Here $|S|$ denotes the number of samples in a time series, $|P|$ denotes the number of points in the point cloud, and T denotes the (minimal) period of the time series computed by an autocorrelation-like function.

As a demonstration of the effectiveness in both accuracy and efficiency, our streamlined algorithm TopCap significantly outperformed traditional deep learning neural networks for the classification of voiced and voiceless consonants from real human speech data. Using speech files from SpeechBox, our topological approach achieved an average accuracy exceeding 96%.

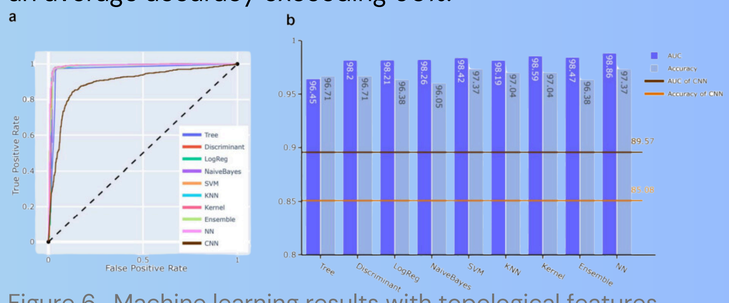


Figure 6. Machine learning results with topological features. a, ROCs of TopCap's traditional machine learning algorithms with topological inputs and of CNN without topological inputs. b, Accuracy and AUC of TopCap vs. those of CNN.

We have been experimenting with more extensive datasets, including LJSpeech, LibriSpeech and TIMIT, as well as extending comparison of our approach to state-of-the-art methods to demonstrate its advantages.

Applying topology to deep learning

Using persistent homology, Carlsson et al. qualitatively analysed approximately 4.5 x 10^6 high-contrast local patches of natural images obtained by Hateren et al. [1]. In their 2008 article, they discovered that, as vectors of pixels, the image data were unevenly distributed over a Klein bottle within the 7-dimensional Euclidean sphere! A decade later, Carlsson and his collaborators utilised TDA to analyse the architecture of CNNs, improving model explainability [3]. Moreover, they used the Klein-bottle distribution as a topological input for designing convolutional layers in neural networks that learn image data and even video data. Both learnings achieved higher accuracies with smaller training sets [2].

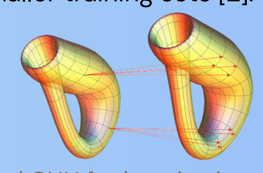


Figure 7. A topological CNN for learning image and video data.

As a warm-up, our research group have reproduced some of their results.

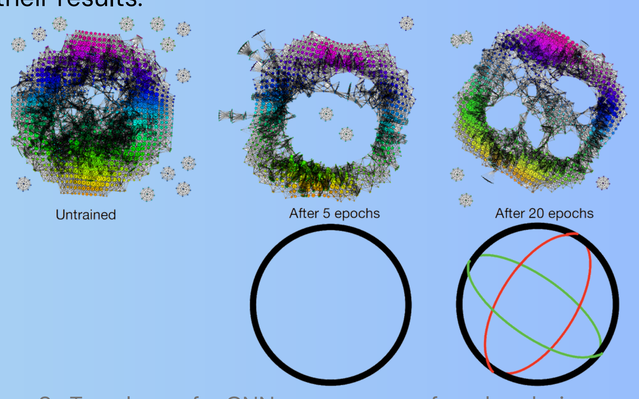


Figure 8. Topology of a CNN: emergence of cycles during a training process (reproduced using GUDHI after Carlsson and Gabrielsson 2018).

However, our preliminary work on distribution space for speech data through "explainable neural networks" has indicated that the situation is quite different from that of image data.

Conjecture. Instead of a universal distribution analogous to the Klein bottle for local image data, specific distributions apply to specific languages (or systems of phones) and are trained through the human brain.

References

[0] Pingyao Feng, Siheng Yi, Qingrui Qu, Zhiwang Yu, and Yifei Zhu, *Topology-enhanced machine learning for consonant recognition*, Preprint available in Research Square (also through QR code).

[1] Gunnar Carlsson, Tigran Ishkhanov, Vin de Silva, and Afra Zomorodian, *On the local behavior of spaces of natural images*, Int. J. Comput. Vis. 76 (2008), no. 1, 1–12.

[2] Ephy R. Love, Benjamin Filippenko, Vasileios Maroulas, and Gunnar Carlsson, *Topological convolutional layers for deep learning*, J. Mach. Learn. Res. 24 (2023), Paper No. 59, 35.

[3] Gunnar Carlsson and Rickard Brüel Gabrielsson, *Topological approaches to deep learning*, Topological data analysis—the Abel Symposium 2018, Abel Symp., vol. 15, Springer, Cham, 2020, 119–146.

For more information, please scan the QR code to visit our website.